SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 1–33

NONLINEAR INSTABILITY OF A CRITICAL TRAVELING WAVE IN THE GENERALIZED KORTEWEG–DE VRIES EQUATION*

ANDREW COMECH†, SCIPIO CUCCAGNA‡, AND DMITRY E. PELINOVSKY§

Abstract. We prove the instability of a "critical" solitary wave of the generalized Korteweg–de Vries equation, the one with the speed at the border between the stability and instability regions. The instability mechanism involved is "purely nonlinear" in the sense that the linearization at a critical soliton does not have eigenvalues with positive real part. We prove that critical solitons correspond generally to the saddle-node bifurcation of two branches of solitons.

Key words. Korteweg–de Vries equation, critical soliton, dynamic instability, orbital stability

AMS subject classifications. 35Q51, 35Q53, 70K50

DOI. 10.1137/060651501

1. Introduction and main results.

We consider the generalized Korteweg–de Vries (KdV) equation in one dimension,

$$\partial_t \boldsymbol{u} = \partial_x \left(-\partial_x^2 \boldsymbol{u} + f(\boldsymbol{u})\right), \qquad \boldsymbol{u} = \boldsymbol{u}(x,t) \in \mathbb{R}, \quad x \in \mathbb{R}, \tag{1.1}$$

where $f \in C^\infty(\mathbb{R})$ is a real-valued function that satisfies

$$f(0) = f'(0) = 0. \tag{1.2}$$

Depending on the nonlinearity f, (1.1) may admit solitary wave solutions, or solitons, of the form $\boldsymbol{u}(x,t) = \boldsymbol{\phi}_c(x - ct)$. Generically, solitons exist for speeds c from (finite or infinite) intervals of a real line. For a particular nonlinearity f, solitons with certain speeds are (orbitally) stable with respect to the perturbations of the initial data, while others are linearly (and also dynamically) unstable. We will study the stability of the critical solitons, the ones with speeds c on the border of stability and instability regions. These solitons are no longer linearly unstable. Still, we will prove their instability, which is the consequence of the higher algebraic multiplicity of the zero eigenvalue of the linearized system.

When $f(\boldsymbol{u}) = -3\boldsymbol{u}^2$, (1.1) turns into the classical KdV equation

$$\partial_t \boldsymbol{u} + \partial_x^3 \boldsymbol{u} + 6\boldsymbol{u}\partial_x \boldsymbol{u} = 0, \tag{1.3}$$

which is well known to have solitons

$$\boldsymbol{u}_c(x,t) = \boldsymbol{\phi}_c(x - ct) = \frac{c}{2\cosh^2\left(\frac{\sqrt{c}}{2}(x - ct)\right)}, \qquad c > 0.$$

*Received by the editors February 4, 2006; accepted for publication (in revised form) October 5, 2006; published electronically April 10, 2007.
http://www.siam.org/journals/sima/39-1/65150.html

†Department of Mathematics, Texas A&M University, College Station, TX 77843 (comech@math.tamu.edu). This author was partially supported by the Max-Planck Institute for Mathematics in the Sciences (Leipzig) and by NSF grants DMS-0434698 and DMS-0621257.

‡DISMI, University of Modena and Reggio Emilia, Reggio Emilia 42100, Italy (cuccagna@unimore.it). This author was fully supported by a special grant of the Italian Ministry of Education, University and Research.

§Department of Mathematics, McMaster University, Hamilton, ON L8S 4K1, Canada (dmpeli@math.mcmaster.ca).

For $f(\boldsymbol{u}) = -\boldsymbol{u}^p$, $p > 1$, we obtain the family of generalized KdV equations (also known as gKdV-k with $k = p - 1$) that have the form

$$\partial_t \boldsymbol{u} + \partial_x^3 \boldsymbol{u} + \partial_x(\boldsymbol{u}^p) = 0. \tag{1.4}$$

They also have solitary wave solutions. All solitary waves of the classical KdV equation and of the subcritical generalized KdV equations ($1 < p < 5$) are orbitally stable; see [Ben72, Bon75, Wei87, ABH87]. Orbital stability is defined in the following sense.

DEFINITION 1.1. *The traveling wave $\boldsymbol{\phi}_c(x - ct)$ is said to be orbitally stable if for any $\epsilon > 0$ there exists $\delta > 0$ so that for any $\boldsymbol{u}_0$ with $\|\boldsymbol{u}_0 - \boldsymbol{\phi}_c\|_{H^1} \le \delta$ there is a solution $\boldsymbol{u}(t)$ with $\boldsymbol{u}(0) = \boldsymbol{u}_0$, defined for all $t \ge 0$, such that*

$$\sup_{t\ge 0} \inf_{s\in\mathbb{R}} \|\boldsymbol{u}(x,t) - \boldsymbol{\phi}(x - s)\|_{H^1} < \epsilon,$$

where $H^1 = H^1(\mathbb{R})$ is the standard Sobolev space. Otherwise the traveling wave is said to be unstable.

Equation (1.1) is a Hamiltonian system, with the Hamiltonian functional

$$E(\boldsymbol{u}) = \int_{\mathbb{R}} \left(\frac{1}{2}(\partial_x \boldsymbol{u})^2 + F(\boldsymbol{u}) \right) dx, \tag{1.5}$$

with $F(\boldsymbol{u})$ the antiderivative of $f(\boldsymbol{u})$ such that $F(0) = 0$. There are two more invariants of motion: the mass

$$I(\boldsymbol{u}) = \int_{\mathbb{R}} \boldsymbol{u}\, dx \tag{1.6}$$

and the momentum

$$\mathscr{N}(\boldsymbol{u}) = \int_{\mathbb{R}} \frac{1}{2}\boldsymbol{u}^2\, dx. \tag{1.7}$$

Assumption 1. There is an open set $\Sigma \subset \mathbb{R}_+$ so that for $c \in \Sigma$ the equation $-c\boldsymbol{\phi}_c = -\boldsymbol{\phi}_c'' + f(\boldsymbol{\phi}_c)$ has a unique solution $\boldsymbol{\phi}_c(x) \in H^\infty(\mathbb{R})$ such that $\boldsymbol{\phi}_c(x) > 0$, $\boldsymbol{\phi}_c(-x) = \boldsymbol{\phi}_c(x)$, $\lim_{|x|\to\infty} \boldsymbol{\phi}_c(x) = 0$. The map $c \mapsto \boldsymbol{\phi}_c \in H^s(\mathbb{R})$ is C^∞ for $c \in \Sigma$ and for any s. Consequently, (1.1) admits traveling wave solutions

$$\boldsymbol{u}(x,t) = \boldsymbol{\phi}_c(x - ct), \qquad c \in \Sigma. \tag{1.8}$$

In Appendix A we specify conditions under which Assumption 1 is satisfied.

Let $\mathscr{N}_c$ and I_c denote $\mathscr{N}(\boldsymbol{\phi}_c)$ and $I(\boldsymbol{\phi}_c)$, respectively. By Assumption 1, $\mathscr{N}_c$ and I_c are C^∞ functions of $c \in \Sigma$. For the general KdV equation (1.1) with smooth $f(\boldsymbol{u})$, Bona, Souganidis, and Strauss [BSS87] show that the traveling wave $\boldsymbol{\phi}_c(x - ct)$ is orbitally stable if

$$\mathscr{N}_c' = \frac{d}{dc}\mathscr{N}_c = \frac{d}{dc}\mathscr{N}(\boldsymbol{\phi}_c) > 0 \tag{1.9}$$

and unstable if instead $\mathscr{N}_c' < 0$. See Figure 1. The criterion (1.9) coincides with the stability condition obtained in [GSS87] in the context of abstract Hamiltonian systems with $\mathbf{U}(1)$ symmetry (the theory developed there does not apply to the generalized KdV equation).

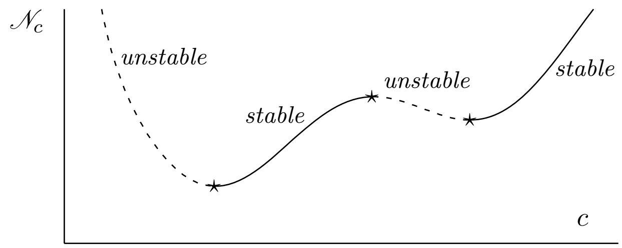


FIG. 1. *Stable and unstable regions on a possible graph of $\mathscr{N}_c$ vs. c. Three critical solitary waves are denoted by stars.*

Remark 1.2. Note that, as one can readily show, the amplitude of solitary waves is monotonically increasing with their speed c, while the momentum $\mathscr{N}_c$ does not have to.

Remark 1.3. For the generalized KdV equations (1.4), the soliton profiles satisfy the scaling relation $\boldsymbol{\phi}_c(x) = c^{\frac{1}{p-1}}\boldsymbol{\phi}_1(c^{\frac{1}{2}}x)$. The values of the momentum functional that correspond to solitons with different speeds c are given by $\mathscr{N}(\boldsymbol{\phi}_c) = \text{const}\, c^{\frac{2}{p-1}-\frac{1}{2}} = \text{const}\, c^{\frac{5-p}{2(p-1)}}$, so that $\frac{d}{dc}\mathscr{N}(\boldsymbol{\phi}_c) > 0$ for $p < 5$, in agreement with the stability criterion (1.9) derived in [BSS87].

In [BSS87] it is stated that critical traveling waves $\boldsymbol{\phi}_{c_\star}(x)$, that is, $c_\star$ such that $\mathscr{N}'_{c_\star} = 0$, are unstable as a consequence of the claim that the set $\{c\colon \boldsymbol{\phi}_c \text{ is stable}\}$ is open. This claim, however, is left unproved in [BSS87]. Moreover, this is not true in general. (This is demonstrated by the dynamical system in $\mathbb{R}^2$ described in the polar coordinates by $\dot\theta = \sin\theta$, $\dot r = 0$. The set of stationary states is the line $y = 0$; the subset of stable stationary points, $x \le 0$, is closed.) The question of stability of critical traveling waves has been left open. We address this question in this paper, proving the instability under certain rather generic assumptions. This result is the analogue of [CP03] for the generalized KdV equation (1.1).

Remark 1.4. We will not consider the L^2-critical KdV equation given by (1.4) with $p = 5$, when $\mathscr{N}_c = \text{const}$. In this case, the solitons are not only unstable but also exhibit a blow-up behavior. This blow-up is considered in a series of papers by Martel and Merle [Mer01, MM01b, MM02a, MM02b].

The analysis of the instability of critical solitary waves (with no linear instability) requires better control of the growth of a particular perturbation. We achieve this by employing the asymptotic stability methods. Pego and Weinstein [PW94] proved that the traveling wave solutions to (1.4) for the subcritical values $p = 2, 3, 4$, and also $p \in (2,5)\backslash E$ with E a finite and possibly empty set, are asymptotically stable in the weighted spaces. Their approach was extended in [Miz01]. For other deep results of stability see [MM01a, MM05]. The proofs extend, under certain spectral hypotheses, to solitary solutions to a generalized KdV equation (1.1) with c such that $\mathscr{N}'_c > 0$.

Substituting $\boldsymbol{u}(x,t) = \boldsymbol{\phi}_c(x-ct) + \boldsymbol{\rho}(x-ct,t)$ into (1.1) and discarding terms nonlinear in $\boldsymbol{\rho}$, we get the linearization at $\boldsymbol{\phi}_c$:

$$\partial_t \boldsymbol{\rho} = \partial_x\big(-\partial_x^2\boldsymbol{\rho} + f'(\boldsymbol{\phi}_c)\boldsymbol{\rho} + c\boldsymbol{\rho}\big) \equiv J\mathcal{H}_c\boldsymbol{\rho}, \tag{1.10}$$

where

$$J = \partial_x, \qquad \mathcal{H}_c = -\partial_x^2 + f'(\boldsymbol{\phi}_c) + c. \tag{1.11}$$

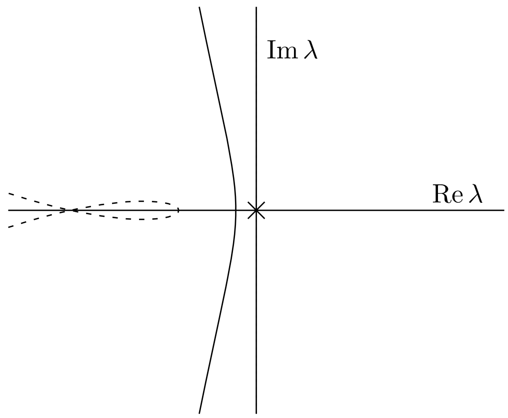


FIG. 2. *Essential spectrum of $J\mathcal{H}_c$, $c = 1$, in the exponentially weighted space $L^2_\mu(\mathbb{R})$ for $\mu = 0.1 < \sqrt{c/3}$ (solid line) and $\mu = 0.65 > \sqrt{c/3}$ (dashed line).*

In (1.10), both $\boldsymbol{\phi}_c(\cdot)$ and $\boldsymbol{\rho}(\cdot, t)$ are evaluated at $x - ct$, but we change the variable and write x instead.

The essential spectrum of $J\mathcal{H}_c$ in $L^2(\mathbb{R})$ coincides with the imaginary axis. $\lambda = 0$ is an eigenvalue (with $\partial_x \boldsymbol{\phi}_c$ being the corresponding eigenvector). To use the asymptotic stability methods from [PW94], we will consider the action of $J\mathcal{H}_c$ in the exponentially weighted spaces. For $s \in \mathbb{R}$ and $\mu \geq 0$, we define

$$H^s_\mu(\mathbb{R}) = \{\boldsymbol{\psi} \in H^s_{loc}(\mathbb{R})\colon\, e^{\mu x}\boldsymbol{\psi}(x) \in H^s(\mathbb{R})\}\,, \qquad \mu \geq 0, \tag{1.12}$$

where $H^s(\mathbb{R})$ is the standard Sobolev space of order s. We also denote $L^2_\mu(\mathbb{R}) = H^0_\mu(\mathbb{R})$. We define the operator $A^\mu_c = e^{\mu x} \circ J\mathcal{H}_c \circ e^{-\mu x}$, where $e^{\pm \mu x}$ are understood as the operators of multiplication by the corresponding functions, so that the action of $J\mathcal{H}_c$ in $L^2_\mu(\mathbb{R})$ corresponds to the action of A^μ_c in $L^2(\mathbb{R})$. The explicit form of A^μ_c is

$$A^\mu_c = e^{\mu x} \circ J\mathcal{H}_c \circ e^{-\mu x} = (\partial_x - \mu)\big[-(\partial_x - \mu)^2 + c - f'(\boldsymbol{\phi}_c)\big]. \tag{1.13}$$

The domain of A^μ_c is given by $D(A^\mu_c) = H^3(\mathbb{R})$. Since the operator $[\partial_x - \mu] f'(\boldsymbol{\phi}_c)$ is relatively compact with respect to $\mathscr{A}^\mu_c = -(\partial_x - \mu)^3 + c(\partial_x - \mu)$, the essential spectrum of A^μ_c coincides with that of $\mathscr{A}^\mu_c$ and thus is given by

$$\sigma_{\mathrm{e}}(A^\mu_c) = \big\{\lambda \in \mathbb{C}\colon\, \lambda = \lambda_{\mathrm{cont}}(k) = (\mu - ik)^3 - c(\mu - ik), \quad k \in \mathbb{R}\big\}\,. \tag{1.14}$$

The essential spectrum of $\mathscr{A}^\mu_c$ is located in the left half-plane for $0 < \mu < \sqrt{c}$ and is simply connected for $0 < \mu < \sqrt{c/3}$; see Figure 2.

We need assumptions about the existence and properties of a critical wave.

Assumption 2. There exists $c_\star \in \Sigma \backslash \partial\Sigma$, $c_\star > 0$, such that $\mathscr{N}'_{c_\star} = 0$.

Remark 1.5. Let us give examples of the nonlinearities that lead to the existence of critical solitary waves. Take $f_-(z) = -Az^p + Bz^q$, with $2 < p < q$, $A > 0$, $B > 0$, or $f_+(z) = Az^p - Bz^q + Cz^r$, with $2 < p < q < r$, $A > 0$, $B > 0$, $C > 0$. In the case of f_+, we require that B be sufficiently large so that $f_+(z)$ takes negative values on a nonempty interval $I \subset \mathbb{R}_+$. Then there will be traveling wave solutions $\boldsymbol{\phi}_c(x - ct)$ with $c \in (0, c_1)$ (also with $c = 0$ in the case of f_+) for some $c_1 > 0$.[1] Elementary

[1] The value of c_1 is determined from the system $f(z_1) + c_1 z_1 = 0$, $F(z_1) + c_1 z_1^2/2 = 0$, with F the primitive of f such that $F(0) = 0$. See Appendix A or [BL83] for more details.

computations show that the value of the momentum $\mathscr{N}_c$ goes to infinity as $c \nearrow c_1$. It also goes to infinity as $c \searrow 0$ if $p > 5$ (also if $p = 5$ in the case of f_+), so that there is a global minimum of $\mathscr{N}_c$ at some point $c_\star \in (0, c_1)$.

Assumption 3. There exists $\mu_0 \in (0, \sqrt{c_\star}/2)$ such that for $0 \le \mu \le \mu_0$ the operator $A^\mu_{c_\star}$ has no L^2-eigenvalues except $\lambda = 0$.

Remark 1.6. We require that $\mu \le \sqrt{c_\star}/2$ so that the inequality $\mu < \sqrt{c/3}$ (needed in the condition of Lemma 4.2) is satisfied for c from an open neighborhood of $c_\star$.

Assumption 4. At the critical value $c_\star$, the nondegeneracy condition $I'_{c_\star} \ne 0$ is satisfied. Here $I_c = I(\boldsymbol{\phi}_c)$ is the value of the mass functional (1.6) on the traveling wave $\boldsymbol{\phi}_c$.

Remark 1.7. If $I'_{c_\star} = 0$, then the eigenvalue $\lambda = 0$ of $J\mathcal{H}_{c_\star}$ corresponds to a Jordan block larger than 3×3. We will not consider this situation.

Our main result is that the critical traveling wave $\boldsymbol{\phi}_{c_\star}(x)$ of the generalized KdV equation (1.1) is (nonlinearly) unstable.

THEOREM 1 (main theorem). *Let Assumptions* 1, 2, 3, *and* 4 *be satisfied, and assume that $\boldsymbol{\phi}_{c_\star}$ is a critical soliton. Assume that there exists an open neighborhood $\mathcal{O}(c_\star) \subset \Sigma$ of $c_\star$ so that $\mathscr{N}'_c$ is strictly negative and nonincreasing for $c \in \mathcal{O}(c_\star)$, $c > c_\star$ (or strictly negative and nondecreasing for $c < c_\star$, or both). Then the critical traveling wave $\boldsymbol{\phi}_{c_\star}(x)$ is orbitally unstable. More precisely, there exists $\epsilon > 0$ such that for any $\delta > 0$ there exists $\boldsymbol{u}_0 \in H^1(\mathbb{R})$ with $\|\boldsymbol{u}_0 - \boldsymbol{\phi}_{c_\star}\|_{H^1} < \delta$ and $t > 0$ so that*

$$\inf_{s\in\mathbb{R}} \|\boldsymbol{u}(\cdot, t) - \boldsymbol{\phi}_{c_\star}(\cdot - s)\|_{H^1} = \epsilon. \tag{1.15}$$

Remark 1.8. For definiteness, we consider the case when $\mathscr{N}'_c$ is strictly negative and nonincreasing for $c > c_\star$, $c \in \mathcal{O}(c_\star)$. The proof for the case when $\mathscr{N}'_c$ is strictly negative and nondecreasing for $c < c_\star$, $c \in \mathcal{O}(c_\star)$, is the same.

Thus, we assume that there exists $\eta_1 > 0$ such that

$$[c_\star, c_\star + \eta_1] \subset \Sigma, \qquad \mathscr{N}'_c < 0 \quad \text{for } c \in (c_\star, c_\star + \eta_1] \subset \Sigma. \tag{1.16}$$

Strategy of the proof and the structure of the paper. In our proof, we develop the method of Pego and Weinstein [PW94] and derive the nonlinear bounds relating the energy estimate and the dissipative estimate. We follow a center manifold approach; that is, we reduce the infinite-dimensional Hamiltonian system to a finite-dimensional system which contains the main features of the dynamics. Specifically, we consider the spectral decomposition near the zero eigenvalue in section 2, and a center manifold reduction is considered in section 3, this part being similar to the approach in [CP03]. Estimates in the energy space and in the weighted space for the error terms are in sections 4 and 5. In this part of our argument we develop the approach of [PW94]. In section 6, we complete the proof of Theorem 1. In section 7, we give an alternative approach to the instability of the critical traveling wave $\boldsymbol{\phi}_{c_\star}(x)$ by a normal form argument [Car81, IA98], under the additional hypothesis that the critical point $c_\star$ of $\mathscr{N}_c$ is nondegenerate:

$$\mathscr{N}''_{c_\star} = \frac{d^2\,\mathscr{N}(\boldsymbol{\phi}_c)}{dc^2}\Big|_{c=c_\star} \ne 0. \tag{1.17}$$

The construction of traveling waves is considered in Appendix A. The details on the Fredholm alternative for $\mathcal{H}_c$ are in Appendix B. An auxiliary technical result is proved in Appendix C.

2. Spectral decomposition in $L^2_\mu(\mathbb{R})$ near $\lambda = 0$. First, we observe that for any $c \in \Sigma$ (see Assumption 1), the linearization operator $J\mathcal{H}_c$ given by (1.11) satisfies the following relations:

$$\mathcal{H}_c \boldsymbol{e}_{1,c} = 0, \qquad \text{where } \boldsymbol{e}_{1,c} = -\partial_x \boldsymbol{\phi}_c(x), \tag{2.1}$$

$$J\mathcal{H}_c \boldsymbol{e}_{2,c} = \boldsymbol{e}_{1,c}, \qquad \text{where } \boldsymbol{e}_{2,c} = \partial_c \boldsymbol{\phi}_c(x). \tag{2.2}$$

Let $\mathscr{S}(\mathbb{R})$ denote the Schwarz space of functions.

Definition 2.1. *Let $\chi_+ \in C^\infty(\mathbb{R})$ be such that $0 \le \chi_+ \le 1$, $\chi_+|_{(-\infty,-1]} = 0$, $\chi_+|_{[0,\infty)} \equiv 1$. Define $\mathscr{S}_{+,m}(\mathbb{R})$, $m \ge 0$, to be the set of functions $u \in C^\infty(\mathbb{R})$ such that $\chi_+ u \in \mathscr{S}(\mathbb{R})$ and for any $N \in \mathbb{Z}$, $N \ge 0$, there exists $C_N > 0$ such that*

$$|u^{(N)}(x)| \le C_N (1 + |x|)^m.$$

Note that for any $m \ge 0$, $\mathrm{Image}(J\mathcal{H}_c|_{\mathbb{S}_{+,m}(\mathbb{R})}) \subset \mathscr{S}_{+,m}(\mathbb{R})$. The algebraic multiplicity of zero eigenvalue of the operator $J\mathcal{H}_c$ considered in $\mathscr{S}_{+,m}(\mathbb{R})$ depends on the values of $\mathscr{N}_c'$ and I_c' as follows.

Proposition 2.2. *Fix $m \ge 0$, and consider the operator $J\mathcal{H}_c$ in $\mathscr{S}_{+,m}(\mathbb{R})$.*

(i) *The eigenvalue $\lambda = 0$ is of geometric multiplicity one, with the kernel generated by $\boldsymbol{e}_{1,c}$.*
(ii) *Assume that $c \in \Sigma$ is such that $\mathscr{N}_c' \ne 0$. Then the eigenvalue $\lambda = 0$ is of algebraic multiplicity two.*
(iii) *Assume that $c_\star \in \Sigma$ is such that $\mathscr{N}_{c_\star}' = 0$, $I_{c_\star}' \ne 0$. Then the eigenvalue $\lambda = 0$ is of algebraic multiplicity three.*

Proof. First of all we claim that in $\mathscr{S}_{+,m}(\mathbb{R})$ we have $\dim\ker J\mathcal{H}_c = 1$.

The differential equation $\mathcal{H}_c \boldsymbol{\psi} = 0$ has two linearly independent solutions. According to (2.1), one of them is $\boldsymbol{e}_{1,c}$, which is odd and exponentially decaying at infinity. The other solution is even and exponentially growing as $|x| \to \infty$ and hence does not belong to $\mathscr{S}_{+,m}(\mathbb{R})$; we denote this solution by $\boldsymbol{\Xi}_c(x)$.

Observe that if $\boldsymbol{v} \in \ker J\mathcal{H}_c$, then $\mathcal{H}_c \boldsymbol{v} = K$, $\boldsymbol{v} \in C^\infty(\mathbb{R})$. Set $\boldsymbol{v} = \frac{K}{c} + \boldsymbol{w}$. Then $\mathcal{H}_c \boldsymbol{w} = -\frac{K}{c} f'(\boldsymbol{\phi}_c)$. Since $\langle f'(\boldsymbol{\phi}_c), \boldsymbol{e}_{1,c} \rangle = 0$, by Lemma B.1 there exists a function $\boldsymbol{w}_0 \in \mathscr{S}_{+,m}(\mathbb{R})$ such that $\mathcal{H}_c \boldsymbol{w}_0 = -\frac{K}{c} f'(\boldsymbol{\phi}_c)$. So $\boldsymbol{w} = \boldsymbol{w}_0 + A\partial_x \boldsymbol{\phi}_c + B\boldsymbol{\Xi}_c$, with A and B constants. Since

$$\boldsymbol{v} = \frac{K}{c} + \boldsymbol{w} = \frac{K}{c} + \boldsymbol{w}_0 + A\partial_x \boldsymbol{\phi}_c + B\boldsymbol{\Xi}_c \in \mathscr{S}_{+,m}(\mathbb{R}),$$

we need $\boldsymbol{v}(x) \to 0$ for $x \to +\infty$, and therefore $B = 0$ and $K = 0$. Hence, $\boldsymbol{v} \in \ker \mathcal{H}_c$, proving that $\ker J\mathcal{H}_c = \ker \mathcal{H}_c$. This proves Proposition 2.2 (i).

Let us introduce the function

$$\boldsymbol{\Theta}_c(x) = \int_{+\infty}^{x} \partial_c \boldsymbol{\phi}_c(y)\, dy. \tag{2.3}$$

Then $\partial_x \boldsymbol{\Theta}_c(x) = \partial_c \boldsymbol{\phi}_c(x)$, $\lim_{x\to-\infty} \boldsymbol{\Theta}_c(x) = -I_c'$; hence $\Theta_c \in \mathscr{S}_{+,0}(\mathbb{R})$. If $\boldsymbol{v}$ satisfies

$$J\mathcal{H}_c \boldsymbol{v} = \partial_c \boldsymbol{\phi}_c(x), \qquad \lim_{x\to+\infty} \boldsymbol{v}(x) = 0, \tag{2.4}$$

then $\boldsymbol{v}(x)$ is the only solution to the problem

$$\mathcal{H}_c \boldsymbol{v} = \boldsymbol{\Theta}_c(x), \qquad \lim_{x\to+\infty} \boldsymbol{v}(x) = 0. \tag{2.5}$$

According to Lemma B.1 (see Appendix B), if $\langle \boldsymbol{e}_{1,c}, \boldsymbol{\Theta}_c\rangle = \langle \boldsymbol{\phi}_c, \partial_c \boldsymbol{\phi}_c\rangle = \mathscr{N}_c' \ne 0$, then $\boldsymbol{v}(x)$ has exponential growth as $x \to -\infty$,

$$\boldsymbol{v}(x) \propto e^{\sqrt{c}|x|}, \qquad x \to -\infty, \tag{2.6}$$

and therefore does not belong to $\mathscr{S}_{+,m}(\mathbb{R})$. This finishes the proof of Proposition 2.2 (ii).

Let us now assume that $\mathscr{N}_{c_\star}' = 0$ for some $c_\star \in \Sigma$. Then, again by Lemma B.1 with $m = 0$, there exists $\boldsymbol{e}_{3,c_\star}(x) \in \mathscr{S}_{+,0}(\mathbb{R})$ such that

$$\mathcal{H}_{c_\star} \boldsymbol{e}_{3,c_\star} = \boldsymbol{\Theta}_{c_\star}(x), \qquad \lim_{x\to+\infty} \boldsymbol{e}_{3,c_\star}(x) = 0. \tag{2.7}$$

Now let us consider $\boldsymbol{w} \in C^\infty(\mathbb{R})$ such that

$$J\mathcal{H}_{c_\star} \boldsymbol{w} = \boldsymbol{e}_{3,c_\star}, \qquad \lim_{x\to+\infty} \boldsymbol{w}(x) = 0. \tag{2.8}$$

Let $\boldsymbol{E}(x) = \int_{+\infty}^x \boldsymbol{e}_{3,c_\star}(y)\,dy$; the function $\boldsymbol{w}(x)$ satisfies $\mathcal{H}_{c_\star}\boldsymbol{w} = \boldsymbol{E}$. Taking the pairing of $\boldsymbol{E}$ with $\boldsymbol{e}_{1,c_\star}$, we get

$$\begin{aligned} \langle \boldsymbol{e}_{1,c_\star}, \boldsymbol{E}\rangle &= -\langle \boldsymbol{\phi}_{c_\star}, \boldsymbol{e}_{3,c_\star}\rangle = \langle \mathcal{H}_{c_\star}\partial_c\boldsymbol{\phi}_{c_\star}, \boldsymbol{e}_{3,c_\star}\rangle = \langle \partial_c\boldsymbol{\phi}_{c_\star}, \mathcal{H}_{c_\star}\boldsymbol{e}_{3,c_\star}\rangle \\ &= \langle \partial_x \boldsymbol{\Theta}_{c_\star}, \boldsymbol{\Theta}_{c_\star}\rangle = \frac{\boldsymbol{\Theta}_{c_\star}^2}{2}\bigg|_{-\infty}^{+\infty} = -\lim_{x\to-\infty}\frac{\boldsymbol{\Theta}_{c_\star}^2(x)}{2} = -\frac{(I_{c_\star}')^2}{2} < 0. \end{aligned} \tag{2.9}$$

(In the first equality, the boundary term does not appear because when $x \to \pm\infty$ the function $\boldsymbol{E}(x)$ grows at most algebraically while $\boldsymbol{\phi}_c$ decays exponentially.) By Lemma B.1, since $\langle \boldsymbol{e}_{1,c_\star}, \boldsymbol{E}\rangle$ is nonzero, $\boldsymbol{w}(x)$ grows exponentially as $x \to -\infty$. This proves that the algebraic multiplicity of the eigenvalue $\lambda = 0$ is exactly three. ◻

Now we would like to consider $J\mathcal{H}_c$ in the weighted space $L^2_\mu(\mathbb{R})$, $\mu > 0$. This is equivalent to considering $A_c^\mu = e^{\mu x} \circ J\mathcal{H}_c \circ e^{-\mu x}$ in $L^2(\mathbb{R})$. In what follows, we always require that

$$0 < \mu < \min(\mu_0, \mu_1), \tag{2.10}$$

with μ_0 from Assumption 3 and μ_1 from Lemma C.1.

We define

$$\boldsymbol{e}_{j,c}^\mu = e^{\mu x}\boldsymbol{e}_{j,c}, \quad j = 1, 2; \qquad \boldsymbol{e}_{3,c_\star}^\mu = e^{\mu x}\boldsymbol{e}_{3,c_\star}. \tag{2.11}$$

From Proposition 2.2, we obtain the following statement.

Corollary 2.3.

(i) *If $\mathscr{N}_c' \ne 0$, then the basis for the generalized kernel of A_c^μ in $L^2(\mathbb{R})$ is formed by the generalized eigenvectors $\{\boldsymbol{e}_{1,c}^\mu, \boldsymbol{e}_{2,c}^\mu\}$.*

(ii) *At $c_\star$ where $\mathscr{N}_{c_\star}' = 0$, $I_{c_\star}' \ne 0$, the basis for the generalized kernel of $A_{c_\star}^\mu$ in $L^2(\mathbb{R})$ is formed by the generalized eigenvectors $\{\boldsymbol{e}_{1,c_\star}^\mu, \boldsymbol{e}_{2,c_\star}^\mu, \boldsymbol{e}_{3,c_\star}^\mu\}$.*

Proof. As follows from Lemma A.1 in Appendix A,

$$|\boldsymbol{e}_{1,c}(x)| \le \mathrm{const}\, e^{-\sqrt{c}|x|}, \qquad x \in \mathbb{R}. \tag{2.12}$$

Applying Lemma A.2 to (2.2) (for both $x \ge 0$ and $x \le 0$), we also see that

$$|\boldsymbol{e}_{2,c}(x)| \le \mathrm{const}(1 + |x|)e^{-\sqrt{c}|x|}, \qquad x \in \mathbb{R}. \tag{2.13}$$

It follows that $\boldsymbol{e}^\mu_{1,c}$, $\boldsymbol{e}^\mu_{2,c} \in L^2(\mathbb{R})$.

If $\mathscr{N}_c' \neq 0$, then by (2.6) $e^{\mu x}\boldsymbol{v}(x) \neq L^2(\mathbb{R})$.

If $\mathscr{N}_c' = 0$ at $c = c_\star$, then $\boldsymbol{e}_{3,c_\star} \in \mathscr{S}_{+,0}(\mathbb{R})$ (belongs to $\mathscr{S}$ for $x \geq 0$ and remains bounded for $x \leq 0$). Moreover, applying Lemma A.2 to (2.7), we see that

$$|\boldsymbol{e}_{3,c_\star}(x)| \leq \mathrm{const}(1+|x|)e^{-\sqrt{c_\star}\,x}, \qquad x \geq 0. \tag{2.14}$$

It follows that $\boldsymbol{e}^\mu_{3,c_\star} \in L^2(\mathbb{R})$. As follows from Proposition 2.2, the function $e^{\mu x}\boldsymbol{w}(x)$ in (2.8) does not belong to $L^2(\mathbb{R})$, so the algebraic multiplicity of $\lambda = 0$ is precisely three. ◻

Lemma 2.4.

(i) *Let $c \in (c_\star, c_\star + \eta_1]$. Then there exists a simple positive eigenvalue λ_c of A^μ_c. This eigenvalue does not depend on μ.*

(ii) *λ_c is a simple eigenvalue of the operator $J\mathcal{H}_c$ considered in $L^2(\mathbb{R})$.*

(iii) *There exists a C^∞ extension of $\boldsymbol{e}_{3,c_\star}$ into an interval $[c_\star, c_\star + \eta_1]$,*

$$c \mapsto \boldsymbol{e}_{3,c} \in H^\infty_\mu(\mathbb{R}), \qquad c \in [c_\star, c_\star + \eta_1],$$

so that the frame

$$\{\boldsymbol{e}^\mu_{j,c} = e^{\mu x}\boldsymbol{e}_{j,c} \in H^\infty(\mathbb{R})\colon j = 1,2,3\}, \qquad c \in [c_\star, c_\star + \eta_1],$$

depends smoothly on c (in L^2), $X^\mu_c = \mathrm{span}\langle \boldsymbol{e}^\mu_{1,c}, \boldsymbol{e}^\mu_{2,c}, \boldsymbol{e}^\mu_{3,c}\rangle$ is the invariant subspace of A^μ_c, and $A^\mu_c|_{X^\mu_c}$ is represented in the frame $\{\boldsymbol{e}^\mu_{j,c}\}$ by the following matrix:

$$A^\mu_c|_{X^\mu_c} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_c \end{bmatrix}, \tag{2.15}$$

where λ_c equals

$$\lambda_c = -\frac{\mathscr{N}_c'}{\langle \boldsymbol{\phi}_c, \boldsymbol{e}_{3,c}\rangle}, \tag{2.16}$$

with $\langle \boldsymbol{\phi}_c, \boldsymbol{e}_{3,c}\rangle > 0$ for $c \in [c_\star, c_\star + \eta_1]$.

Proof. Due to the restriction (2.10) on μ, the essential spectrum of A^μ_c for $c \geq c_\star$ is given by (1.14) and is located strictly to the left of the imaginary axis. By Assumption 3, the discrete spectrum of $A^\mu_{c_\star}$ consists of the isolated eigenvalue $\lambda = 0$, which is of algebraic multiplicity three by Corollary 2.3. We choose a closed contour $\gamma \subset \rho(A^\mu_{c_\star})$ in $\mathbb{C}^1$ so that the interval $[0, \Lambda]$ of the real axis is strictly inside γ, where

$$\Lambda = \sup_{c \in \Sigma} \sup_{x \in \mathbb{R}} |f''(\boldsymbol{\phi}_c(x))\boldsymbol{\phi}_c'(x)|. \tag{2.17}$$

Remark 2.5. The value of Λ is chosen so that all point eigenvalues of the operator $J\mathcal{H}_c$, $c \in \Sigma$, are bounded by Λ. Indeed, if $\boldsymbol{\psi}$ satisfies $J\mathcal{H}_c\boldsymbol{\psi} = \lambda\boldsymbol{\psi}$ with $\lambda \in \mathbb{R}$, then $\boldsymbol{\psi} \in H^\infty(\mathbb{R})$ and can be assumed to be real-valued. Therefore, we have

$$\begin{aligned}\lambda\langle \boldsymbol{\psi}, \boldsymbol{\psi}\rangle &= \langle \boldsymbol{\psi}, \partial_x(-\partial_x^2 + f'(\boldsymbol{\phi}_c) + c)\boldsymbol{\psi}\rangle \\ &= -\langle \boldsymbol{\psi}', f'(\boldsymbol{\phi}_c)\boldsymbol{\psi}\rangle = -\langle \boldsymbol{\psi}\boldsymbol{\psi}', f'(\boldsymbol{\phi}_c)\rangle = \frac{1}{2}\int_{\mathbb{R}} \boldsymbol{\psi}^2 \partial_x f'(\boldsymbol{\phi}_c)\,dx,\end{aligned}$$

so that $|\lambda| \le \sup_{x\in\mathbb{R}} |f''(\boldsymbol{\phi}_c(x))\boldsymbol{\phi}_c'(x)|/2$.

We notice that for c from an open neighborhood of $c_\star$, γ belongs to the resolvent set $\rho(A_c^\mu)$. Indeed, we have

$$\frac{1}{A_c^\mu - z} = \frac{1}{A_{c_\star}^\mu - z + (A_c^\mu - A_{c_\star}^\mu)} = \frac{1}{(A_{c_\star}^\mu - z)} \frac{1}{(1 + (A_{c_\star}^\mu - z)^{-1}(A_c^\mu - A_{c_\star}^\mu))}. \tag{2.18}$$

Since $A_{c_\star}^\mu - z$, $z \in \gamma$, is invertible in L^2 and is smoothing of order three, while $A_c^\mu - A_{c_\star}^\mu$ depends continuously on c as a differential operator of order 1, the operator $(A_{c_\star}^\mu - z)^{-1}(A_c^\mu - A_{c_\star}^\mu)$ is bounded by $1/2$ as an operator in L^2 for all $z \in \gamma$ and for all c sufficiently close to $c_\star$. We assume that $\eta_1 > 0$ is small enough so that

$$\gamma \in \rho(A_c^\mu) \quad \text{for} \quad c \in [c_\star, c_\star + \eta_1]. \tag{2.19}$$

Integrating $(A_c^\mu - z)^{-1}$ along γ, we get a projection

$$P_c^\mu = -\frac{1}{2\pi i} \oint_\gamma \frac{dz}{A_c^\mu - z}, \qquad c \in [c_\star, c_\star + \eta_1]. \tag{2.20}$$

Since $\operatorname{rank} P_{c_\star}^\mu = 3$, we also have

$$\operatorname{rank} P_c^\mu = 3, \qquad c \in [c_\star, c_\star + \eta_1].$$

The three-dimensional spectral subspace $\operatorname{Range} P_{c_\star}^\mu$ corresponds to the eigenvalue $\lambda = 0$ that has algebraic multiplicity three. According to Corollary 2.3, when $\mathscr{N}_c' \ne 0$, $\lambda = 0$ is of algebraic multiplicity two, and therefore $X_c^\mu \equiv \operatorname{Range} P_c^\mu$ splits into a two-dimensional spectral subspace of A_c^μ corresponding to $\lambda = 0$ (it is spanned by $\{\boldsymbol{e}_{1,c}^\mu, \boldsymbol{e}_{2,c}^\mu\}$) and a one-dimensional subspace that corresponds to a nonzero eigenvalue.

For $c \in [c_\star, c_\star + \eta_1]$, we define

$$\tilde{\boldsymbol{e}}_{3,c}^\mu = P_c^\mu \boldsymbol{e}_{3,c_\star}^\mu, \qquad c \in [c_\star, c_\star + \eta_1]. \tag{2.21}$$

Note that $\tilde{\boldsymbol{e}}_{3,c}^\mu \in L^2(\mathbb{R})$ since P_c^μ is continuous in L^2. In the frame $\{\boldsymbol{e}_{1,c}^\mu, \boldsymbol{e}_{2,c}^\mu, \tilde{\boldsymbol{e}}_{3,c}^\mu\}$ we can write

$$A_c^\mu \tilde{\boldsymbol{e}}_{3,c}^\mu = a_c \boldsymbol{e}_{1,c}^\mu + b_c \boldsymbol{e}_{2,c}^\mu + \lambda_c \tilde{\boldsymbol{e}}_{3,c}^\mu. \tag{2.22}$$

Since the frame $\{\boldsymbol{e}_{1,c}^\mu, \boldsymbol{e}_{2,c}^\mu, \tilde{\boldsymbol{e}}_{3,c}^\mu\}$ and also $A_c^\mu \tilde{\boldsymbol{e}}_{3,c}^\mu$ depend smoothly on c (as functions from $[c_\star, c_\star + \eta_1]$ to $L^2(\mathbb{R})$; recall that f is smooth), the coefficients a_c, b_c, and λ_c are smooth functions of c for $c \in [c_\star, c_\star + \eta_1]$. It is also important to point out that a_c, b_c, and λ_c do not depend on $\mu > 0$, since if the relation (2.22) holds for certain values of a_c, b_c, and λ_c for a particular value $\mu > 0$, then, by the definition of A_c^μ, $\boldsymbol{e}_{1,c}^\mu$, $\boldsymbol{e}_{2,c}^\mu$, and $\tilde{\boldsymbol{e}}_{3,c}^\mu$, the relation (2.22) also holds for μ' from an open neighborhood of μ.

According to the construction of $\boldsymbol{e}_{3,c_\star}$ in Proposition 2.2, $a_{c_\star} = \lambda_{c_\star} = 0$ and $b_{c_\star} = 1$. We define

$$\boldsymbol{e}_{3,c}^\mu = \frac{1}{b_c + a_c \lambda_c} (\tilde{\boldsymbol{e}}_{3,c}^\mu - a_c \boldsymbol{e}_{2,c}^\mu).$$

Then $\boldsymbol{e}_{3,c}^\mu \in L^2(\mathbb{R})$ for $c \in [c_\star, c_\star + \eta_1]$. We compute

$$A_c^\mu \boldsymbol{e}_{3,c}^\mu = \boldsymbol{e}_{2,c}^\mu + \lambda_c \boldsymbol{e}_{3,c}^\mu. \tag{2.23}$$

Thus, in the frame $\{\boldsymbol{e}^\mu_{j,c}\colon\ j=1,2,3\}$ the operator $A^\mu_c|_{\mathrm{Range}\,P^\mu_c}$ has the desired matrix form (2.15). Conjugating by means of $e^{\mu x}$ we get a corresponding frame $\{\boldsymbol{e}_{j,c}\colon j=1,2,3\}$ in L^2_μ, with $\boldsymbol{e}_{3,c}$ satisfying

$$J\mathcal{H}_c\boldsymbol{e}_{3,c}=\boldsymbol{e}_{2,c}+\lambda_c\boldsymbol{e}_{3,c},\qquad \boldsymbol{e}_{3,c}\in L^2_\mu(\mathbb{R}). \tag{2.24}$$

For $c\in[c_\star,c_\star+\eta_1]$ and $z\notin\sigma(A^\mu_c)$, $R^\mu_c(z)=(A^\mu_c-z)^{-1}$ is a pseudodifferential operator of order -3, and hence P^μ_c is smoothing of order three in the Sobolev spaces $H^s(\mathbb{R})$. The bootstrapping argument applied to the relations $\boldsymbol{e}^\mu_{j,c}=P^\mu_c\boldsymbol{e}^\mu_{j,c}$ shows that $\boldsymbol{e}^\mu_{j,c}\in H^\infty(\mathbb{R})$. By definition (1.12), this means that

$$\boldsymbol{e}_{j,c}\in H^\infty_\mu(\mathbb{R}),\qquad j=1,2,3,\quad c\in[c_\star,c_\star+\eta_1]. \tag{2.25}$$

Using (2.24), we compute

$$\begin{aligned}0&=\langle\mathcal{H}_c\boldsymbol{e}_{1,c},\boldsymbol{e}_{3,c}\rangle=-\langle\mathcal{H}_cJ\boldsymbol{\phi}_c,\boldsymbol{e}_{3,c}\rangle\\&=\langle\boldsymbol{\phi}_c,J\mathcal{H}_c\boldsymbol{e}_{3,c}\rangle=\langle\boldsymbol{\phi}_c,\boldsymbol{e}_{2,c}\rangle+\lambda_c\langle\boldsymbol{\phi}_c,\boldsymbol{e}_{3,c}\rangle,\quad c\in[c_\star,c_\star+\eta_1].\end{aligned}$$

We conclude that

$$\lambda_c=-\frac{\langle\boldsymbol{\phi}_c,\boldsymbol{e}_{2,c}\rangle}{\langle\boldsymbol{\phi}_c,\boldsymbol{e}_{3,c}\rangle},\qquad c\in[c_\star,c_\star+\eta_1],$$

where $\langle\boldsymbol{\phi}_c,\boldsymbol{e}_{2,c}\rangle=\langle\boldsymbol{\phi}_c,\partial_c\boldsymbol{\phi}_c\rangle=\mathcal{N}_c'<0$. Note that $\langle\boldsymbol{\phi}_c,\boldsymbol{e}_{3,c}\rangle>0$ for $c_\star<c\le c_\star+\eta_1$, since $\langle\boldsymbol{\phi}_{c_\star},\boldsymbol{e}_{3,c_\star}\rangle>0$ by (2.9) and $\langle\boldsymbol{\phi}_c,\boldsymbol{e}_{3,c}\rangle$ does not change sign for $c_\star<c\le c_\star+\eta_1$ (this follows from the inequality $|\langle\boldsymbol{\phi}_c,\boldsymbol{e}_{3,c}\rangle|>|\mathcal{N}_c'|/\Lambda>0$; see Remark 2.5). This finishes the proof of the lemma. ∎

Remark 2.6. According to Assumption 3, we may assume that η_1 is small enough so that for $c\in[c_\star,c_\star+\eta_1]$ and $0\le\mu\le\mu_0$ there is no discrete spectrum of A^μ_c except $\lambda=0$ and $\lambda=\lambda_c$. It follows that P^μ_c is the spectral projector that corresponds to the discrete spectrum of A^μ_c.

LEMMA 2.7. *If $\lambda_c>0$, then $\boldsymbol{e}_{3,c}\in H^\infty(\mathbb{R})$.*

Proof. By Lemma 2.4, $\lambda_c>0$ is a simple eigenvalue of $J\mathcal{H}_c$ considered in $L^2(\mathbb{R})$. By (2.1), (2.2), and (2.24),

$$\boldsymbol{\psi}_c=\boldsymbol{e}_{c,1}+\lambda_c\boldsymbol{e}_{c,2}+\lambda_c^2\boldsymbol{e}_{c,3}\in C^\infty(\mathbb{R}) \tag{2.26}$$

satisfies $J\mathcal{H}_c\boldsymbol{\psi}_c=\lambda_c\boldsymbol{\psi}_c$, and also $\lim_{x\to+\infty}\boldsymbol{\psi}_c(x)=0$. Thus, $\boldsymbol{\psi}_c$ coincides with an L^2 eigenvector of $J\mathcal{H}_c\boldsymbol{\psi}_c$ that corresponds to λ_c. Therefore, $\boldsymbol{\psi}_c\in H^\infty(\mathbb{R})$. Since $\boldsymbol{e}_{c,1},\boldsymbol{e}_{c,2}\in H^1(\mathbb{R})$ and $\lambda_c\ne0$, the statement of the lemma follows from the relation (2.26). ∎

Let us also introduce the dual basis that consists of eigenvectors of the adjoint operator $(J\mathcal{H}_c)^*=-\mathcal{H}_cJ=-\mathcal{H}_c\partial_x$ which we consider in the weighted space

$$L^2_{-\mu}(\mathbb{R})=\left\{\boldsymbol{\psi}\in L^2_{loc}(\mathbb{R})\colon e^{-\mu x}\boldsymbol{\psi}(x)\in L^2(\mathbb{R})\right\},\quad \mu>0. \tag{2.27}$$

For any $c\in\Sigma$, the generalized kernel of $(J\mathcal{H}_c)^*$ contains at least two linearly independent vectors:

$$-\mathcal{H}_c\partial_x\boldsymbol{g}_{1,c}=0,\qquad -\mathcal{H}_c\partial_x\boldsymbol{g}_{2,c}=\boldsymbol{g}_{1,c}, \tag{2.28}$$

where

$$\boldsymbol{g}_{1,c}(x) = -\int_{-\infty}^{x} \boldsymbol{e}_{1,c}(y,c)\,dy = \boldsymbol{\phi}_c(x), \tag{2.29}$$

$$\boldsymbol{g}_{2,c}(x) = \int_{-\infty}^{x} \boldsymbol{e}_{2,c}(y,c)\,dy = \int_{-\infty}^{x} \partial_c \boldsymbol{\phi}_c(y)\,dy. \tag{2.30}$$

The lower limit of integration ensures that $\lim_{x\to-\infty} \boldsymbol{g}_{2,c}(x) = 0$, so that $\boldsymbol{g}_{2,c} \in L^2_{-\mu}(\mathbb{R})$.

PROPOSITION 2.8. *Assume that $c_\star \in \Sigma$ is such that $\mathscr{N}'_{c_\star} = 0$, $I'_{c_\star} \neq 0$. The eigenvalue $\lambda = 0$ of the operator $-\mathcal{H}_{c_\star}\partial_x$ is of algebraic multiplicity three in $L^2_{-\mu}(\mathbb{R})$, and there exists $\boldsymbol{g}_{3,c_\star} \in H^\infty_{-\mu}(\mathbb{R})$ such that*

$$-\mathcal{H}_{c_\star}\partial_x \boldsymbol{g}_{3,c_\star} = \boldsymbol{g}_{2,c_\star}.$$

Proof. The argument repeats the steps of the proof of Proposition 2.2. The function $\boldsymbol{g}_{3,c_\star}$ is given by

$$\boldsymbol{g}_{3,c_\star}(x) = -\int_{-\infty}^{x} \tilde{\boldsymbol{e}}_{3,c_\star}(y)dy, \tag{2.31}$$

where $\tilde{\boldsymbol{e}}_{3,c_\star}(x)$ satisfies

$$\mathcal{H}_{c_\star}\tilde{\boldsymbol{e}}_{3,c_\star} = \int_{-\infty}^{x} \boldsymbol{e}_{2,c_\star}(y)\,dy, \qquad \lim_{x\to-\infty} \tilde{\boldsymbol{e}}_{3,c_\star}(x) = 0. \tag{2.32}$$

Since $\int_{-\infty}^{x} \boldsymbol{e}_{2,c_\star}(y)\,dy$ remains bounded as $x \to +\infty$, while $\langle \boldsymbol{g}_{2,c_\star}, \boldsymbol{\phi}_{c_\star}\rangle = 0$, the function $\tilde{\boldsymbol{e}}_{3,c_\star}(x)$ remains bounded as $x \to +\infty$. This follows from Lemma B.1 of Appendix B (after the reflection $x \to -x$). Therefore, $\boldsymbol{g}_{3,c_\star}(x)$ has a linear growth as $x \to +\infty$; $\boldsymbol{g}_{3,c_\star} \in \mathscr{S}_{-,1}(\mathbb{R})$ (defined similarly to $\mathscr{S}_{+,1}$ in Definition 2.1). □

As in Lemma 2.4, one can show that there is an extension of $\boldsymbol{g}_{3,c_\star}$ into an interval $[c_\star, c_\star + \eta_1]$,

$$c \mapsto \boldsymbol{g}_{3,c} \in H^\infty_{-\mu}(\mathbb{R}), \qquad c \in [c_\star, c_\star + \eta_1],$$

so that, similarly to (2.24) and (2.25),

$$-\mathcal{H}_c\partial_x \boldsymbol{g}_{3,c} = \boldsymbol{g}_{2,c}(x) + \lambda_c \boldsymbol{g}_{3,c}, \qquad \boldsymbol{g}_{3,c} \in H^\infty_{-\mu}(\mathbb{R}), \quad c \in [c_\star, c_\star + \eta_1]. \tag{2.33}$$

Using the bases $\{\boldsymbol{e}_{j,c} \in H^\infty_{\mu}(\mathbb{R})\colon\ j = 1,2,3\}$, $\{\boldsymbol{g}_{j,c} \in H^\infty_{-\mu}(\mathbb{R})\colon\ j = 1,2,3\}$, we can write the projection operator $e^{-\mu x} \circ P^\mu_c \circ e^{\mu x}$ that corresponds to the discrete spectrum of $J\mathcal{H}_c$ in the form

$$(e^{-\mu x} \circ P^\mu_c \circ e^{\mu x})\boldsymbol{\psi} = \sum_{j,k=1}^{3} \mathcal{T}^{jk}_c \langle \boldsymbol{g}_{k,c}, \boldsymbol{\psi}\rangle \boldsymbol{e}_{j,c}, \tag{2.34}$$

with $\mathcal{T}^{jk}_c$ being the inverse of the matrix

$$\mathcal{T}_c = \{\mathcal{T}_{jk,c}\}, \quad \mathcal{T}_{jk,c} = \langle \boldsymbol{g}_{j,c}, \boldsymbol{e}_{k,c}\rangle, \quad c \in [c_\star, c_\star + \eta_1], \quad 1 \le j,k \le 3. \tag{2.35}$$

Let us introduce the functions

$$\alpha_c = \langle \boldsymbol{g}_{1,c}, \boldsymbol{e}_{3,c}\rangle, \qquad \beta_c = \langle \boldsymbol{g}_{2,c}, \boldsymbol{e}_{3,c}\rangle, \qquad \gamma_c = \langle \boldsymbol{g}_{3,c}, \boldsymbol{e}_{3,c}\rangle. \tag{2.36}$$

Since $\boldsymbol{e}_{j,c} \in L^2_\mu(\mathbb{R})$ and $\boldsymbol{g}_{j,c} \in L^2_{-\mu}(\mathbb{R})$, α_c, β_c, and γ_c are continuous functions of c for $c \in [c_\star, c_\star + \eta_1]$. Recalling that $\langle \boldsymbol{g}_{2,c}, \boldsymbol{e}_{1,c}\rangle = \langle \boldsymbol{g}_{2,c}, J\mathcal{H}_c \boldsymbol{e}_{2,c}\rangle = -\langle \mathcal{H}_c J \boldsymbol{g}_{2,c}, \boldsymbol{e}_{2,c}\rangle = \langle \boldsymbol{g}_{1,c}, \boldsymbol{e}_{2,c}\rangle = \langle \boldsymbol{\phi}_c, \partial_c \boldsymbol{\phi}_c\rangle = \mathscr{N}_c'$, $\langle \boldsymbol{g}_{1,c}, \boldsymbol{e}_{1,c}\rangle = -\langle \boldsymbol{\phi}_c, \partial_x \boldsymbol{\phi}_c\rangle = 0$, we may write the matrix $\mathcal{T}$ in the following form:

$$\mathcal{T}_c = \begin{bmatrix} 0 & \mathscr{N}_c' & \alpha_c \\ \mathscr{N}_c' & \frac{1}{2}(I_c')^2 & \beta_c \\ \alpha_c & \beta_c & \gamma_c \end{bmatrix}. \tag{2.37}$$

Note that $\mathcal{T}_{c_\star}$ is nondegenerate, because $\mathscr{N}_{c_\star}' = 0$ by the choice of $c_\star$, while $\alpha_{c_\star} = \langle \boldsymbol{g}_{1,c_\star}, \boldsymbol{e}_{3,c_\star}\rangle = \langle \boldsymbol{\phi}_{c_\star}, \boldsymbol{e}_{3,c_\star}\rangle = \frac{1}{2}(I_{c_\star}')^2 > 0$ by (2.9).

3. Center manifold reduction. We first discuss the existence of a solution $\boldsymbol{u}(t)$ that corresponds to perturbed initial data. We will rely on the well-posedness results due to T. Kato.

LEMMA 3.1. *For any $\mu > 0$, $s \geq 2$, and $\boldsymbol{u}_0 \in H^s(\mathbb{R}) \cap L^2_{2\mu}(\mathbb{R})$ with $\|\boldsymbol{u}_0\|_{H^1} < 2\|\boldsymbol{\phi}_{c_\star}\|_{H^1}$, there exists a function*

$$\boldsymbol{u}(t) \in C([0,\infty), H^s(\mathbb{R}) \cap L^2_{2\mu}(\mathbb{R})), \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0, \tag{3.1}$$

which solves (1.1) *for $0 \leq t < t_1$, where t_1 is finite or infinite, defined by*

$$t_1 = \sup\{t \geq 0 : \ \|\boldsymbol{u}(t)\|_{H^1} < 2\|\boldsymbol{\phi}_{c_\star}\|_{H^1}\}. \tag{3.2}$$

Proof. According to [Kat83, Theorem 10.1], equation (1.1) is globally well-posed in $H^s(\mathbb{R}) \cap L^2_{2\mu}(\mathbb{R})$ for any $s \geq 2$, $\mu > 0$ (for the initial data with arbitrarily large norm) if f satisfies

$$\lim_{|z|\to\infty} |z|^{-4} f'(z) \geq 0. \tag{3.3}$$

We modify the nonlinearity $f(z)$ for $|z| > 2\|\boldsymbol{\phi}_{c_\star}\|_{H^1}$ so that (3.3) is satisfied; let us call this modified nonlinearity $\tilde{f}(z)$. Thus, for any $\boldsymbol{u}_0 \in H^s(\mathbb{R}) \cap L^2_{2\mu}(\mathbb{R})$ with $\|\boldsymbol{u}_0\|_{H^1} < 2\|\boldsymbol{\phi}_{c_\star}\|_{H^1}$, there exists a function

$$\boldsymbol{u}(t) \in C([0,\infty), H^s(\mathbb{R}) \cap L^2_{2\mu}(\mathbb{R})), \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0, \tag{3.4}$$

that solves the equation with the modified nonlinearity

$$\partial_t \boldsymbol{u} = \partial_x\big(-\partial_x^2 \boldsymbol{u} + \tilde{f}(\boldsymbol{u})\big). \tag{3.5}$$

For $0 \leq t < t_1$, with t_1 defined by (3.2), one has $\|\boldsymbol{u}(t)\|_{L^\infty} \leq \|\boldsymbol{u}(t)\|_{H^1} < 2\|\boldsymbol{\phi}_{c_\star}\|_{H^1}$. Therefore, for $0 \leq t < t_1$, $\boldsymbol{u}(t)$ solves both (3.5) and (1.1) since $\tilde{f}(z) = f(z)$ for $|z| \leq 2\|\boldsymbol{\phi}_{c_\star}\|_{H^1}$. $\square$

We fix μ satisfying (2.10). For the initial data $\boldsymbol{u}_0 \in H^2(\mathbb{R}) \cap L^2_{2\mu}(\mathbb{R})$ with $\|\boldsymbol{u}_0\|_{H^1} < 2\|\boldsymbol{\phi}_{c_\star}\|_{H^1}$ there is a function $\boldsymbol{u} \in C([0,\infty), H^2(\mathbb{R}) \cap L^2_{2\mu}(\mathbb{R}))$ that solves (1.1) for $0 \leq t < t_1$, with t_1 from (3.2). We will approximate the solution $\boldsymbol{u}(x,t)$ by a traveling wave $\boldsymbol{\phi}_c$ moving with the variable speed $c = c(t)$. Thus, we decompose the solution $\boldsymbol{u}(x,t)$ into the traveling wave $\boldsymbol{\phi}_c(x)$ and the perturbation $\boldsymbol{\rho}(x,t)$ as follows:

$$\boldsymbol{u}(x,t) = \boldsymbol{\phi}_{c(t)}\left(x - \xi(t) - \int_0^t c(t')\,dt'\right) + \boldsymbol{\rho}\left(x - \xi(t) - \int_0^t c(t')\,dt',\, t\right). \tag{3.6}$$

The functions $\xi(t)$ and $c(t)$ are yet to be chosen.

Using (3.6), we rewrite the generalized KdV equation (1.1) as an equation on $\boldsymbol{\rho}$:

$$\dot{\boldsymbol{\rho}} - J\mathcal{H}_c\boldsymbol{\rho} = -\dot{\xi}\boldsymbol{e}_{1,c} - \dot{c}\boldsymbol{e}_{2,c} + \dot{\xi}\partial_x\boldsymbol{\rho} + J\boldsymbol{N}, \tag{3.7}$$

with $\mathcal{H}_c$ given by (1.11) and with $J\boldsymbol{N}$ given by

$$J\boldsymbol{N} = \partial_x\left[f(\boldsymbol{\phi}_c + \boldsymbol{\rho}) - f(\boldsymbol{\phi}_c) - \boldsymbol{\rho}f'(\boldsymbol{\phi}_c)\right], \tag{3.8}$$

where we changed coordinates, denoting $y = x - \xi(t) - \int_0^t c(t')\,dt'$ by x. By Proposition 2.2 (iii), the eigenvalue $\lambda = 0$ of operator $J\mathcal{H}_{c_\star}$ in $L^2_\mu(\mathbb{R})$ has algebraic multiplicity three. We decompose the perturbation $\boldsymbol{\rho}(x,t)$ as follows:

$$\boldsymbol{\rho}(x,t) = \zeta(t)\boldsymbol{e}_{3,c(t)}(x) + \boldsymbol{v}(x,t), \tag{3.9}$$

where $\boldsymbol{e}_{3,c}$ is constructed in Lemma 2.4. The inclusions $\boldsymbol{\phi}_c \in H^2(\mathbb{R})\cap L^2_{2\mu}(\mathbb{R}) \subset H^1_\mu(\mathbb{R})$ and $\boldsymbol{e}_{3,c} \in H^1_\mu(\mathbb{R})$ show that $\boldsymbol{v}(\cdot,t) \in H^1_\mu(\mathbb{R})$.

We would like to choose $\xi(t)$, $c(t) = c_\star + \eta(t)$, and $\zeta(t)$ so that

$$\boldsymbol{v}(x,t) = \boldsymbol{u}\left(x + \xi(t) + \int_0^t (c_\star + \eta(t'))\,dt',\, t\right) - \boldsymbol{\phi}_{c_\star+\eta(t)}(x) - \zeta(t)\boldsymbol{e}_{3,c_\star+\eta(t)}(x) \tag{3.10}$$

represents the part of the perturbation that corresponds to the continuous spectrum of $J\mathcal{H}_c$.

Proposition 3.2. *There exist $\eta_1 > 0$, $\zeta_1 > 0$, and $\delta_1 > 0$ such that if η_0 and ζ_0 satisfy*

$$|\eta_0| < \eta_1, \qquad |\zeta_0| < \zeta_1, \qquad \|\boldsymbol{\phi}_{c_\star+\eta_0} + \zeta_0\boldsymbol{e}_{3,c_\star+\eta_0} - \boldsymbol{\phi}_{c_\star}\|_{H^1} < \|\boldsymbol{\phi}_{c_\star}\|_{H^1}, \tag{3.11}$$

then there is $T_1 \in \mathbb{R}_+ \cup \{+\infty\}$ such that the following hold:

(i) *There exists $\boldsymbol{u} \in C([0,\infty), H^2(\mathbb{R}) \cap L^2_{2\mu}(\mathbb{R}))$ so that*

$$\boldsymbol{u}(0) = \boldsymbol{\phi}_{c_\star+\eta_0} + \zeta_0\boldsymbol{e}_{3,c_\star+\eta_0} \tag{3.12}$$

and $\boldsymbol{u}(t)$ solves (1.1) for $0 \le t < T_1$.

(ii) *There exist functions*

$$\xi, \eta, \zeta \in C([0,\infty)), \qquad \xi(0) = 0, \quad \eta(0) = \eta_0, \quad \zeta(0) = \zeta_0, \tag{3.13}$$

such that the function $\boldsymbol{v}(t)$ defined by (3.10) satisfies

$$e^{\mu x}\boldsymbol{v}(x,t) \in \ker P^\mu_{c_\star+\eta(t)}, \qquad 0 \le t < T_1. \tag{3.14}$$

(iii) *The following inequalities hold for $0 \le t < T_1$:*

$$\|\boldsymbol{u}(t)\|_{H^1} < 2\|\boldsymbol{\phi}_{c_\star}\|_{H^1}, \quad |\eta(t)| < \eta_1, \quad |\zeta(t)| < \zeta_1, \quad \|\boldsymbol{v}(t)\|_{H^1_\mu} < \delta_1. \tag{3.15}$$

(iv) *If one cannot choose $T_1 = \infty$, then at least one of the inequalities in (3.15) turns into an equality at $t = T_1$.*

Proof. Since $\boldsymbol{u}_0 = \boldsymbol{\phi}_{c_\star+\eta_0} + \zeta_0\boldsymbol{e}_{3,c_\star+\eta_0} \in H^2(\mathbb{R})\cap L^2_{2\mu}(\mathbb{R})$ and the conditions (3.11) are satisfied, by Lemma 3.1, there is a function $\boldsymbol{u}(t) \in C([0,\infty), H^2(\mathbb{R})\cap L^2_{2\mu}(\mathbb{R}))$ and $t_1 \in \mathbb{R}_+ \cup \{+\infty\}$ such that $\boldsymbol{u}(t)$ solves (1.1) for $0 \le t < t_1$ and, if $t_1 < \infty$, then

$\|\boldsymbol{u}(t_1)\|_{H^1} = 2\|\boldsymbol{\phi}_{c_\star}\|_{H^1}$. We thus need to construct $\xi(t)$, $\eta(t)$, and $\zeta(t)$ so that $\boldsymbol{v}(x,t)$ defined by (3.10) satisfies the constraints

$$\langle \boldsymbol{g}_{1,c_\star+\eta(t)}, \boldsymbol{v}(t)\rangle = \langle \boldsymbol{g}_{2,c_\star+\eta(t)}, \boldsymbol{v}(t)\rangle = \langle \boldsymbol{g}_{3,c_\star+\eta(t)}, \boldsymbol{v}(t)\rangle = 0. \tag{3.16}$$

Let us note that $\boldsymbol{v}(0) = 0$ by (3.10), (3.12), and (3.13). Since $J\mathcal{H}_c\boldsymbol{e}_{3,c} = \lambda_c\boldsymbol{e}_{3,c} + \boldsymbol{e}_{2,c}$,

$$\partial_t(\zeta\boldsymbol{e}_{3,c}) - J\mathcal{H}_c(\zeta\boldsymbol{e}_{3,c}) = \dot\zeta\boldsymbol{e}_{3,c} + \dot\eta\zeta\partial_c\boldsymbol{e}_{3,c} - \zeta(\lambda_c\boldsymbol{e}_{3,c} + \boldsymbol{e}_{2,c}). \tag{3.17}$$

Therefore, (3.7) can be written as the following equation on $\boldsymbol{v}(t) = \boldsymbol{\rho} - \zeta\boldsymbol{e}_{3,c}$:

$$\dot{\boldsymbol{v}} - J\mathcal{H}_c\boldsymbol{v} = -\dot\xi\boldsymbol{e}_{1,c} - (\dot\eta - \zeta)\,\boldsymbol{e}_{2,c} - (\dot\zeta - \lambda_c\zeta)\boldsymbol{e}_{3,c} - \dot\eta\zeta\partial_c\boldsymbol{e}_{3,c} + \dot\xi\partial_x\boldsymbol{\rho} + J\boldsymbol{N}. \tag{3.18}$$

Differentiating the constraints (3.16) and using the evolution equation (3.18), we derive the center manifold reduction,

$$\mathcal{T}_c\begin{bmatrix}\dot\xi\\ \dot\eta-\zeta\\ \dot\zeta-\lambda_c\zeta\end{bmatrix} - \dot\eta\begin{bmatrix}\langle\partial_c\boldsymbol{g}_{1,c},\boldsymbol{v}\rangle\\ \langle\partial_c\boldsymbol{g}_{2,c},\boldsymbol{v}\rangle\\ \langle\partial_c\boldsymbol{g}_{3,c},\boldsymbol{v}\rangle\end{bmatrix} = -\dot\eta\zeta\begin{bmatrix}\langle\boldsymbol{g}_{1,c},\partial_c\boldsymbol{e}_{3,c}\rangle\\ \langle\boldsymbol{g}_{2,c},\partial_c\boldsymbol{e}_{3,c}\rangle\\ \langle\boldsymbol{g}_{3,c},\partial_c\boldsymbol{e}_{3,c}\rangle\end{bmatrix} + \dot\xi\begin{bmatrix}\langle\boldsymbol{g}_{1,c},\partial_x\boldsymbol{\rho}\rangle\\ \langle\boldsymbol{g}_{2,c},\partial_x\boldsymbol{\rho}\rangle\\ \langle\boldsymbol{g}_{3,c},\partial_x\boldsymbol{\rho}\rangle\end{bmatrix} + \begin{bmatrix}\langle\boldsymbol{g}_{1,c},J\boldsymbol{N}\rangle\\ \langle\boldsymbol{g}_{2,c},J\boldsymbol{N}\rangle\\ \langle\boldsymbol{g}_{3,c},J\boldsymbol{N}\rangle\end{bmatrix}, \tag{3.19}$$

where the matrix $\mathcal{T}_c$ is given by (2.35). The above can be rewritten as

$$\mathcal{S}\begin{bmatrix}\dot\xi\\ \dot\eta-\zeta\\ \dot\zeta-\lambda_c\zeta\end{bmatrix} = \begin{bmatrix}-\zeta^2\langle\boldsymbol{g}_{1,c},\partial_c\boldsymbol{e}_{3,c}\rangle + \zeta\langle\partial_c\boldsymbol{g}_{1,c},\boldsymbol{v}\rangle + \langle\boldsymbol{g}_{1,c},J\boldsymbol{N}\rangle\\ -\zeta^2\langle\boldsymbol{g}_{2,c},\partial_c\boldsymbol{e}_{3,c}\rangle + \zeta\langle\partial_c\boldsymbol{g}_{2,c},\boldsymbol{v}\rangle + \langle\boldsymbol{g}_{2,c},J\boldsymbol{N}\rangle\\ -\zeta^2\langle\boldsymbol{g}_{3,c},\partial_c\boldsymbol{e}_{3,c}\rangle + \zeta\langle\partial_c\boldsymbol{g}_{3,c},\boldsymbol{v}\rangle + \langle\boldsymbol{g}_{3,c},J\boldsymbol{N}\rangle\end{bmatrix}, \tag{3.20}$$

where $c = c_\star + \eta$ and

$$\mathcal{S}(\eta,\zeta,\boldsymbol{v}) = \mathcal{T}_c + \begin{bmatrix}-\langle\boldsymbol{g}_{1,c},\partial_x(\zeta\boldsymbol{e}_{3,c}+\boldsymbol{v})\rangle & \zeta\langle\boldsymbol{g}_{1,c},\partial_c\boldsymbol{e}_{3,c}\rangle - \langle\partial_c\boldsymbol{g}_{1,c},\boldsymbol{v}\rangle & 0\\ -\langle\boldsymbol{g}_{2,c},\partial_x(\zeta\boldsymbol{e}_{3,c}+\boldsymbol{v})\rangle & \zeta\langle\boldsymbol{g}_{2,c},\partial_c\boldsymbol{e}_{3,c}\rangle - \langle\partial_c\boldsymbol{g}_{2,c},\boldsymbol{v}\rangle & 0\\ -\langle\boldsymbol{g}_{3,c},\partial_x(\zeta\boldsymbol{e}_{3,c}+\boldsymbol{v})\rangle & \zeta\langle\boldsymbol{g}_{3,c},\partial_c\boldsymbol{e}_{3,c}\rangle - \langle\partial_c\boldsymbol{g}_{3,c},\boldsymbol{v}\rangle & 0\end{bmatrix}. \tag{3.21}$$

Note that the matrix $\mathcal{S}(\eta,\zeta,\boldsymbol{v})$ depends continuously on $(\eta,\zeta,\boldsymbol{v}) \in \mathbb{R}^2 \times H^1_\mu(\mathbb{R})$. Since the matrix $\mathcal{T}_{c_\star}$ is nonsingular (see (2.37)), the matrix $\mathcal{S}(\eta,\zeta,\boldsymbol{v})$ is invertible for sufficiently small values of $|\eta|$, $|\zeta|$, and $\|\boldsymbol{v}\|_{H^1_\mu}$.

Thus, there exist $\eta_1 > 0$, $\zeta_1 > 0$, and $\delta_1 > 0$ so that the matrix $\mathcal{S}(\eta,\zeta,\boldsymbol{v})$ is invertible if

$$|\eta| \le 2\eta_1, \qquad |\zeta| \le 2\zeta_1, \qquad \|\boldsymbol{v}\|_{H^1_\mu} \le 2\delta_1. \tag{3.22}$$

For such η, ζ, and $\boldsymbol{v}$, we can write

$$\begin{bmatrix}\dot\xi\\ \dot\eta-\zeta\\ \dot\zeta-\lambda_c\zeta\end{bmatrix} = \begin{bmatrix}R_1(\eta,\zeta,\boldsymbol{v})\\ R_2(\eta,\zeta,\boldsymbol{v})\\ R_3(\eta,\zeta,\boldsymbol{v})\end{bmatrix}, \tag{3.23}$$

where the right-hand side is given by

$$\begin{bmatrix}R_1(\eta,\zeta,\boldsymbol{v})\\ R_2(\eta,\zeta,\boldsymbol{v})\\ R_3(\eta,\zeta,\boldsymbol{v})\end{bmatrix} = \mathcal{S}(\eta,\zeta,\boldsymbol{v})^{-1}\begin{bmatrix}-\zeta^2\langle\boldsymbol{g}_{1,c},\partial_c\boldsymbol{e}_{3,c}\rangle + \zeta\langle\partial_c\boldsymbol{g}_{1,c},\boldsymbol{v}\rangle + \langle\boldsymbol{g}_{1,c},J\boldsymbol{N}\rangle\\ -\zeta^2\langle\boldsymbol{g}_{2,c},\partial_c\boldsymbol{e}_{3,c}\rangle + \zeta\langle\partial_c\boldsymbol{g}_{2,c},\boldsymbol{v}\rangle + \langle\boldsymbol{g}_{2,c},J\boldsymbol{N}\rangle\\ -\zeta^2\langle\boldsymbol{g}_{3,c},\partial_c\boldsymbol{e}_{3,c}\rangle + \zeta\langle\partial_c\boldsymbol{g}_{3,c},\boldsymbol{v}\rangle + \langle\boldsymbol{g}_{3,c},J\boldsymbol{N}\rangle\end{bmatrix}. \tag{3.24}$$

Assume that η_0 and ζ_0 are such that the conditions (3.11) are satisfied. Let $\varrho_0 \in C^\infty_{comp}(\mathbb{R})$ be such that $0 \le \varrho_0(s) \le 1$, $\varrho_0(s) \equiv 1$ for $|s| \le 1$, and $\varrho_0(s) \equiv 0$ for $|s| \ge 2$. Define a continuous matrix-valued function $\tilde{\mathcal{S}} : \mathbb{R}^2 \times H^1_\mu \to GL(3)$ by

$$\tilde{\mathcal{S}}(\eta, \zeta, \boldsymbol{v}) = \mathcal{S}(\varrho\eta, \varrho\zeta, \varrho\boldsymbol{v}), \qquad \text{where} \quad \varrho = \varrho_0(\eta/\eta_1)\varrho_0(\zeta/\zeta_1)\varrho_0(\|\boldsymbol{v}\|_{H^1_\mu}/\delta_1).$$

This function coincides with $\mathcal{S}$ (defined in (3.21)) for $|\eta| < \eta_1$, $|\zeta| < \zeta_1$, and $\|\boldsymbol{v}\|_{H^1_\mu} < \delta_1$, and has uniformly bounded inverse. The system (3.23) with the right-hand side as in (3.24) but with $\tilde{\mathcal{S}}$ instead of $\mathcal{S}$, and with $\boldsymbol{v}$ given by the ansatz (3.10), defines differentiable functions $\xi(t)$, $\eta(t)$, and $\zeta(t)$ for all $t \ge 0$. Note that $\boldsymbol{v}(t)$ defined by (3.10) is a continuous function of time and is valued in $H^1_\mu(\mathbb{R})$, since so are $\boldsymbol{u}$, $\boldsymbol{\phi}_c$, and $\boldsymbol{e}_{3,c}$. Define $t_2 \in \mathbb{R}_+ \cup \{+\infty\}$ by

$$t_2 = \sup\{t \ge 0\colon |\eta(t)| < \eta_1,\ |\zeta(t)| < \zeta_1,\ \|\boldsymbol{v}(\cdot,t)\|_{H^1_\mu} < \delta_1\}. \tag{3.25}$$

For $t \in (0, t_2)$, the solution $(\xi(t), \eta(t), \zeta(t))$ also solves (3.23), since the inequalities $|\eta(t)| < \eta_1$, $|\zeta(t)| < \zeta_1$, and $\|\boldsymbol{v}(\cdot,t)\|_{H^1_\mu} < \delta_1$ ensure that $\tilde{\mathcal{S}}$ coincides with $\mathcal{S}$. Thus, Proposition 3.2 is proved with

$$T_1 = \min(t_1, t_2) \in \mathbb{R}_+ \cup \{+\infty\}, \tag{3.26}$$

where t_1, t_2 are from (3.2) and (3.25). ∎

4. Energy and dissipative estimates. We adapt the analysis from [PW94]. In this section, we formulate two lemmas that are the analogue of [PW94, Proposition 6.1]. Lemma 4.1 is based on the energy conservation and allows us to control $\|\boldsymbol{\rho}\|_{H^1}$ in terms of $\|\boldsymbol{v}\|_{H^1_\mu}$. Lemma 4.3 bounds $\|\boldsymbol{v}\|_{H^1_\mu}$ in terms of $\|\boldsymbol{\rho}\|_{H^1}$ and is based on dissipative estimates on the semigroup generated by A^μ_c (see Lemma 4.2).

Let $\eta_1 > 0$, $\zeta_1 > 0$, and $\delta_1 > 0$ not be larger than in Proposition 3.2, and assume that δ_1 satisfies

$$\delta_1 < \frac{\min(1, c_\star)}{4\sup_{|z|\le 2\|\boldsymbol{\phi}_{c_\star}\|_{H^1}} |f''(z)|}. \tag{4.1}$$

Let $\eta_0 > 0$ and ζ_0 be such that the conditions (3.11) are satisfied. According to Proposition 3.2, there exists $T_1 \in \mathbb{R}_+ \cup \{+\infty\}$ such that there is a solution $\boldsymbol{u} \in C((0, T_1), H^2(\mathbb{R}) \cap L^2_{2\mu}(\mathbb{R}))$ to (1.1) with the initial data

$$\boldsymbol{u}(0) = \boldsymbol{u}_0 := \boldsymbol{\phi}_{c_\star+\eta_0} + \zeta_0 \boldsymbol{e}_{3,c_\star+\eta_0},$$

and functions $\xi(t)$, $\eta(t)$, and $\zeta(t)$ and $\boldsymbol{v}(t)$ (given by (3.10)), defined for $0 \le t < T_1$, such that (3.14) and (3.15) are satisfied. For given η_0 and ζ_0, define the following function of η:

$$\mathscr{Y}(\eta) = \|\boldsymbol{\rho}_0\|_{H^1} + \|\boldsymbol{\rho}_0\|_{H^1}^{1/2}|\eta - \eta_0|^{1/2} + |\mathscr{N}_{c_\star+\eta} - \mathscr{N}_{c_\star+\eta_0}|^{1/2}, \tag{4.2}$$

where $\boldsymbol{\rho}_0 \equiv \zeta_0 \boldsymbol{e}_{3,c_\star+\eta_0}$.

Lemma 4.1. *There exists $C_1 > 0$ such that if at some moment $0 \le t < T_1$,*

$$\|\boldsymbol{\rho}(t)\|_{H^1} \le \delta_1,$$

then

$$\|\boldsymbol{\rho}(t)\|_{H^1} \le C_1 \left(\mathscr{Y}(\eta(t)) + |\zeta(t)| + \|\boldsymbol{v}(t)\|_{H^1_\mu}\right), \tag{4.3}$$

where $\mathscr{Y}(\eta)$ *is given by* (4.2).

Proof. Let us introduce the effective Hamiltonian $\mathscr{L}_c$:

$$\mathscr{L}_c(\boldsymbol{u}) = E(\boldsymbol{u}) + c\mathscr{N}(\boldsymbol{u}), \quad \mathscr{L}_c'(\boldsymbol{\phi}_c) = E'(\boldsymbol{\phi}_c) + c\mathscr{N}'(\boldsymbol{\phi}_c) = 0, \quad \mathscr{L}_c''(\boldsymbol{\phi}_c) = \mathcal{H}_c, \tag{4.4}$$

where E and $\mathscr{N}$ are the energy and momentum functionals defined in (1.5) and (1.7). Using the Taylor series expansion for $\mathscr{L}_c$ at $\boldsymbol{\phi}_c$, we have

$$\begin{aligned} \mathscr{L}_c(\boldsymbol{u}(t)) &= \mathscr{L}_c(\boldsymbol{\phi}_c) + \frac{1}{2}\langle \boldsymbol{\rho}, \mathcal{H}_c\boldsymbol{\rho}\rangle + \int_{\mathbb{R}} g(\boldsymbol{\phi}_c, \boldsymbol{\rho})\boldsymbol{\rho}^3\,dx \\ &= \mathscr{L}_c(\boldsymbol{\phi}_c) + \frac{1}{2}\langle \boldsymbol{\rho}, (-\partial_x^2 + c)\boldsymbol{\rho}\rangle + \frac{1}{2}\langle \boldsymbol{\rho}, f'(\boldsymbol{\phi}_c)\boldsymbol{\rho}\rangle + \int_{\mathbb{R}} g(\boldsymbol{\phi}_c, \boldsymbol{\rho})\boldsymbol{\rho}^3\,dx, \end{aligned} \tag{4.5}$$

where

$$g(\boldsymbol{\phi}_c, \boldsymbol{\rho}) = \frac{1}{2}\int_0^1 (1-s)^2 f''(\boldsymbol{\phi}_c + s\boldsymbol{\rho})\,ds. \tag{4.6}$$

For the second term in (4.5), there is the following bound from below:

$$\frac{1}{2}\int_{\mathbb{R}} \left((\partial_x\boldsymbol{\rho})^2 + c\boldsymbol{\rho}^2\right)dx \geq m\|\boldsymbol{\rho}\|_{H^1}^2, \qquad m = \frac{1}{2}\min(1, c_\star) > 0. \tag{4.7}$$

The bound for the third term in the right-hand side of (4.5) follows from the inequalities

$$\int_{\mathbb{R}} |f'(\boldsymbol{\phi}_c)|\boldsymbol{\rho}^2\,dx \leq \|e^{-2\mu x} f'(\boldsymbol{\phi}_c)\|_{L^\infty}\|\boldsymbol{\rho}\|_{L^2_\mu}^2 \leq b\left[|\zeta|\|\boldsymbol{e}_{3,c}\|_{L^2_\mu} + \|\boldsymbol{v}(t)\|_{L^2_\mu}\right]^2, \tag{4.8}$$

where $b = \sup_{c\in[c_\star, c_\star+\eta_1]}\|e^{-2\mu x} f'(\boldsymbol{\phi}_c)\|_{L^\infty} < \infty$ due to (2.10), the assumption (1.2) that $f'(0) = 0$, and due to Lemma A.1 from Appendix A. We bound the last term in (4.5) by

$$\int_{\mathbb{R}} |g(\boldsymbol{\phi}_c, \boldsymbol{\rho})\boldsymbol{\rho}^3|\,dx \leq \|g(\boldsymbol{\phi}_c, \boldsymbol{\rho})\|_{L^\infty}\|\boldsymbol{\rho}\|_{H^1}^3 \leq \delta_1\|g(\boldsymbol{\phi}_c, \boldsymbol{\rho})\|_{L^\infty}\|\boldsymbol{\rho}\|_{H^1}^2. \tag{4.9}$$

According to (4.1), g from (4.6) satisfies $\delta_1\|g(\boldsymbol{\phi}_c, \boldsymbol{\rho})\|_{L^\infty} < \frac{\min(1,c_\star)}{4} = \frac{m}{2}$, and this leads to

$$\int_{\mathbb{R}} |g(\boldsymbol{\phi}_c, \boldsymbol{\rho})\boldsymbol{\rho}^3|\,dx \leq \frac{m}{2}\|\boldsymbol{\rho}\|_{H^1}^2. \tag{4.10}$$

Combining (4.5) with the bounds (4.7), (4.8), and (4.10), we obtain

$$\frac{m}{2}\|\boldsymbol{\rho}\|_{H^1}^2 \leq |\mathscr{L}_c(\boldsymbol{u}) - \mathscr{L}_c(\boldsymbol{\phi}_c)| + \frac{b}{2}\left[|\zeta|\|\boldsymbol{e}_{3,c}\|_{L^2_\mu} + \|\boldsymbol{v}\|_{H^1_\mu}\right]^2,$$

so that, for some $C > 0$,

$$\|\boldsymbol{\rho}\|_{H^1} \leq C\left[|\mathscr{L}_c(\boldsymbol{u}) - \mathscr{L}_c(\boldsymbol{\phi}_c)|^{1/2} + |\zeta| + \|\boldsymbol{v}\|_{H^1_\mu}\right]. \tag{4.11}$$

Now let us estimate $|\mathscr{L}_c(\boldsymbol{u}(t)) - \mathscr{L}_c(\boldsymbol{\phi}_c)|$. Note that $\mathscr{L}_c(\boldsymbol{u}(t)) = \mathscr{L}_c(\boldsymbol{u}_0)$ since the value of the energy functional E given by (1.5) and the value of the momentum functional $\mathscr{N}$ given by (1.7) are conserved along the trajectories of (1.1). Thus, we can write

$$|\mathscr{L}_c(\boldsymbol{u}(t)) - \mathscr{L}_c(\boldsymbol{\phi}_c)| \le |\mathscr{L}_c(\boldsymbol{u}_0) - \mathscr{L}_c(\boldsymbol{\phi}_{c_0})| + |\mathscr{L}_c(\boldsymbol{\phi}_c) - \mathscr{L}_c(\boldsymbol{\phi}_{c_0})|. \tag{4.12}$$

Using the definition (4.4) of the functional $\mathscr{L}_c$, we express the first term in the right-hand side of (4.12) as

$$\mathscr{L}_c(\boldsymbol{u}_0) - \mathscr{L}_c(\boldsymbol{\phi}_{c_0}) = \mathscr{L}_{c_0}(\boldsymbol{u}_0) - \mathscr{L}_{c_0}(\boldsymbol{\phi}_{c_0}) + (\eta - \eta_0)(\mathscr{N}(\boldsymbol{u}_0) - \mathscr{N}(\boldsymbol{\phi}_{c_0})). \tag{4.13}$$

Since $\mathscr{L}'_{c_0}(\boldsymbol{\phi}_{c_0}) = 0$, there exists $k > 0$ such that $|\mathscr{L}_{c_0}(\boldsymbol{u}_0) - \mathscr{L}_{c_0}(\boldsymbol{\phi}_{c_0})| \le k\|\boldsymbol{\rho}_0\|^2_{H^1}$, where $\boldsymbol{\rho}_0 = \boldsymbol{u}_0 - \boldsymbol{\phi}_{c_0}$; this allows us to bound (4.13) by

$$|\mathscr{L}_c(\boldsymbol{u}_0) - \mathscr{L}_c(\boldsymbol{\phi}_{c_0})| \le \mathrm{const}(\|\boldsymbol{\rho}_0\|^2_{H^1} + |\eta - \eta_0|\|\boldsymbol{\rho}_0\|_{H^1}). \tag{4.14}$$

For the second term in the right-hand side of (4.12), we have

$$|\mathscr{L}_c(\boldsymbol{\phi}_c) - \mathscr{L}_c(\boldsymbol{\phi}_{c_0})| \le |E_c - E_{c_0}| + c|\mathscr{N}_c - \mathscr{N}_{c_0}|.$$

From the relation

$$\frac{d}{dc}E_c = -c\frac{d}{dc}\mathscr{N}_c$$

we conclude that $|E_c - E_{c_0}| \le \max(c, c_0)|\mathscr{N}_c - \mathscr{N}_{c_0}|$, since $\mathscr{N}_c'$ is sign-definite for $c_\star < c \le c_\star + \eta_1$ by (1.16). Therefore, there is the following bound for the second term in the right-hand side of (4.12):

$$|\mathscr{L}_c(\boldsymbol{\phi}_c) - \mathscr{L}_c(\boldsymbol{\phi}_{c_0})| \le 2\max(c, c_0)|\mathscr{N}_c - \mathscr{N}_{c_0}|. \tag{4.15}$$

Using the bounds (4.14) and (4.15) in (4.12), we obtain

$$|\mathscr{L}_c(\boldsymbol{u}(t)) - \mathscr{L}_c(\boldsymbol{\phi}_c)| \le \mathrm{const}\left(\|\boldsymbol{\rho}_0\|^2_{H^1} + |\eta - \eta_0|\|\boldsymbol{\rho}_0\|_{H^1} + |\mathscr{N}_c - \mathscr{N}_{c_0}|\right).$$

Substituting this result into (4.11), we obtain the bound (4.3). ◻

LEMMA 4.2 (see [PW94]). *Let Assumption* 3 *be satisfied, and pick* $\mu \in (0, \sqrt{c/3})$. *Let* $Q_c^\mu = I - P_c^\mu$, *where* P_c^μ *introduced in* (2.20) *is the spectral projection that corresponds to the discrete spectrum of* A_c^μ *(see Remark* 2.6*). Then* A_c^μ *is the generator of a strongly continuous linear semigroup on* $H^s(\mathbb{R})$ *for any real* s, *and there exist constants* $a > 0$ *and* $b > 0$ *such that for all* $\boldsymbol{v} \in L^2(\mathbb{R})$ *and* $t > 0$ *the following estimate is satisfied:*

$$\|e^{A_c^\mu t} Q_c^\mu \boldsymbol{v}\|_{H^1} \le a t^{-1/2} e^{-bt} \|\boldsymbol{v}\|_{L^2}. \tag{4.16}$$

We require that η_1 be small enough, so that

$$\eta_1 \sup_{c \in [c_\star, c_\star + \eta_1]} \|\partial_c Q_c^\mu\|_{H^1 \to H^1} \le \frac{1}{2}. \tag{4.17}$$

LEMMA 4.3. *There exists* $C_2 > 0$ *such that if*

$$\eta_1 + \zeta_1 + \delta_1 < C_2 \tag{4.18}$$

and

$$\sup_{s\in[0,t]}|\eta(s)|\le\eta_1,\quad \sup_{s\in[0,t]}|\zeta(s)|\le\zeta_1,\quad \sup_{s\in[0,t]}\|\boldsymbol{\rho}(s)\|_{H^1}\le\delta_1,\quad \sup_{s\in[0,t]}\|\boldsymbol{v}(s)\|_{H^1_\mu}\le\delta_1, \tag{4.19}$$

then

$$\|\boldsymbol{v}(t)\|_{H^1_\mu}\le C_2\sup_{s\in[0,t]}\left[\zeta^2(s)+|\zeta(s)|\|\boldsymbol{\rho}(s)\|_{H^1}\right]. \tag{4.20}$$

Proof. Using the center manifold reduction (3.23), we rewrite the evolution equation (3.18) in the following form:

$$\dot{\boldsymbol{v}}-J\mathcal{H}_c\boldsymbol{v}=-\sum_{j=1}^{3}R_j\boldsymbol{e}_{j,c}-\zeta(\zeta+R_2)\partial_c\boldsymbol{e}_{3,c}+R_1\partial_x(\zeta\boldsymbol{e}_{3,c}+\boldsymbol{v})+J\boldsymbol{N}, \tag{4.21}$$

where $c=c(t)=c_\star+\eta(t)$, $\zeta=\zeta(t)$, and the nonlinear terms $R_j(t)$ are given by (3.24). We set

$$\boldsymbol{\omega}(x,t)=e^{\mu x}\boldsymbol{v}(x,t),\qquad \boldsymbol{e}^\mu_{j,c}(x)=e^{\mu x}\boldsymbol{e}_{j,c}(x),\qquad c\in[c_\star,c_\star+\eta_1],\quad j=1,2,3,$$

and consider A^μ_c given by (1.13). Equation (4.21) takes the following form:

$$\dot{\boldsymbol{\omega}}-A^\mu_c\boldsymbol{\omega}=\boldsymbol{G}, \tag{4.22}$$

where

$$\boldsymbol{G}(x,t)=-\sum_{j=1}^{3}R_j\boldsymbol{e}^\mu_{j,c}-\zeta(\zeta+R_2)\partial_c\boldsymbol{e}^\mu_{3,c}+R_1(\partial_x-\mu)(\zeta\boldsymbol{e}^\mu_{3,c}+\boldsymbol{\omega})+e^{\mu x}J\boldsymbol{N}. \tag{4.23}$$

As follows from (4.22),

$$\partial_t(Q^\mu_{c_\star}\boldsymbol{\omega})=Q^\mu_{c_\star}\dot{\boldsymbol{\omega}}=Q^\mu_{c_\star}\left(A^\mu_c\boldsymbol{\omega}+\boldsymbol{G}\right)=A^\mu_{c_\star}Q^\mu_{c_\star}\boldsymbol{\omega}+Q^\mu_{c_\star}(A^\mu_c-A^\mu_{c_\star})\boldsymbol{\omega}+Q^\mu_{c_\star}\boldsymbol{G}.$$

We may write $Q^\mu_{c_\star}\boldsymbol{\omega}$ as follows:

$$Q^\mu_{c_\star}\boldsymbol{\omega}(t)=\int_0^t e^{A^\mu_{c_\star}(t-s)}\mathfrak{G}(s)\,ds, \tag{4.24}$$

where

$$\mathfrak{G}(x,t)=Q^\mu_{c_\star}(A^\mu_c-A^\mu_{c_\star})\boldsymbol{\omega}(x,t)+Q^\mu_{c_\star}\boldsymbol{G}(x,t). \tag{4.25}$$

Using the dissipative estimate given by (4.16), we get

$$\|Q^\mu_{c_\star}\boldsymbol{\omega}(t)\|_{H^1}\le C\int_0^t(t-s)^{-1/2}e^{-b(t-s)}\|\mathfrak{G}(s)\|_{L^2}\,ds \tag{4.26}$$

$$\le Ce^{-bt/2}\sup_{s\in[0,t]}e^{bs/2}\|\mathfrak{G}(s)\|_{L^2}\int_0^t(t-s)^{-1/2}e^{-b(t-s)/2}\,ds \tag{4.27}$$

$$\le C\sup_{s\in[0,t]}e^{bs/2}\|\mathfrak{G}(s)\|_{L^2}. \tag{4.28}$$

Since $\boldsymbol{\omega} = Q^\mu_c \boldsymbol{\omega} = Q^\mu_{c_\star}\boldsymbol{\omega} + (Q^\mu_c - Q^\mu_{c_\star})\boldsymbol{\omega}$, we have

$$\|\boldsymbol{\omega}\|_{H^1} \le \|Q^\mu_{c_\star}\boldsymbol{\omega}\|_{H^1} + |\eta| \sup_{c\in[c_\star, c_\star+\eta_1]} \|\partial_c Q^\mu_c\|_{H^1\to H^1}\|\boldsymbol{\omega}\|_{H^1} \le \|Q^\mu_{c_\star}\boldsymbol{\omega}\|_{H^1} + \frac{1}{2}\|\boldsymbol{\omega}\|_{H^1},$$

where we used (4.17). It follows that $\|\boldsymbol{\omega}\|_{H^1} \le 2\|Q^\mu_{c_\star}\boldsymbol{\omega}\|_{H^1}$. Hence, we have

$$\|\boldsymbol{\omega}(t)\|_{H^1} \le Ce^{-bt/2} \sup_{s\in[0,t]} e^{bs/2}\|\mathfrak{G}(s)\|_{L^2}. \tag{4.29}$$

We now need the bound on $\|\mathfrak{G}\|_{L^2}$. We start with

$$\|\mathfrak{G}\|_{L^2} \le \|Q^\mu_{c_\star}(A^\mu_c - A^\mu_{c_\star})\boldsymbol{\omega}\|_{L^2} + \|Q^\mu_{c_\star}\boldsymbol{G}\|_{L^2}. \tag{4.30}$$

We estimate the first term in the right-hand side of (4.30) as follows:

$$\|Q^\mu_{c_\star}(A^\mu_c - A^\mu_{c_\star})\boldsymbol{\omega}(t)\|_{L^2} \le \|Q^\mu_{c_\star}(A^\mu_c - A^\mu_{c_\star})\|_{H^1\text{-}L^2}\|\boldsymbol{\omega}(t)\|_{H^1} \le C|\eta|\|\boldsymbol{\omega}(t)\|_{H^1}. \tag{4.31}$$

Since $\boldsymbol{e}^\mu_{j,c}$, $1\le j\le 3$, depend continuously on c while $Q^\mu_{c_\star}\boldsymbol{e}^\mu_{j,c_\star} = 0$, there are bounds $\|Q^\mu_{c_\star}\boldsymbol{e}^\mu_{j,c}\|_{H^1} \le C|\eta|$. This allows us to derive the following bound for the second term in the right-hand side of (4.30):

$$\begin{aligned}\|Q^\mu_{c_\star}\boldsymbol{G}\|_{L^2} &\le C\left(|\eta| \sup_{1\le j\le 3}|R_j| + |\zeta||\zeta+R_2| + |R_1|(|\zeta| + \|\boldsymbol{\omega}\|_{H^1}) + \|J\boldsymbol{N}\|_{L^2_\mu}\right)\\ &\le C\left(\zeta^2 + (|\eta| + |\zeta| + \|\boldsymbol{\omega}\|_{H^1})\sup_{1\le j\le 3}|R_j| + \|J\boldsymbol{N}\|_{L^2_\mu}\right).\end{aligned}$$

Using the representation (3.24) and the inclusions $\partial_c\boldsymbol{e}_{3,c} \in H^\infty_\mu(\mathbb{R})$, $\boldsymbol{g}_i \in H^\infty_{-\mu}(\mathbb{R})$, $\partial_c\boldsymbol{g}_i \in H^\infty_{-\mu}(\mathbb{R})$, we obtain the following estimates on R_j:

$$|R_j(\eta,\zeta,\boldsymbol{v})| \le C\left(\zeta^2 + |\zeta|\|\boldsymbol{v}\|_{H^1_\mu} + \|J\boldsymbol{N}\|_{L^2_\mu}\right), \qquad j = 1,2,3. \tag{4.32}$$

Taking into account (4.32), we get

$$\begin{aligned}\|Q^\mu_{c_\star}\boldsymbol{G}\|_{L^2_\mu} &\le C\left(\zeta^2 + (|\eta| + |\zeta| + \|\boldsymbol{\omega}\|_{H^1})(\zeta^2 + |\zeta|\|\boldsymbol{\omega}\|_{H^1} + \|J\boldsymbol{N}\|_{L^2_\mu}) + \|J\boldsymbol{N}\|_{L^2_\mu}\right)\\ &\le C\left(\zeta^2 + (|\eta| + \|\boldsymbol{\omega}\|_{H^1})|\zeta|\|\boldsymbol{\omega}\|_{H^1} + \|J\boldsymbol{N}\|_{L^2_\mu}\right).\end{aligned} \tag{4.33}$$

In the last inequality, we used the uniform boundedness of $|\eta|$, $|\zeta|$, and $\|\boldsymbol{\omega}\|_{H^1}$ that follows from (4.19).

Summing up (4.31) and (4.33), we obtain the following bound on $\|\mathfrak{G}\|_{L^2_\mu}$:

$$\|\mathfrak{G}\|_{L^2_\mu} \le C\left[\zeta^2 + (|\eta| + |\zeta|)\|\boldsymbol{\omega}\|_{H^1} + \|J\boldsymbol{N}\|_{L^2_\mu}\right]. \tag{4.34}$$

Using the integral representation for the nonlinearity (3.8),

$$J\boldsymbol{N} = \partial_x[f(\boldsymbol{\phi}_c + \boldsymbol{\rho}) - f(\boldsymbol{\phi}_c) - f'(\boldsymbol{\phi}_c)\boldsymbol{\rho}] = \partial_x\left[\frac{\boldsymbol{\rho}^2}{2}\int_0^1 (1-s)^2 f''(\boldsymbol{\phi}_c + s\boldsymbol{\rho})\,ds\right], \tag{4.35}$$

we obtain the bound

$$\|J\boldsymbol{N}\|_{L^2_\mu} \le C\|\boldsymbol{\rho}\|_{H^1_\mu}\|\boldsymbol{\rho}\|_{H^1} \le C\left(|\zeta|\|\boldsymbol{e}_{3,c}\|_{H^1_\mu} + \|\boldsymbol{v}\|_{H^1_\mu}\right)\|\boldsymbol{\rho}\|_{H^1},$$

with the constant C that depends on $\|\boldsymbol{\phi}_c\|_{H^1}$ and on the bounds on $f''(z)$ and $f'''(z)$ for $|z| \leq \|\boldsymbol{u}\|_{L^\infty}$, which is bounded by $2\|\boldsymbol{\phi}_{c_\star}\|_{H^1}$. This bound allows us to rewrite (4.34) as

(4.36)
$$\|\mathfrak{G}\|_{L^2_\mu} \leq C\left[\zeta^2 + (|\eta| + |\zeta| + \|\boldsymbol{\rho}\|_{H^1})\|\boldsymbol{\omega}\|_{H^1} + |\zeta|\|\boldsymbol{\rho}\|_{H^1}\right] \leq C\left[g_0 + g_1\|\boldsymbol{\omega}\|_{H^1}\right],$$

where

$$g_0(t) = \zeta^2(t) + |\zeta(t)|\|\boldsymbol{\rho}(t)\|_{H^1}, \qquad g_1(t) = |\eta(t)| + |\zeta(t)| + \|\boldsymbol{\rho}(t)\|_{H^1}. \tag{4.37}$$

Thus, (4.29) could be written as

$$2e^{bt/2}\|\boldsymbol{\omega}(t)\|_{H^1} \leq C_2 \sup_{s\in[0,t]} e^{bs/2}\left[g_0(s) + g_1(s)\|\boldsymbol{\omega}(s)\|_{H^1}\right] \tag{4.38}$$

for some $C_2 > 0$. Since the right-hand side is monotonically increasing with t, we also have

$$\sup_{s\in[0,t]} 2e^{bs/2}\|\boldsymbol{\omega}(s)\|_{H^1} \leq C_2 \sup_{s\in[0,t]} e^{bs/2}\left[g_0(s) + g_1(s)\|\boldsymbol{\omega}(s)\|_{H^1}\right]. \tag{4.39}$$

The function g_1 from (4.37) satisfies $C_2 \sup_{s\in[0,t]} g_1(s) < 1$ (this follows from the assumptions (4.18) and (4.19)), and therefore

$$\|\boldsymbol{\omega}(t)\|_{H^1} \leq C_2 e^{-bt/2} \sup_{s\in[0,t]} e^{bs/2} g_0(s) \leq C_2 \sup_{s\in[0,t]} \left[\zeta^2(s) + |\zeta(s)|\|\boldsymbol{\rho}(s)\|_{H^1}\right].$$

Since $\boldsymbol{\omega} = e^{\mu x}\boldsymbol{v}$, the last inequality yields (4.20). ◻

5. Nonlinear estimates. Now we close the estimates using the bounds on $\|\boldsymbol{\rho}\|_{H^1}$ (Lemma 4.1) and on $\|\boldsymbol{v}\|_{H^1_\mu}$ (Lemma 4.3) from the previous section.

We assume that $\eta_1 > 0$, $\zeta_1 > 0$, and $\delta_1 > 0$ are sufficiently small: not larger than in Proposition 3.2, satisfy the bounds (4.1), (4.17), and (4.18), and also that ζ_1 satisfies

$$\zeta_1 < \frac{1}{3\max(1, C_1)C_2}. \tag{5.1}$$

Define

$$C_3 = 2C_1, \qquad C_4 = 2C_2\max(1, C_3), \tag{5.2}$$

with C_1 and C_2 as in Lemmas 4.1 and 4.3. Choosing smaller values of η_1 and ζ_1 if necessary, we may assume that

$$C_3\left(\zeta_1 + 2\eta_1 + (\mathscr{N}_{c_\star+\eta_1} - \mathscr{N}_{c_\star})^{1/2}\right) < \delta_1, \tag{5.3}$$

$$C_4\left(\zeta_1^2 + 2\eta_1\zeta_1 + \zeta_1(\mathscr{N}_{c_\star+\eta_1} - \mathscr{N}_{c_\star})^{1/2}\right) < \delta_1. \tag{5.4}$$

Define

$$\eta_M(t) = \sup_{0\leq s\leq t} \eta(s), \tag{5.5}$$

$$\zeta_M(t) = \sup_{0\leq s\leq t} |\zeta(s)|. \tag{5.6}$$

PROPOSITION 5.1. *Assume that the initial data $\eta_0 > 0$ and ζ_0 are such that the following inequalities are satisfied:*

$$\eta_0 \in (0, \eta_1), \qquad |\zeta_0| < \zeta_1, \qquad \|\boldsymbol{\rho}_0\|_{H^1} < \min(\eta_1, \delta_1). \tag{5.7}$$

Then for $0 \le t < T_1$ the functions $\boldsymbol{\rho}(t)$, $\boldsymbol{v}(t)$ satisfy the bounds

$$\|\boldsymbol{\rho}(t)\|_{H^1} \le C_3 \left[\zeta_M(t) + \mathscr{Y}(\eta_M(t))\right], \tag{5.8}$$

$$\|\boldsymbol{v}(t)\|_{H^1_\mu} \le C_4 \left[\zeta_M(t)^2 + \zeta_M(t)\mathscr{Y}(\eta_M(t))\right], \tag{5.9}$$

where C_3, C_4 are defined by (5.2), *$\mathscr{Y}(\eta) = \|\boldsymbol{\rho}_0\|_{H^1} + \|\boldsymbol{\rho}_0\|_{H^1}^{1/2}|\eta - \eta_0|^{1/2} + |\mathscr{N}_{c_\star+\eta} - \mathscr{N}_{c_\star+\eta_0}|^{1/2}$ is introduced in* (4.2), *and η_M, ζ_M are defined in* (5.5) *and* (5.6).

Proof. Let

$$S = \{t \in [0, T_1)\colon \|\boldsymbol{\rho}(t)\|_{H^1} < \delta_1\}.$$

S is nonempty since $\|\boldsymbol{\rho}(0)\|_{H^1} < \delta_1$ by (5.7). According to Proposition 3.2 and representation (3.6), $\|\boldsymbol{\rho}(t)\|_{H^1}$ is a continuous function of t. Since the inequality in the definition of S is sharp, S is an open subset of $[0, T_1)$. Let us assume that $T_2 \in (0, T_1)$ is such that

$$\|\boldsymbol{\rho}(t)\|_{H^1} < \delta_1, \qquad 0 \le t < T_2. \tag{5.10}$$

It is enough to prove that $T_2 \in S$ (then the connected subset of S that contains $t = 0$ is both open and closed in $[0, T_1)$ and hence coincides with $[0, T_1)$). Since $\|\boldsymbol{v}(t)\|_{H^1_\mu} < \delta_1$ for $0 \le t < T_1$, Lemmas 4.1 and 4.3 are both applicable for $t \le T_2$. The estimate (4.3) on $\|\boldsymbol{\rho}(t)\|_{H^1}$ together with the estimate (4.20) on $\|\boldsymbol{v}(t)\|_{H^1_\mu}$ give

$$\begin{aligned}\|\boldsymbol{\rho}(t)\|_{H^1} &\le C_1\Big(\mathscr{Y}(\eta(t)) + |\zeta(t)| + \|\boldsymbol{v}(t)\|_{H^1_\mu}\Big)\\ &\le C_1\Big(\mathscr{Y}(t) + |\zeta(t)| + C_2 \sup_{s\in[0,t]}\left[\zeta^2 + |\zeta|\|\boldsymbol{\rho}\|_{H^1}\right]\Big).\end{aligned}$$

For $0 \le t \le T_2$, define $M(t) = \sup_{s\in[0,t]} \|\boldsymbol{\rho}(s)\|_{H^1}$. We have

$$M(t) \le C_1\Big(\sup_{s\in[0,t]}(\mathscr{Y}(\eta(s)) + |\zeta(s)|) + C_2 \sup_{s\in[0,t]}\left[\zeta^2(s) + |\zeta(s)|M(t)\right]\Big).$$

We carry the term $C_1C_2|\zeta|M(t)$ to the left-hand side of the inequality, taking into account that $C_1C_2|\zeta(t)| \le C_1C_2\zeta_1 \le \frac{1}{3}$ for all $0 \le t < T_1$ by (5.1). This results in the following relation:

$$\|\boldsymbol{\rho}(t)\|_{H^1} \le M(t) \le \frac{3}{2}C_1\Big(\sup_{s\in[0,t]}(\mathscr{Y}(\eta(s)) + |\zeta(s)|) + C_2\sup_{s\in[0,t]}\zeta^2(s)\Big).$$

Since $C_2\zeta^2 \le C_2\zeta_1|\zeta| \le |\zeta|/3$ by (5.1), we obtain

$$\|\boldsymbol{\rho}(t)\|_{H^1} \le \frac{3}{2}C_1 \sup_{s\in[0,t]}\left(\mathscr{Y}(\eta(s)) + \frac{4}{3}|\zeta(s)|\right) \le C_3 \sup_{s\in[0,t]}\left(\mathscr{Y}(\eta(s)) + |\zeta(s)|\right), \quad t \in [0, T_2],$$

with $C_3 = 2C_1$. This proves (5.8) for $t \in [0, T_2]$. It then follows that

$$\|\boldsymbol{\rho}(T_2)\|_{H^1} \le C_3\left[\zeta_1 + \mathscr{Y}(\eta_1)\right] \le C_3\left[\zeta_1 + 2\eta_1 + \left(\mathscr{N}_{c_\star+\eta_1} - \mathscr{N}_{c_\star}\right)^{1/2}\right] < \delta_1,$$

where we took into account the definition of $\mathscr{Y}(\eta)$ in (4.2), the bound $\|\boldsymbol{\rho}_0\|_{H^1} < \eta_1$ from (5.7), and inequality (5.3). Hence, $T_2 \in S$. It follows that S coincides with $[0, T_1)$.

Using the bound (5.8) in (4.20) and recalling the definition of C_4 in (5.2), we derive the bound (5.9) on $\|\boldsymbol{v}(t)\|_{H^1_\mu}$. ◻

COROLLARY 5.2. *Assume that the conditions of Proposition* 5.1 *are satisfied. If* $\eta_1 > 0$ *and* $\zeta_1 > 0$ *were chosen sufficiently small, then there exists a constant* $C_5 > 0$ *so that for* $0 \le t < T_1$ *the function* $\boldsymbol{v}(t)$ *satisfies the bound*

$$\|\boldsymbol{v}(t)\|_{H^1_{\mu/2}} \le C_5 \left[\zeta_M^2(t) + \zeta_M(t)\mathscr{Y}(\eta_M(t))\right], \tag{5.11}$$

where η_M, ζ_M *are defined in* (5.5), (5.6).

Proof. The bound (5.11) is proved in the same way as (5.9). We may need to take smaller values of η_1 and ζ_1 so that Lemmas 4.1 and 4.3 become applicable for the new exponential weight. Note that the exponential weight does not enter the definition (4.2) of the function $\mathscr{Y}(\eta)$. ◻

LEMMA 5.3. *Assume that the bounds* (5.9) *and* (5.11) *are satisfied for* $0 \le t < T_1$. *Then there exists* $C_6 > 0$ *so that the terms* R_2 *and* R_3 *defined in* (3.24) *satisfy for* $0 \le t < T_1$ *the bounds*

$$|R_j(\eta, \zeta, \boldsymbol{v})| \le C_6 \zeta_M^2, \qquad j = 2, 3. \tag{5.12}$$

Proof. By (4.32),

$$|R_j(\eta, \zeta, \boldsymbol{v})| \le C\left(\zeta^2 + |\zeta|\|\boldsymbol{v}\|_{H^1_\mu} + \|J\boldsymbol{N}\|_{L^2_\mu}\right), \qquad j = 2, 3. \tag{5.13}$$

According to (5.9), the second term in the right-hand side of (5.13) is bounded by $C\zeta^2$ as long as $\eta \in (0, \eta_1)$ and $|\zeta| \le \zeta_1$. We now need a bound on $\|J\boldsymbol{N}\|_{L^2_\mu}$. Using the representation (4.35) for the nonlinearity, we obtain the bounds

$$\|J\boldsymbol{N}\|_{L^2_\mu} \le C\|\boldsymbol{\rho}\|^2_{H^1_{\mu/2}} \le C\left(\zeta^2\|\boldsymbol{e}_{3,c}\|^2_{H^1_{\mu/2}} + \|\boldsymbol{v}\|^2_{H^1_{\mu/2}}\right). \tag{5.14}$$

The constant depends on $\|\boldsymbol{\phi}_c\|_{H^1}$ and on the bounds on $f''(z)$ and $f'''(z)$ for $|z| \le \|\boldsymbol{u}\|_{L^\infty}$, which is bounded by $2\|\boldsymbol{\phi}_{c_\star}\|_{H^1}$. As follows from (5.11),

$$\|\boldsymbol{v}(t)\|_{H^1_{\mu/2}} \le C_5(\zeta_1 + \mathscr{Y}(\eta_1))\zeta_M(t). \tag{5.15}$$

Using this bound in (5.14), we get $\|J\boldsymbol{N}\|_{L^2_\mu} \le C\zeta_M^2$. The bound (5.12) follows. ◻

6. Choosing the initial perturbation. In this section, we show how to choose the initial perturbation that indeed leads to the instability and conclude the proof of Theorem 1.

We choose $\eta_1 > 0$, $\zeta_1 > 0$, and $\delta_1 > 0$ small enough so that (4.1), (4.17), (4.18) are satisfied, and so that Lemmas 4.1 and 4.3 apply to both exponential weights μ and $\mu/2$. Taking $\eta_1 > 0$, $\zeta_1 > 0$ smaller if necessary, we may assume that the conditions (5.1), (5.3), and (5.4) are satisfied, and moreover that

$$C_6 \zeta_1 < 1/2, \tag{6.1}$$

where $C_6 > 0$ is from Lemma 5.3.

Let

$$\lambda(\eta)=\lambda_{c_\star+\eta},\qquad \Lambda(\eta)=\int_0^\eta \lambda(\eta')\,d\eta'. \tag{6.2}$$

Let us recall that, according to (1.16), we assume that there exists $\eta_1>0$ so that $\mathscr{N}_c'<0$ and is nonincreasing for $c_\star<c\le c_\star+\eta_1$. Thus, we assume that $\lambda(\eta)>0$ for $0<\eta\le\eta_1$ (according to (2.16), $\mathscr{N}_c'$ and λ_c are of opposite sign).

LEMMA 6.1. *One can choose $\eta_1>0$ sufficiently small so that for $0<\eta\le\eta_1$ one has*

$$3C_6e^{2C_6\eta}\Lambda(\eta)<\lambda(\eta). \tag{6.3}$$

Proof. By (2.16), $\lambda_c=-\frac{\mathrm{N}_c'}{B_c}$, where

$$B_c=\langle\boldsymbol{\phi}_c,\boldsymbol{e}_{3,c}\rangle. \tag{6.4}$$

Since $B_{c_\star}>0$ by (2.9), we may assume that $\eta_1>0$ is small enough so that

$$B_{c_\star}/2\le B_c\le 2B_{c_\star},\qquad c\in[c_\star,c_\star+\eta_1]. \tag{6.5}$$

According to Theorem 1, $\mathscr{N}_c'<0$ and is nonincreasing for $c\in(c_\star,c_\star+\eta_1)$. Therefore, using inequalities (6.5), we obtain

$$\Lambda(\eta)=\int_{c_\star}^{c_\star+\eta}\lambda_c\,dc=\int_{c_\star}^{c_\star+\eta}\frac{-\mathscr{N}_c'}{B_c}\,dc\le-\frac{2\eta\mathscr{N}_{c_\star+\eta}'}{B_{c_\star}}\le 4\eta\lambda(\eta),\qquad 0\le\eta\le\eta_1,$$

where $\lambda(\eta)>0$ for $0<\eta\le\eta_1$. We take $\eta_1>0$ so small that $12\,\eta_1C_6e^{2C_6\eta_1}<1$; then (6.3) is satisfied. □

Taking $\eta_1>0$ smaller if necessary, we may assume that Lemma C.1 is satisfied and that

$$\lambda(\eta)/C_6<\zeta_1. \tag{6.6}$$

Remark 6.2. Inequality (6.6) ensures that $\eta(t)$ reaches η_1 prior to $\zeta(t)$ reaching ζ_1 (see Lemma 6.4 and Figure 3).

Since $\Lambda(\eta)=\mathrm{o}(\eta)$, we may also assume that $\eta_1>0$ is small enough so that

$$K_1\Lambda(\eta_1)\le\kappa\eta_1/2, \tag{6.7}$$

where $K_1=K_1(\eta_1,\zeta_1)$ is defined below in (6.26) and $\kappa>0$ is from Lemma C.1.

LEMMA 6.3. *For any $\delta\in(0,\min(\eta_1,\delta_1))$, one can choose the initial data $\eta_0\in(0,\eta_1)$, $\zeta_0\in(0,\zeta_1)$ so that the following estimates are satisfied:*

$$\|\zeta_0\boldsymbol{e}_{3,c_\star+\eta_0}\|_{H^1}<\min(\eta_1,\delta_1), \tag{6.8}$$

$$\|(\boldsymbol{\phi}_{c_\star+\eta_0}+\zeta_0\boldsymbol{e}_{3,c_\star+\eta_0})-\boldsymbol{\phi}_{c_\star}\|_{H^1\cap H^1_\mu}<\delta<\min(\eta_1,\delta_1), \tag{6.9}$$

$$\zeta_0<\Lambda(\eta_0). \tag{6.10}$$

Proof. Pick $\eta_0\in(0,\eta_1)$ so that

$$\|\boldsymbol{\phi}_{c_\star+\eta_0}-\boldsymbol{\phi}_{c_\star}\|_{H^1\cap H^1_\mu}<\delta/2. \tag{6.11}$$

For given $\eta_0 > 0$, we take $\zeta_0 \in (0, \zeta_1)$ small enough so that

$$\zeta_0 \|\boldsymbol{e}_{3,c_\star+\eta_0}\|_{H^1 \cap H^1_\mu} < \delta/2. \tag{6.12}$$

Note that $\|\boldsymbol{e}_{3,c_\star+\eta_0}\|_{H^1}$ for $\eta_0 > 0$ is finite by Lemma 2.7. Inequality (6.12) implies that (6.8) is satisfied. Together with (6.11), it also guarantees that (6.9) holds. We then require that $\zeta_0 > 0$ be small enough so that (6.10) takes place. ∎

We rewrite the last two equations from the system (3.23):

$$\begin{cases} \dot\eta = \zeta + R_2(\eta, \zeta, \boldsymbol{v}), \\ \dot\zeta = \lambda(\eta)\zeta + R_3(\eta, \zeta, \boldsymbol{v}). \end{cases} \tag{6.13}$$

LEMMA 6.4. *For $0 \le t < T_1$, with $T_1 > 0$ as in Proposition 3.2,*

$$\dot\eta \ge \zeta_0/2, \qquad \dot\zeta \ge 0, \tag{6.14}$$

$$\zeta_0 \le \zeta(t) < 3e^{2C_6\eta(t)}\Lambda(\eta(t)). \tag{6.15}$$

Proof. According to Proposition 3.2, the trajectory $(\eta(t), \zeta(t))$ that starts at (η_0, ζ_0) satisfies the inequalities $\eta(t) < \eta_1$ and $\zeta(t) < \zeta_1$ for $0 \le t < T_1$. We define the region $\Omega \subset \mathbb{R}_+ \times \mathbb{R}_+$ by

$$\Omega = \{(\eta, \zeta)\colon \zeta_0 \le \zeta \le \lambda(\eta)/C_6,\ \eta_0 \le \eta \le \eta_1\}. \tag{6.16}$$

Define $T_\Omega \in \mathbb{R}_+ \cup \{+\infty\}$ by

$$T_\Omega = \sup\{\, t \in [0, T_1)\colon\ (\eta(t), \zeta(t)) \in \Omega, \quad \dot\zeta(t) \ge 0\,\}. \tag{6.17}$$

Let us argue that $T_\Omega > 0$. At $t = 0$, $(\eta(0), \zeta(0)) = (\eta_0, \zeta_0) \in \Omega$. From (6.13), we compute $\dot\eta(0) \ge \zeta_0 - C_6\zeta_0^2 > 0$, where we applied the bounds (5.12) and the inequality $C_6\zeta_0 < 1/2$ that follows from (6.1) and the choice $\zeta_0 < \zeta_1$. Similarly, $\dot\zeta(0) \ge \lambda(\eta_0)\zeta_0 - C_6\zeta_0^2 > 0$ due to the inequality $C_6\zeta_0 < \lambda(\eta_0)$ that follows from (6.10) and (6.3). Therefore, $(\eta(t), \zeta(t)) \in \Omega$ and $\dot\zeta(t) > 0$ for times $t > 0$ from a certain open neighborhood of $t = 0$, proving that $T_\Omega > 0$.

The monotonicity of $\zeta(t)$ for $t < T_\Omega$ implies that $\zeta_M(t) := \sup_{s\in(0,t)} |\zeta(s)| = \zeta(t)$ for $0 \le t < T_\Omega$, and (5.12) takes the form

$$|R_j(\eta, \zeta, \boldsymbol{v})| \le C_6\zeta^2, \qquad j = 2, 3, \qquad 0 \le t < T_\Omega. \tag{6.18}$$

Using (6.13) and (6.18), and taking into account (6.1) and monotonicity of $\zeta(t)$ for $0 \le t < T_\Omega$, we compute

$$\dot\eta(t) = \zeta(t) + R_2 \ge \zeta(t) - C_6\zeta^2(t) = \zeta(t)(1 - C_6\zeta(t)) > \zeta_0/2, \tag{6.19}$$

which is valid for $0 \le t < T_\Omega$. This allows us to consider ζ as a function of η (as long as $0 \le t < T_\Omega$). By (6.13), (6.18), and (6.1),

$$\frac{d\zeta}{d\eta} = \frac{\lambda(\eta)\zeta + R_3}{\zeta + R_2} \le \frac{\lambda(\eta)\zeta + C_6\zeta^2}{\zeta - C_6\zeta^2} = \frac{\lambda(\eta) + C_6\zeta}{1 - C_6\zeta} \le 2(\lambda(\eta) + C_6\zeta), \tag{6.20}$$

which is valid for $0 \le t < T_\Omega$. Thus, $\frac{d\zeta}{d\eta} - 2C_6\zeta < 2\lambda(\eta)$ for $0 \le t < T_\Omega$. Multiplying both sides of this relation by $e^{-2C_6\eta}$ and integrating, we get Gronwall's inequality:

$$\int_{\eta_0}^{\eta} \frac{d}{d\eta'}\left(e^{-2C_6\eta'}\zeta(\eta')\right) d\eta' < 2\int_{\eta_0}^{\eta} e^{-2C_6\eta'}\lambda(\eta')\, d\eta' \le 2e^{-2C_6\eta_0}\Lambda(\eta), \tag{6.21}$$

$$\zeta < e^{2C_6\eta}\Big(2e^{-2C_6\eta_0}\Lambda(\eta) + e^{-2C_6\eta_0}\zeta_0\Big) < 3e^{2C_6\eta}\Lambda(\eta), \qquad 0 \le t < T_\Omega. \tag{6.22}$$

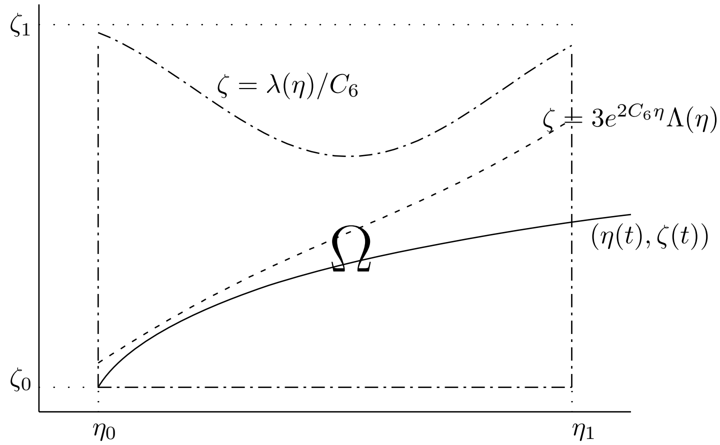


FIG. 3. *The trajectory $(\eta(t), \zeta(t))$ (the solid line) stays in the part of the region Ω below the dashed line $\zeta = 3e^{2C_6\eta}\Lambda(\eta)$.*

See Figure 3. We used the inequality $\zeta_0 < \Lambda(\eta_0) \le \Lambda(\eta)$ that follows from (6.10) and monotonicity of $\Lambda(\eta)$.

Now let us argue that $T_\Omega = T_1$. If $T_\Omega = \infty$, we are done; therefore we only need to consider the case $T_\Omega < \infty$. By (6.17), the moment T_Ω is characterized by

$$\text{either} \quad T_\Omega = T_1 \quad \text{or} \quad (\eta(T_\Omega), \zeta(T_\Omega)) \in \partial\Omega \quad \text{or} \quad \dot{\eta}(T_\Omega) = 0, \tag{6.23}$$

or any combination of these three conditions. By continuity, the bound (6.22) is also valid at T_Ω (the last inequality in (6.22) remains strict); therefore,

$$\zeta(T_\Omega) < 3e^{2C_6\eta(T_\Omega)}\Lambda(\eta(T_\Omega)) < \lambda(\eta(T_\Omega))/C_6. \tag{6.24}$$

In the last inequality, we used Lemma 6.1. Inequality (6.24) also leads to

$$\dot{\zeta} = \lambda(\eta)\zeta + R_3 \ge \zeta(\lambda(\eta) - C_6\zeta) > 0, \qquad 0 \le t \le T_\Omega. \tag{6.25}$$

Using (6.24) and (6.25) in (6.23), we conclude that either $T_\Omega = T_1$ or $\eta(T_\Omega) = \eta_1$, and hence again $T_\Omega = T_1$ (by (3.15), $\eta(t) < \eta_1$ for $0 \le t < T_1$). The bounds (6.14) and (6.15) for $0 \le t < T_\Omega = T_1$ follow from (6.19) and (6.22) (note that $\dot{\zeta} \ge 0$ for $0 \le t < T_\Omega = T_1$ by (6.17)). $\square$

LEMMA 6.5. *Assume that $\|\boldsymbol{\rho}_0\|_{H^1} < \eta_1$. There exists $C_7 > 0$ so that*

$$\|\boldsymbol{\rho}(t)\|_{L^2_\mu} \le C_7\Lambda(\eta), \qquad 0 \le t < T_1.$$

Proof. Using the estimate (6.15) from Lemma 6.4 and the estimate (5.9) from Proposition 5.1 (where $\eta_M(t) = \eta(t)$ and $\zeta_M(t) = \zeta(t)$ due to (6.14) and positivity of η_0 and ζ_0), we obtain

$$\|\boldsymbol{\rho}(t)\|_{L^2_\mu} \le |\zeta|\|\boldsymbol{e}_{3,c}\|_{L^2_\mu} + \|\boldsymbol{v}\|_{L^2_\mu} \le |\zeta|\left(\|\boldsymbol{e}_{3,c}\|_{L^2_\mu} + C_4[\zeta + \mathscr{Y}(\eta)]\right).$$

Now the statement of the lemma follows from the bound (6.15). The value of C_7 could be taken equal to $K_1 = K_1(\eta_1, \zeta_1)$, which we define by

$$K_1 = 3e^{2C_6\eta_1}\left(\sup_{c\in[c_\star, c_\star+\eta_1]} \|\boldsymbol{e}_{3,c}\|_{L^2_\mu} + C_4\big[\zeta_1 + \big\{2\eta_1 + |\mathscr{N}_{c_\star+\eta_1} - \mathscr{N}_{c_\star}|^{\frac{1}{2}}\big\}\big]\right). \tag{6.26}$$

Note that the term in the braces dominates $\mathscr{Y}(\eta)$, which was defined in (4.2) (when estimating $\mathscr{Y}(\eta)$, we used the bound $\|\boldsymbol{\rho}_0\|_{H^1} < \eta_1$). ∎

Conclusion of the proof of Theorem 1. In Theorem 1, let us take

$$\epsilon = \min(\kappa\eta_1/2, \|\boldsymbol{\phi}_{c_\star}\|_{H^1}) > 0. \tag{6.27}$$

Pick $\delta > 0$ arbitrarily small. To comply with the requirements of Lemmas 6.3 and 6.5, we may assume that δ is smaller than $\min(\eta_1, \delta_1)$. Fix $\mu \in (0, \min(\mu_0, \mu_1))$, with μ_0 from Assumption 3 and μ_1 as in Lemma C.1. Let η_0 and ζ_0 satisfy all the inequalities in Lemma 6.3; then the conditions (3.11) of Proposition 3.2 are satisfied. Let

$$\boldsymbol{u}_0 = \boldsymbol{\phi}_{c_\star+\eta_0} + \zeta_0 \boldsymbol{e}_{3,c_\star+\eta_0},$$

so that $\boldsymbol{u}_0 \in H^2(\mathbb{R}) \cap L^2_{2\mu}(\mathbb{R})$ by (2.25) and $\|\boldsymbol{u}_0 - \boldsymbol{\phi}_{c_\star}\|_{H^1} < \delta$ by (6.9). Proposition 3.2 states that there exist $T_1 \in \mathbb{R}_+ \cup \{+\infty\}$ and a function $\boldsymbol{u}(t) \in C([0,\infty), H^2(\mathbb{R}) \cap L^2_{2\mu}(\mathbb{R}))$, $\boldsymbol{u}(0) = \boldsymbol{u}_0$, so that for $0 \le t < T_1$ the function $\boldsymbol{u}(t)$ solves (1.1) and all the inequalities (3.15) are satisfied.

LEMMA 6.6. *In Proposition* 3.2*, one can only take $T_1 < \infty$.*

Proof. If we had $T_1 = +\infty$, then $\dot\eta \ge \zeta_0/2$ for $t \in \mathbb{R}_+$ by Lemma 6.4; hence $\eta(t)$ would reach η_1 in finite time, contradicting the bound $\eta(t) < \eta_1$ for $0 \le t < T_1$ from Proposition 3.2 (iii). ∎

Since $T_1 < \infty$, Proposition 3.2 (iv) states that at least one of the inequalities in (3.15) turns into an equality at $t = T_1$. As follows from the bound (5.9) and inequality (5.4), $\|\boldsymbol{v}(T_1)\|_{H^1_\mu} < \delta_1$. Also, by (6.15) (where the bound from above does not have to be strict at T_1),

$$\zeta(T_1) \le 3e^{2C_6\eta_1}\Lambda(\eta(T_1)) \le 3e^{2C_6\eta_1}\Lambda(\eta_1) < \lambda(\eta)/C_6 < \zeta_1. \tag{6.28}$$

We took into account the monotonicity of $\Lambda(\eta)$ and inequalities (6.3) and (6.6). Therefore, either $\|\boldsymbol{u}(T_1)\|_{H^1} = 2\|\boldsymbol{\phi}_{c_\star}\|_{H^1}$ or $\eta(T_1) = \eta_1$ (or both). In the first case,

$$\inf_{s\in\mathbb{R}} \|\boldsymbol{u}(\cdot, T_1) - \boldsymbol{\phi}_{c_\star}(\cdot - s)\|_{H^1} \ge \|\boldsymbol{u}(\cdot, T_1)\|_{H^1} - \|\boldsymbol{\phi}_{c_\star}\|_{H^1} \ge \|\boldsymbol{\phi}_{c_\star}\|_{H^1} \ge \epsilon; \tag{6.29}$$

hence the instability of $\boldsymbol{\phi}_{c_\star}$ follows. We are left to consider the case $\eta(T_1) = \eta_1$. According to (3.6),

$$\begin{aligned}\inf_{s\in\mathbb{R}} \|\boldsymbol{u}(\cdot, t) - \boldsymbol{\phi}_{c_\star}(\cdot - s)\|_{L^2} &\ge \inf_{s\in\mathbb{R}} \|\boldsymbol{u}(\cdot, t) - \boldsymbol{\phi}_{c_\star}(\cdot - s)\|_{L^2(\mathbb{R}, \min(1, e^{\mu x})\,dx)} \\ &\ge \inf_{s\in\mathbb{R}} \|\boldsymbol{\phi}_{c(t)}(\cdot) - \boldsymbol{\phi}_{c_\star}(\cdot - s)\|_{L^2(\mathbb{R}, \min(1, e^{\mu x})\,dx)} - \|\boldsymbol{\rho}(t)\|_{L^2_\mu}.\end{aligned} \tag{6.30}$$

Applying Lemmas C.1 and 6.5 to the two terms in the right-hand side of (6.30), we see that

$$\inf_{s\in\mathbb{R}} \|\boldsymbol{u}(\cdot, t) - \boldsymbol{\phi}_{c_\star}(\cdot - s)\|_{L^2} \ge \kappa\eta - C_7\Lambda(\eta), \qquad 0 \le t < T_1, \quad \kappa > 0. \tag{6.31}$$

Since $C_7\Lambda(\eta_1) \le \kappa\eta_1/2$ by (6.7),

$$\inf_{s\in\mathbb{R}} \|\boldsymbol{u}(\cdot, T_1) - \boldsymbol{\phi}_{c_\star}(\cdot - s)\|_{L^2} \ge \kappa\eta_1/2 \ge \epsilon, \tag{6.32}$$

and again the instability of $\boldsymbol{\phi}_{c_\star}$ follows.

This completes the proof of Theorem 1.

7. Nondegenerate case: Normal form. In this section, we prove that the critical soliton with speed $c_\star$ generally corresponds to the saddle-node bifurcation of two branches of noncritical solitons. We assume for simplicity that $c_\star$ is a nondegenerate critical point of $\mathscr{N}_c$, in the sense that

$$\mathscr{N}'_{c_\star} = 0, \qquad \mathscr{N}''_{c_\star} \neq 0. \tag{7.1}$$

We rewrite the last two equations from system (3.23):

$$\begin{bmatrix} \dot\eta \\ \dot\zeta \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & \lambda_c \end{bmatrix} \begin{bmatrix} \eta \\ \zeta \end{bmatrix} + \begin{bmatrix} R_2(\eta,\zeta,\boldsymbol{v}) \\ R_3(\eta,\zeta,\boldsymbol{v}) \end{bmatrix}. \tag{7.2}$$

As follows from (2.9) and (2.16),

$$\lambda_c = \lambda_{c_\star+\eta} = \lambda'_{c_\star}\eta + \mathrm{O}(\eta^2), \qquad \lambda'_{c_\star} = -\frac{2\mathscr{N}''_{c_\star}}{(I'_{c_\star})^2}, \tag{7.3}$$

where $\lambda'_{c_\star} \neq 0$ by (7.1). System (7.2) has the nonlinear terms $R_j(\eta,\zeta,\boldsymbol{v})$, $j = 2, 3$, estimated in Lemma 5.3 for monotonically increasing functions $\eta(t)$, $|\zeta(t)|$ on a local existence interval $0 < t < T_1$. It follows from (3.24) that

$$R_2(0,0,0) = R_3(0,0,0) = 0,$$

so that the point $(\eta,\zeta) = (0,0)$ is a critical point of (7.2) when $\boldsymbol{v} = 0$. This critical point corresponds to the critical traveling wave $\boldsymbol{\phi}_{c_\star}(x)$ itself. The following result establishes a local equivalence between system (7.2) and the truncated system $\ddot\eta = \lambda'_{c_\star}\eta\dot\eta$, thus guaranteeing the instability of the critical point $(\eta,\zeta) = (0,0)$.

Proposition 7.1. *Assume that conditions* (7.1) *are satisfied. Consider the subset of trajectories* $(\eta(t),\zeta(t))$ *of system* (7.2) *that lie inside the* ϵ*-neighborhood* $\mathcal{D}_\epsilon \subset \mathbb{R}^2$ *of the origin and satisfy the condition that both functions* $\eta(t)$ *and* $|\zeta(t)|$ *are monotonically increasing. For sufficiently small* $\epsilon > 0$ *this subset of the trajectories is topologically equivalent to a subset of the trajectories of the truncated normal form,*

$$\dot x = \frac{1}{2}\lambda'_{c_\star} x^2 + E_1, \tag{7.4}$$

where E_1 *is constant.*

Proof. Since $\zeta = \dot\eta - R_2(\eta,\zeta,\boldsymbol{v})$, we can rewrite system (7.2) in the equivalent form

$$\frac{d}{dt}\left(\dot\eta - \frac{1}{2}\lambda'_{c_\star}\eta^2 - R_2(\eta,\zeta,\boldsymbol{v})\right) = R(\eta,\zeta,\boldsymbol{v}), \tag{7.5}$$

where

$$R(\eta,\zeta,\boldsymbol{v}) \equiv R_3(\eta,\zeta,\boldsymbol{v}) - \lambda_c R_2(\eta,\zeta,\boldsymbol{v})) + (\lambda_c - \lambda'_{c_\star}\eta)\zeta.$$

It follows from Lemma 5.3 and (7.3) that there exists a constant $C > 0$ such that $|R| \le C(\zeta^2 + \eta^2|\zeta|)$. The integral form of (7.5) is

$$\dot\eta - \frac{1}{2}\lambda'_{c_\star}\eta^2 - E_1 = \tilde R(t), \tag{7.6}$$

where

$$\tilde{R}(t) \equiv R_2(\eta(t), \zeta(t), \boldsymbol{v}(t)) + \int_0^t R(\eta(t'), \zeta(t'), \boldsymbol{v}(t'))\, dt'$$

and E_1 is the constant of integration. Using Lemma 5.3, the bound $|\zeta| \le \dot{\eta} + C_6\zeta^2$, and integration by parts, we obtain that

$$\int_0^t \zeta^2\, dt' \le \eta|\zeta| + C_6 \int_0^t |\zeta|^3\, dt' \le \eta|\zeta| + C_6\eta|\zeta|^2 + C_6^2 \int_0^t |\zeta|^4\, dt' \le \cdots \le \frac{\eta|\zeta|}{1 - C_6|\zeta|}$$

and

$$\int_0^t \eta^2|\zeta|\, dt' \le \frac{\eta^3}{3} + C_6 \int_0^t \eta^2\zeta^2\, dt' \le \cdots \le \frac{\eta^3}{3(1 - C_6|\zeta|)}.$$

Thus, if $|\zeta|$ is sufficiently small, there exists a constant $\tilde{C} > 0$ such that $|\tilde{R}| \le \tilde{C}(\zeta^2 + |\zeta|\eta + \eta^3)$. The topological equivalence of (7.6) with the above estimate on $|\tilde{R}|$ in the disk $(\eta, \zeta) \in \mathcal{D}_\epsilon$ to the truncated normal form (7.4) with sufficiently small E_1 is proved in [Kuz98, Lemma 3.1]. By definition, two systems are said to be topologically equivalent if there exists a homeomorphism between solutions of these systems. We note that this equivalence holds for a family of trajectories which corresponds to monotonically increasing functions $\eta(t)$, $|\zeta(t)|$ in a subset of the small disk near $(\eta, \zeta) = (0, 0)$. □

Corollary 7.2. *The critical point* $(0, 0)$ *of system* (7.2) *is unstable in the sense that there exists* $\epsilon > 0$ *such that for any* $\delta > 0$ *there are* $(\eta(0), \zeta(0)) \in \mathcal{D}_\delta$ *and* $t_* = t_*(\delta, \epsilon) < \infty$ *such that* $(\eta(t_*), \zeta(t_*)) \notin \mathcal{D}_\epsilon$.

Proof. The normal form equation (7.4) shows that the critical point $x = 0$ is semistable at $E_1 = 0$, such that the trajectory with any $x(0) \neq 0$ of the same sign as $\lambda'_{c_\star}$ escapes the local neighborhood of the point $x = 0$ in a local time $t \in [0, T]$. By Proposition 7.1, local dynamics of (7.4) for $x(t)$ is equivalent to local dynamics of (7.2) for (η, ζ). □

Remark 7.3. The truncated normal form (7.4) is rewritten for $c = c_\star + x$:

$$\dot{c} = \frac{1}{2}\lambda'_{c_\star}(c - c_\star)^2 + E_1. \tag{7.7}$$

The normal form (7.7) corresponds to the standard saddle-node bifurcation. It was derived and studied in [PG96] by using the asymptotic multiscale expansion method. When $E = 0$, the critical point $c = c_\star$ is a degenerate saddle point, which is nonlinearly unstable. Assume for definiteness that $\lambda'_{c_\star} > 0$ (which implies that $\mathscr{N}''_{c_\star} < 0$). Then there are no fixed points for $E_1 > 0$ and two fixed points for $E_1 < 0$ in the normal form equation (7.7). Therefore, there exist initial perturbations (with $E_1 > 0$ and any c_0 or with $E_1 = 0$ and $c_0 > c_\star$) which are arbitrarily close to the traveling wave with $c = c_\star$, but the norm $|c - c_\star|$ exceeds some a priori fixed value at $t = t_* > 0$. Two fixed points exist for $E_1 < 0$:

$$c = c_E^\pm = c_\star \pm \sqrt{\frac{E_1}{\mathscr{N}''_{c_\star}}}|I'_{c_\star}|, \tag{7.8}$$

so that $c = c_E^+$ is an unstable saddle point and $c = c_E^-$ is a stable node. The two fixed points correspond to two branches of traveling waves with $\mathscr{N}_c < \mathscr{N}_{\max}$, where

$\mathscr{N}_{\max} = \mathscr{N}(\boldsymbol{\phi}_{c_\star})$. The left branch with $c_E^- < c_\star$ corresponds to $\mathscr{N}'_{c_E^-} > 0$ and the right branch with $c_E^+ > c_\star$ corresponds to $\mathscr{N}'_{c_E^+} < 0$. According to the stability theory for traveling waves [PW92], the left branch is orbitally stable, while the right branch is linearly unstable.

Appendix A. Existence of solitary waves. Let us discuss the existence of standing waves. We assume that f is smooth. Let F denote the primitive of f such that $F(0) = 0$. Thus, by (1.2),

$$F(0) = F'(0) = F''(0) = 0. \tag{A.1}$$

The wave profile $\boldsymbol{\phi}_c$ is to satisfy the equation

$$u'' - cu = f(u), \qquad c > 0.$$

Multiplying the above relation by u' and integrating, and taking into account that we need $\lim_{|x|\to\infty} u(x) = 0$, we get

$$\frac{du(x)}{dx} = \pm\sqrt{cu^2 + 2F(u)}. \tag{A.2}$$

There will be a strictly positive continuous solution exponentially decaying at infinity if there exists $\boldsymbol{\xi}_c > 0$ such that $c\frac{u^2}{2} + F(u) > 0$ for $0 < u < \boldsymbol{\xi}_c$, and also

$$c\frac{\boldsymbol{\xi}_c^2}{2} + F(\boldsymbol{\xi}_c) = 0, \qquad c\boldsymbol{\xi}_c + f(\boldsymbol{\xi}_c) < 0.$$

The last two conditions imply that the map $c \mapsto \boldsymbol{\xi}_c$ is invertible and smooth (as F is). One immediately sees that $\boldsymbol{\phi}_c \in C^\infty(\mathbb{R})$ and, due to the exponential decay at infinity, $\boldsymbol{\phi}_c \in H^\infty(\mathbb{R})$. For each c, the solution $\boldsymbol{\phi}_c$ is unique (up to translations of the origin) and (after a suitable translation of the origin) satisfies the following properties: it is strictly positive, symmetric, and monotonically decreasing (strictly) away from the origin. This result follows from the implicit representation

$$x = \pm\int_{\boldsymbol{\phi}_c}^{\boldsymbol{\xi}_c} \frac{du}{\sqrt{cu^2 + 2F(u)}}. \tag{A.3}$$

See [BL83, section 6] for the exhaustive treatment of this subject.

LEMMA A.1. *There exist positive constants C_1, C_2, C_1', and C_2' such that*

$$C_1 e^{-\sqrt{c}|x|} \le |\boldsymbol{\phi}_c(x)| \le C_2 e^{-\sqrt{c}|x|}, \qquad x \in \mathbb{R}, \tag{A.4}$$

$$C_1' e^{-\sqrt{c}|x|} \le |\partial_x \boldsymbol{\phi}_c(x)| \le C_2' e^{-\sqrt{c}|x|}, \qquad |x| \ge 1. \tag{A.5}$$

Proof. Since $\lim_{|x|\to\infty} \boldsymbol{\phi}_c(x) = 0$, there exists $x_1 > 0$ so that $\frac{|F(\boldsymbol{\phi}_c(x))|}{\boldsymbol{\phi}_c^2(x)} < \frac{c}{4}$ for $|x| \ge x_1$. Then, for $x > x_1$, we get from (A.3)

$$x - x_1 = \int_{\boldsymbol{\phi}_c(x)}^{\boldsymbol{\phi}_c(x_1)} \frac{du}{\sqrt{cu^2 + 2F(u)}}.$$

It follows that

$$\begin{aligned} \int_{\boldsymbol{\phi}_c(x)}^{\boldsymbol{\phi}_c(x_1)} \frac{du}{c^{1/2}u} - \int_{\boldsymbol{\phi}_c(x)}^{\boldsymbol{\phi}_c(x_1)} \frac{|F(u)|}{c^{3/2}u^3}\,du \le x - x_1 \\ \le \int_{\boldsymbol{\phi}_c(x)}^{\boldsymbol{\phi}_c(x_1)} \frac{du}{c^{1/2}u} + \int_{\boldsymbol{\phi}_c(x)}^{\boldsymbol{\phi}_c(x_1)} \frac{|F(u)|}{c^{3/2}u^3}\,du. \end{aligned} \tag{A.6}$$

By (A.1), $|F(u)|/u^3$ is bounded for u small, and we conclude from (A.6) that

$$\ln \boldsymbol{\phi}_c(x) - C_3 \le c^{1/2}(x - x_1) \le \ln \boldsymbol{\phi}_c(x) + C_3, \tag{A.7}$$

where $C_3 = c^{-1} \int_0^{\boldsymbol{\phi}_c(x_1)} |F(u)| u^{-3}\, du$. Inequalities (A.7) immediately prove (A.4). Bounds (A.5) immediately follow from (A.2). □

We also need the following result that gives the rate of decay of $\boldsymbol{e}_{2,c} = \partial_c \boldsymbol{\phi}_c$ and $\boldsymbol{e}_{3,c_\star}$ at infinity.

Lemma A.2. *Let $R \in C^\infty(\mathbb{R})$ satisfy the bound $|R(x)| \le C_1 e^{-\sqrt{c}|x|}$ for $x \ge 0$ for some $c > 0$, $C_1 > 0$. Let $u \in C^\infty(\mathbb{R})$ satisfy*

$$u'' - cu = R, \qquad \lim_{x\to+\infty} u(x) = 0. \tag{A.8}$$

Then there exists $C_2 > 0$ (that depends on c, C_1, and u) such that

$$|u(x)| \le C_2(1 + |x|)e^{-\sqrt{c}|x|}, \qquad x \ge 0. \tag{A.9}$$

Remark A.3. C_2 depends not only on c and C_1 but also on u because the solution to (A.8) is defined up to $\mathrm{const}\, e^{-\sqrt{c}\,x}$.

Proof. First, we notice that if $P \in C^\infty(\mathbb{R})$, $P(x) \ge 0$ for $x \ge 0$, and if $v \in C^\infty(\mathbb{R})$ solves

$$v'' - cv = P(x), \qquad v(0) = 0, \qquad \lim_{x\to+\infty} v(x) = 0, \tag{A.10}$$

then $v(x) \le 0$ for $x \ge 0$. (The existence of a point $x_0 > 0$ where u assumes a positive maximum contradicts the equation in (A.10).)

Now we consider the functions u_- and u_+ that satisfy

$$u_\pm''(x) - cu_\pm = \pm C_1 e^{-\sqrt{c}|x|}, \qquad u_\pm(0) = u(0), \qquad \lim_{x\to+\infty} u_\pm(x) = 0. \tag{A.11}$$

Both $u_\pm$ can be written explicitly; they satisfy (A.9). Since $v = u - u_-$ and $v = u_+ - u$ satisfy (A.10) with $P(x) = C_1 e^{-\sqrt{c}|x|} + R(x)$ and $P(x) = C_1 e^{-\sqrt{c}|x|} - R(x)$, respectively, we conclude that $u_+(x) \le u(x) \le u_-(x)$ for $x \ge 0$, and hence u also satisfies (A.9). □

Appendix B. Fredholm alternative for $\boldsymbol{\mathcal{H}_c}$.

Lemma B.1 (Fredholm alternative). *Let $R(x) \in \mathscr{S}_{+,m}(\mathbb{R})$, $m \ge 0$ (see Definition* 2.1*). If*

$$\int_{\mathbb{R}} \boldsymbol{e}_{1,c}(x) R(x)\, dx = 0, \tag{B.1}$$

then the equation

$$\mathcal{H}_c u = R \tag{B.2}$$

has a solution $u \in \mathscr{S}_{+,m}(\mathbb{R})$. (This solution is unique if we impose the constraint $\langle \boldsymbol{e}_{1,c}, u\rangle = 0$.) Otherwise, any solution $u(x)$ to (B.2) *such that $\lim_{x\to+\infty} u(x) = 0$ grows exponentially at $-\infty$:*

$$\lim_{x\to-\infty} e^{-\sqrt{c}|x|} u(x) \ne 0.$$

Proof. Let us pick an even function $R_+ \in H^\infty(\mathbb{R})$ so that $R_+(x) = R(x)$ for $x \ge 1$. Since R_+ is even and therefore orthogonal to the kernel of the operator $\mathcal{H}_c$, there is a solution $u_+ \in H^\infty(\mathbb{R})$ to the equation

$$\mathcal{H}_c u_+ = R_+. \tag{B.3}$$

Denote by u the solution to the ordinary differential equation

$$\mathcal{H}_c u \equiv -u'' + (f'(\boldsymbol{\phi}_c) + c)u = R \tag{B.4}$$

such that $u|_{x=1} = u_+|_{x=1}$, $u'|_{x=1} = u'_+|_{x=1}$. Then $u \in C^\infty(\mathbb{R})$ coincides with u_+ for $x \ge 1$ and thus satisfies

$$\lim_{x\to+\infty} u(x) = 0. \tag{B.5}$$

We take the pairing of (B.4) with $\boldsymbol{e}_{1,c}$:

$$\int_x^\infty \boldsymbol{e}_{1,c}(y)\mathcal{H}_c u(y)\,dy = \int_x^\infty \boldsymbol{e}_{1,c}(y)R(y)\,dy \equiv r(x), \qquad x \in \mathbb{R}. \tag{B.6}$$

Since

$$\boldsymbol{e}_{1,c}\mathcal{H}_c u = u\mathcal{H}_c\boldsymbol{e}_{1,c} - \boldsymbol{e}_{1,c}\partial_x^2 u + u\partial_x^2\boldsymbol{e}_{1,c} = -\partial_x(\boldsymbol{e}_{1,c}u') + \partial_x(u\partial_x\boldsymbol{e}_{1,c}),$$

where we took into account that $\mathcal{H}_c\boldsymbol{e}_{1,c} = 0$, we obtain from (B.6) the relation

$$\boldsymbol{e}_{1,c}(x)u'(x) - u(x)\partial_x\boldsymbol{e}_{1,c}(x) = r(x). \tag{B.7}$$

The boundary term at $x = +\infty$ does not contribute into (B.7) due to the limit (B.5). We will use this relation to find the behavior of $u(x)$ as $x \to -\infty$. For $x \le -1$, we divide the relation (B.7) by $\boldsymbol{e}_{1,c}^2$ (we can do this since $\boldsymbol{e}_{1,c}(x) = -\partial_x\boldsymbol{\phi}_c(x) \ne 0$ for $x \ne 0$), getting

$$\partial_x\left(\frac{u(x)}{\boldsymbol{e}_{1,c}(x)}\right) = \frac{r(x)}{\boldsymbol{e}_{1,c}^2(x)}. \tag{B.8}$$

Therefore, for $x \le -1$,

$$u(x) - \boldsymbol{e}_{1,c}(x)\frac{u(-1)}{\boldsymbol{e}_{1,c}(-1)} = \boldsymbol{e}_{1,c}(x)\int_{-1}^x \frac{(r(y) - r_-) + r_-}{\boldsymbol{e}_{1,c}^2(y)}\,dy, \tag{B.9}$$

where $r_- = \lim_{x\to-\infty} r(x)$.

Since $R \in \mathscr{S}_{+,m}(\mathbb{R})$, $|R(x)| \le C(1+|x|)^m$, $m \in \mathbb{Z}$, $m \ge 0$. Using Lemma A.1, we see that

$$|r(x) - r_-| = \left|\int_{-\infty}^x R(y)\boldsymbol{e}_{1,c}(y)\,dy\right| \le \mathrm{const}\, e^{-\sqrt{c}|x|}(1+|x|)^m, \quad x \le -1. \tag{B.10}$$

At the same time, Lemma A.1 also shows that

$$\int_{-1}^x \frac{dy}{\boldsymbol{e}_{1,c}^2(y)} \ge \mathrm{const}\, e^{2\sqrt{c}|x|}, \qquad x \le -1. \tag{B.11}$$

Therefore, if $r_- \neq 0$, the right-hand side of (B.9) grows exponentially as $x \to -\infty$. The same is true for $u(x)$, since the second term in the left-hand side of (B.9) decays exponentially when $|x| \to \infty$ by Lemma A.1. If instead $r_- = 0$, Lemma A.1 and the bound (B.10) show that the right-hand side of (B.9) is bounded by $\mathrm{const}(1+|x|)^m$, proving a similar bound for $u(x)$. Using (B.4) to get the bounds on the derivatives $u^{(N)}$, we conclude that $u \in \mathscr{S}_{+,m}(\mathbb{R})$. ◻

Appendix C. Nondegeneracy of $\inf_{s\in\mathbb{R}} \|\boldsymbol{\phi}_c(\cdot) - \boldsymbol{\phi}_{c_\star}(\cdot - s)\|$ at $c_\star$.

LEMMA C.1. *If $\eta_1 > 0$ is sufficiently small, there exist $\mu_1 > 0$ and $\kappa > 0$ so that*

$$\inf_{s\in\mathbb{R}} \|\boldsymbol{\phi}_c(\cdot) - \boldsymbol{\phi}_{c_\star}(\cdot - s)\|_{L^2(\mathbb{R},\min(1,e^{\mu x})\,dx)} \geq \kappa|c - c_\star|, \quad c \in [c_\star, c_\star + \eta_1], \quad \mu \in [0, \mu_1].$$

Proof. Consider the function

$$g_\mu(c,s) = \|\boldsymbol{\phi}_c(\cdot) - \boldsymbol{\phi}_{c_\star}(\cdot - s)\|^2_{L^2(\mathbb{R},\min(1,e^{\mu x})\,dx)}. \tag{C.1}$$

It is a smooth nonnegative function of c and s for $c \in [c_\star, c_\star + \eta_1]$ and $s \in \mathbb{R}$. It also depends smoothly on the parameter $\mu \geq 0$. Zero is its absolute minimum, achieved at the point $(c,s) = (c_\star, 0)$. We also note that the point $(c_\star, 0)$ is nondegenerate when $\mu = 0$:

$$\partial_c^2 g_0(c,s)|_{(c_\star,0)} = 2\|\partial_c \boldsymbol{\phi}_c|_{c=c_\star}\|^2_{L^2} > 0, \qquad \partial_s^2 g_0(c,s)|_{(c_\star,0)} = 2\|\partial_x \boldsymbol{\phi}_{c_\star}\|^2_{L^2} > 0,$$
$$\partial_c \partial_s g_0(c,s)|_{(c_\star,0)} = -2(\partial_c \boldsymbol{\phi}_c|_{c=c_\star}, \partial_x \boldsymbol{\phi}_{c_\star}) = 0.$$

By continuity, the quadratic form $g''_\mu|_{(c_\star,0)}$ is nondegenerate for $0 \leq \mu \leq \mu_1$, with some $\mu_1 > 0$. Therefore, there exist $\kappa > 0$ and an open neighborhood $\Omega \subset \mathbb{R}^2$ of the point $(c_\star, 0)$ such that

$$g_\mu(c,s) \geq \kappa^2((c - c_\star)^2 + s^2), \qquad (c,s) \in \Omega, \qquad 0 \leq \mu \leq \mu_1. \tag{C.2}$$

Moreover, we claim that

$$\mathit{\Gamma} \equiv \inf_{\mu\in(0,\mu_1)} \inf_{(c,s)\in[c_\star,c_\star+\eta_1]\times\mathbb{R})\backslash\Omega} g_\mu(c,s) > 0. \tag{C.3}$$

To prove (C.3), we only need to note that $(c_\star, 0)$ is the only point where $g_\mu(c,s)$ takes the zero value and that $\lim_{|s|\to\infty} g_\mu(c,s) \geq \inf_{c\in[c_\star,c_\star+\eta_1]} \|\boldsymbol{\phi}_c\|^2_{L^2(\mathbb{R},\min(1,e^{\mu_1 x})\,dx)} > 0$.

Now, we assume that $\eta_1 > 0$ is small enough so that $\kappa^2\eta_1^2 < \mathit{\Gamma}$. Then, by (C.2) (valid for $(c,s) \in \Omega$) and (C.3) (valid for $(c,s) \in ([c_\star, c_\star + \eta_1] \times \mathbb{R})\backslash\Omega$), we conclude that

$$\inf_{s\in\mathbb{R}} g_\mu(c,s) \geq \kappa^2(c - c_\star)^2, \qquad c \in [c_\star, c_\star + \eta_1], \quad \mu \in [0, \mu_1]. \tag{C.4}$$

This proves the lemma. ◻

REFERENCES

[ABH87] J. P. ALBERT, J. L. BONA, AND D. B. HENRY, *Sufficient conditions for stability of solitary-wave solutions of model equations for long waves*, Phys. D, 24 (1987), pp. 343–366.

[Ben72] T. B. BENJAMIN, *The stability of solitary waves*, Proc. Roy. Soc. London Ser. A, 328 (1972), pp. 153–183.

[BL83] H. BERESTYCKI AND P.-L. LIONS, *Nonlinear scalar field equations.* I. *Existence of a ground state*, Arch. Rational Mech. Anal., 82 (1983), pp. 313–345.

[Bon75] J. Bona, *On the stability theory of solitary waves*, Proc. Roy. Soc. London Ser. A, 344 (1975), pp. 363–374.

[BSS87] J. L. Bona, P. E. Souganidis, and W. A. Strauss, *Stability and instability of solitary waves of Korteweg-de Vries type*, Proc. Roy. Soc. London Ser. A, 411 (1987), pp. 395–412.

[Car81] J. Carr, *Applications of Centre Manifold Theory*, Appl. Math. Sci. 35, Springer-Verlag, New York, 1981.

[CP03] A. Comech and D. Pelinovsky, *Purely nonlinear instability of standing waves with minimal energy*, Comm. Pure Appl. Math., 56 (2003), pp. 1565–1607.

[GSS87] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry.* I, J. Funct. Anal., 74 (1987), pp. 160–197.

[IA98] G. Iooss and M. Adelmeyer, *Topics in Bifurcation Theory and Applications*, 2nd ed., Adv. Ser. Nonlinear Dynam. 3, World Scientific, River Edge, NJ, 1998.

[Kat83] T. Kato, *On the Cauchy problem for the (generalized) Korteweg-de Vries equation*, in Studies in Applied Mathematics, Adv. Math. Suppl. Stud. 8, Academic Press, New York, 1983, pp. 93–128.

[Kuz98] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, 2nd ed., Appl. Math. Sci. 112, Springer-Verlag, New York, 1998.

[Mer01] F. Merle, *Existence of blow-up solutions in the energy space for the critical generalized KdV equation*, J. Amer. Math. Soc., 14 (2001), pp. 555–578.

[Miz01] T. Mizumachi, *Large time asymptotics of solutions around solitary waves to the generalized Korteweg–de Vries equations*, SIAM J. Math. Anal., 32 (2001), pp. 1050–1080.

[MM01a] Y. Martel and F. Merle, *Asymptotic stability of solitons for subcritical generalized KdV equations*, Arch. Ration. Mech. Anal., 157 (2001), pp. 219–254.

[MM01b] Y. Martel and F. Merle, *Instability of solitons for the critical generalized Korteweg-de Vries equation*, Geom. Funct. Anal., 11 (2001), pp. 74–123.

[MM02a] Y. Martel and F. Merle, *Blow up in finite time and dynamics of blow up solutions for the L^2-critical generalized KdV equation*, J. Amer. Math. Soc., 15 (2002), pp. 617–664.

[MM02b] Y. Martel and F. Merle, *Stability of blow-up profile and lower bounds for blow-up rate for the critical generalized KdV equation*, Ann. of Math. (2), 155 (2002), pp. 235–280.

[MM05] Y. Martel and F. Merle, *Asymptotic stability of solitons of the subcritical gKdV equations revisited*, Nonlinearity, 18 (2005), pp. 55–80.

[PG96] D. E. Pelinovsky and R. H. J. Grimshaw, *An asymptotic approach to solitary wave instability and critical collapse in long-wave KdV-type evolution equations*, Phys. D, 98 (1996), pp. 139–155.

[PW92] R. L. Pego and M. I. Weinstein, *Eigenvalues, and instabilities of solitary waves*, Philos. Trans. Roy. Soc. London Ser. A, 340 (1992), pp. 47–94.

[PW94] R. L. Pego and M. I. Weinstein, *Asymptotic stability of solitary waves*, Comm. Math. Phys., 164 (1994), pp. 305–349.

[Wei87] M. I. Weinstein, *Existence and dynamic stability of solitary wave solutions of equations arising in long wave propagation*, Comm. Partial Differential Equations, 12 (1987), pp. 1133–1173.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 34–56

ON THE BLOWUP FOR THE L^2-CRITICAL FOCUSING NONLINEAR SCHRÖDINGER EQUATION IN HIGHER DIMENSIONS BELOW THE ENERGY CLASS*

MONICA VISAN† AND XIAOYI ZHANG‡

Abstract. We consider the focusing mass-critical nonlinear Schrödinger equation and prove that blowup solutions to this equation with initial data in $H^s(\mathbb{R}^d)$, $s > s_0(d)$ and $d \geq 3$, concentrate at least the mass of the ground state at the blowup time. This extends recent work by Colliander et al. [*Math. Res. Lett.*, 12 (2005), pp. 357–375], Hmidi and Keraani [*Remarks on the Blowup for the L^2-Critical Nonlinear Schrödinger Equations*, preprint], and Tzirakis [*SIAM J. Math. Anal.*, 37 (2006), pp. 1923–1946] on the blowup of the two-dimensional and one-dimensional mass-critical focusing nonlinear Schrödinger equation below the energy space to all dimensions $d \geq 3$.

Key words. nonlinear Schrödinger equation, blowup, I-method.

AMS subject classifications. 35Q55

DOI. 10.1137/060663969

1. Introduction.

We consider the initial value problem for the focusing L_x^2-critical nonlinear Schrödinger equation

$$\begin{cases} iu_t + \Delta u = -|u|^{\frac{4}{d}}u, \\ u(0,x) = u_0(x) \in H^s(\mathbb{R}^d),\ s \geq 0, \end{cases} \tag{1.1}$$

where $u(t,x)$ is a complex-valued function in spacetime $\mathbb{R} \times \mathbb{R}^d$, $d \geq 3$.

It is well known (see, for example, [6]) that the Cauchy problem (1.1) is locally well posed in $H^s(\mathbb{R}^d)$ for $s \geq 0$. Moreover, the unique solution obeys conservation of *mass*:

$$M(u(t)) := \int_{\mathbb{R}^d} |u(t,x)|^2\,dx = M(u_0).$$

If $s \geq 1$, the *energy* is also finite and conserved:

$$E(u(t)) := \int_{\mathbb{R}^d} \left(\frac{1}{2}|\nabla u(t,x)|^2 - \frac{d}{2(d+2)}|u(t,x)|^{2+\frac{4}{d}} \right) dx = E(u_0).$$

This equation has a natural scaling. More precisely, the map

$$u(t,x) \mapsto u_\lambda(t,x) := \lambda^{\frac{d}{2}} u(\lambda^2 t, \lambda x) \tag{1.2}$$

maps a solution to (1.1) to another solution to (1.1). The reason why this equation is called L_x^2-critical (or mass-critical) is because the scaling (1.2) also leaves the mass invariant.

Equation (1.1) is subcritical for $s > 0$. In this case, (1.1) is well posed in $H^s(\mathbb{R}^d)$ and the life span of the local solution depends only on the H_x^s-norm of the initial data

*Received by the editors June 28, 2006; accepted for publication (in revised form) October 13, 2006; published electronically April 10, 2007.
http://www.siam.org/journals/sima/39-1/66396.html
†School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540 (mvisan@ias.edu).
‡Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing 100080, China (zh.xiaoyi@gmail.com). This author was supported by NSF grant 10601060 (China).

(see [6]). Denote by $T^* > 0$ the maximal forward time of existence. As a consequence of the local well-posedness theory, we have the following blowup criterion:

$$\text{either} \quad T^* = \infty \quad \text{or} \quad T^* < \infty \quad \text{and} \quad \lim_{t \to T^*} \|u(t)\|_{H^s_x} = \infty.$$

The blowup behavior for solutions from H^1_x initial data has received a lot of attention. The results are closely related to the ground state Q which is the unique positive radial solution to the elliptic equation

$$\Delta Q - Q + |Q|^{\frac{4}{d}} Q = 0.$$

Using the sharp Gagliardo–Nirenberg inequality (see [39]),

$$\|u\|^{\frac{4}{d}+2}_{L_x^{\frac{4}{d}+2}} \leq C_d \|u\|^{\frac{4}{d}}_{L^2_x} \|\nabla u\|^2_{L^2_x} \quad \text{with } C_d := \frac{d+2}{d\,\|Q\|^{\frac{4}{d}}_{L^2_x}}, \tag{1.3}$$

it is not hard to see that the mass of the ground state is the minimal mass required for the solution to develop a singularity. Indeed, in the case $\|u_0\|_{L^2_x} < \|Q\|_{L^2_x}$, (1.3) combined with the conservation of energy implies that the solution to (1.1) is global. This is sharp since the pseudoconformal invariance of (1.1) allows us to build a solution with mass equal to that of the ground state that blows up at time T^*:

$$u(t,x) := |T^* - t|^{-\frac{d}{2}} e^{[i(T^*-t)^{-1} - i|x|^2(T^*-t)^{-1}]} Q\Big(\frac{x}{T^* - t}\Big).$$

Moreover, Merle [28] showed that up to the symmetries of (1.1), this is the only blowup solution with minimal mass. Furthermore, any blowup solution must concentrate at least the mass of the ground state at the blowup time; more precisely, as shown in [31], there exists $x(t) \in \mathbb{R}^d$ such that

$$\forall R > 0, \; \lim_{t \to T^*} \int_{|x - x(t)| \leq R} |u(t,x)|^2 dx \geq \int_{\mathbb{R}^d} Q^2 dx. \tag{1.4}$$

Of course, the goal is to establish all these properties for blowup solutions from data in L^2_x rather than H^1_x. Unfortunately, all the methods used in the H^1_x setting break down at the L^2_x level. Moreover, as (1.1) is L^2_x-critical, even the local well-posedness theory in L^2_x is substantially different from that in H^s_x for $s > 0$. Specifically, the life span of the local solution depends on the profile of the initial data, rather than on its L^2_x-norm (see [6]). In particular, this leads to the following blowup criterion:

$$\text{either} \quad T^* = \infty \quad \text{or} \quad T^* < \infty \quad \text{and} \quad \|u\|_{L^{2+\frac{4}{d}}_{t,x}([0,T^*) \times \mathbb{R}^d)} = \infty.$$

From the global theory for small data (see [6]), we know that if the mass of the initial data is sufficiently small, then there exists a unique global solution to (1.1). However, for large (but finite) mass initial data, which is also sufficiently smooth and decaying, the viriel identity guarantees that finite time blowup occurs; see [18, 42].

The first blowup result for general L^2_x initial data belongs to Bourgain [2], who proved the following parabolic concentration of mass at the blowup time:

$$\lim_{t \nearrow T^*} \sup_{\substack{\text{cubes } I \subset \mathbb{R}^2 \\ \text{side}(I) < (T^* - t)^{\frac{1}{2}}}} \Big(\int_I |u(t,x)|^2 \, dx \Big)^{\frac{1}{2}} \geq c(\|u_0\|_{L^2_x}) > 0, \tag{1.5}$$

where $c(\|u_0\|_{L^2_x})$ is a small constant depending on the mass of the initial data. This result was extended to dimension $d = 1$ by Keraani [25], and to dimensions $d \geq 3$ by Begout and Vargas [1]. The conjecture is that rather than $c(\|u_0\|_{L^2_x})$ on the right-hand side of (1.5), one should have the mass of the ground state as in (1.4); however, this appears to be a very difficult problem.

The goal of this paper is to reproduce as much of the H^1_x theory as we can at lower regularity (but, unfortunately, well above L^2_x) in dimensions three and higher. In two dimensions, this was pursued by Colliander et al. [13]; using the *I-method* (introduced by Colliander et al. [10]), they established a mass concentration property (at the blowup time) for radial blowup solutions from initial data in $H^s(\mathbb{R}^2)$ with $s > \frac{1+\sqrt{11}}{5}$:

$$\limsup_{t \nearrow T^*} \|u\|_{L^2(|x|<(T^*-t)^{\frac{s}{2}}\gamma(T^*-t))} \geq \|Q\|_{L^2_x}, \tag{1.6}$$

where $\gamma(r) \to \infty$ arbitrarily slowly as $r \to 0$. While the right-hand side in (1.6) is the conjectured one, we should remark that the width of concentration, that is, $(T^*-t)^{\frac{s}{2}-}$, is larger than the expected concentration width, $(T^*-t)^{\frac{1}{2}-}$. Moreover, the lim sup in (1.6) is not present in the H^1_x theory (see (1.4)); the reason for its appearance is basically the lack of information on the blowup rate of the H^s_x-norm.

Hmidi and Keraani [21] removed the radial assumption used in [13]. Moreover, they showed that, up to symmetries of the equation, the ground state Q is the profile for blowup solutions with minimal mass and initial data in $H^s(\mathbb{R}^2)$ with $s > \frac{1+\sqrt{11}}{5}$. We will make use of the key innovation of their work, namely, Lemma 2.11 below.

In one dimension, recent work by Tzirakis [36] established the analogue of (1.6) for arbitrary initial data in $H^s(\mathbb{R})$, $s > \frac{10}{11}$.

Our contribution is to treat (nonradial) data in $H^s(\mathbb{R}^d)$ for $s > s_0(d)$, where

$$s_0(d) := \begin{cases} \dfrac{1+\sqrt{13}}{5} & \text{for } d = 3, \\ \dfrac{8-d+\sqrt{9d^2+64d+64}}{2(d+10)} & \text{for } d \geq 4. \end{cases} \tag{1.7}$$

We prove the following theorem.

THEOREM 1.1. *Assume $d \geq 3$ and $s > s_0(d)$. Let $u_0 \in H^s(\mathbb{R}^d)$ such that the corresponding solution u to* (1.1) *blows up at time $0 < T^* < \infty$. Then there exists a function $V \in H^1_x$ such that $\|V\|_2 \geq \|Q\|_2$ and there exist sequences $\{t_n, \rho_n, x_n\}_{n\geq 1} \subset \mathbb{R}_+ \times \mathbb{R}_+^* \times \mathbb{R}^d$ satisfying*

$$t_n \nearrow T^* \text{ as } n \to \infty \quad \text{and} \quad \rho_n \lesssim (T^*-t_n)^{\frac{s}{2}} \quad \forall\, n \geq 1$$

such that

$$\rho_n^{\frac{d}{2}} u(t_n, \rho_n \cdot + x_n) \rightharpoonup V \quad \text{weakly as } n \to \infty.$$

As a consequence of Theorem 1.1, we establish the following mass concentration property for blowup solutions.

THEOREM 1.2. *Assume $d \geq 3$ and $s > s_0(d)$. Let $u_0 \in H^s(\mathbb{R}^d)$ such that the corresponding solution u to* (1.1) *blows up at time $0 < T^* < \infty$. Let $\alpha(t) > 0$ be such that*

$$\lim_{t \nearrow T^*} \frac{(T^*-t)^{\frac{s}{2}}}{\alpha(t)} = 0.$$

Then there exists $x(t) \in \mathbb{R}^d$ *such that*

$$\limsup_{t \nearrow T^*} \int_{|x-x(t)| \leq \alpha(t)} |u(x,t)|^2 dx \geq \int_{\mathbb{R}^d} Q^2 dx.$$

Under the additional hypothesis that the mass of the initial data equals the mass of the ground state, we may upgrade Theorem 1.1 to the following.

THEOREM 1.3. *Assume* $d \geq 3$ *and* $s > s_0(d)$. *Let* $u_0 \in H^s(\mathbb{R}^d)$ *with* $\|u_0\|_2 = \|Q\|_2$ *such that the corresponding solution* u *to* (1.1) *blows up at time* $0 < T^* < \infty$. *Then there exist sequences* $\{t_n, \theta_n, \rho_n, x_n\}_{n \geq 1} \subset \mathbb{R}_+ \times S^1 \times \mathbb{R}_+^* \times \mathbb{R}^d$ *satisfying*

$$t_n \nearrow T^* \text{ as } n \to \infty \quad \text{and} \quad \rho_n \lesssim (T^* - t_n)^{\frac{s}{2}} \quad \forall\, n \geq 1$$

such that

$$\rho_n^{\frac{d}{2}} e^{i\theta_n} u(t_n, \rho_n x + x_n) \to Q \quad \text{strongly in } H_x^{\tilde{s}-},$$

where

$$\tilde{s} := \frac{2d + 8s + s^2 d(2 - \min\{1, \frac{4}{d}\})}{4d + 16s - s^2(d \min\{1, \frac{4}{d}\} + 8)}.$$

In a nutshell, Theorem 1.3 says that up to the symmetries for (1.1), the ground state is the profile for blowup solutions with minimal mass and initial data in H_x^s, $s > s_0(d)$. Alas, we only show that this is true along a sequence of times.

To prove Theorems 1.1 through 1.3, we will rely on the *I*-method and Lemma 2.11. The idea behind the *I*-method is to smooth out the initial data in order to access the theory available at H_x^1 regularity. To this end, one introduces the Fourier multiplier I, which is the identity on low frequencies and behaves like a fractional integral operator of order $1 - s$ on high frequencies. Thus, the operator I maps H_x^s to H_x^1. However, even though we do have energy conservation for (1.1), Iu is not a solution to (1.1), and hence we expect an energy increment. The key is to prove that on intervals of local well-posedness, the modified energy $E(Iu)$ is an "almost conserved" quantity and grows much slower than the modified kinetic energy $\|\nabla Iu\|_{L_x^2}^2$. This requires delicate estimates on the commutator between I and the nonlinearity. In dimensions one and two, the nonlinearity is algebraic and one can write the commutator explicitly using the Fourier transform and control it by multilinear analysis and bilinear estimates (see [13, 36]). However, in dimensions $d \geq 3$ this method fails. Instead, we will have to rely on more rudimentary tools such as Strichartz and fractional chain rule estimates in order to control the commutator.

The remainder of this paper is organized as follows: In section 2, we introduce notation and prove some lemmas that will be useful. In section 3, we revisit the H_x^s local well-posedness theory for (1.1). Section 4 is devoted to controlling the modified energy increment. In sections 5 through 7 we prove Theorems 1.1 through 1.3.

After this work was submitted, we were informed of an independent paper attacking the same problem [14]. However, there appear to be several gaps in their argument, for example, in estimating (4.19). While it seems possible to remedy their errors, this would result in a larger value for $s_0(d)$ than that claimed in their paper, which is already inferior to that given here. We contend that all overlap between this paper and [14] can be attributed to the precursors [13] and [21].

2. Preliminaries. We will often use the notation $X \lesssim Y$ whenever there exists some constant C so that $X \leq CY$. Similarly, we will use $X \sim Y$ if $X \lesssim Y \lesssim X$. We use $X \ll Y$ if $X \leq cY$ for some small constant c. The derivative operator ∇ refers to the space variable only. We use $A\pm$ to denote $A \pm \varepsilon$ for any sufficiently small $\varepsilon > 0$; the implicit constant in an inequality involving this notation is permitted to depend on ε.

We use $L^r_x(\mathbb{R}^d)$ to denote the Banach space of functions $f : \mathbb{R}^d \to \mathbb{C}$ whose norm

$$\|f\|_r := \Big(\int_{\mathbb{R}^d} |f(x)|^r dx \Big)^{\frac{1}{r}}$$

is finite, with the usual modifications when $r = \infty$.

We use $L^q_t L^r_x$ to denote the spacetime norm

$$\|u\|_{q,r} := \|u\|_{L^q_t L^r_x(\mathbb{R}\times\mathbb{R}^d)} := \Big(\int_{\mathbb{R}} \Big(\int_{\mathbb{R}^d} |u(t,x)|^r dx \Big)^{q/r} dt \Big)^{1/q},$$

with the usual modifications when either q or r are infinity, or when the domain $\mathbb{R}\times\mathbb{R}^d$ is replaced by some smaller spacetime region. When $q = r$ we abbreviate $L^q_t L^r_x$ by $L^q_{t,x}$.

We define the Fourier transform on $\mathbb{R}^d$ to be

$$\hat{f}(\xi) := \int_{\mathbb{R}^d} e^{-2\pi i x\cdot\xi} f(x) dx.$$

We will make use of the fractional differentiation operators $|\nabla|^s$ defined by

$$\widehat{|\nabla|^s f}(\xi) := |\xi|^s \hat{f}(\xi).$$

These define the homogeneous Sobolev norms

$$\|f\|_{\dot{H}^s_x} := \||\nabla|^s f\|_{L^2_x}$$

and the more common inhomogeneous Sobolev norms

$$\|f\|_{H^{s,p}_x} := \|\langle\nabla\rangle^s f\|_{L^p_x}, \quad \text{where } \langle\nabla\rangle := (1 + |\nabla|^2)^{\frac{1}{2}}.$$

We will often denote $H^{s,2}_x$ by H^s_x.

Let $F(z) := -|z|^{\frac{4}{d}} z$ be the function that defines the nonlinearity in (1.1). Then

$$F_z(z) := \frac{\partial F}{\partial z}(z) = -\frac{2+d}{d}|z|^{\frac{4}{d}} \quad \text{and} \quad F_{\bar{z}}(z) := \frac{\partial F}{\partial \bar{z}}(z) = -\frac{2}{d}|z|^{\frac{4}{d}} \frac{z}{\bar{z}}.$$

We write F' for the vector $(F_z, F_{\bar{z}})$ and adopt the notation

$$w \cdot F'(z) := wF_z(z) + \bar{w}F_{\bar{z}}(z).$$

In particular, we observe the chain rule

$$\nabla F(u) = \nabla u \cdot F'(u).$$

Clearly $F'(z) = O(|z|^{\frac{4}{d}})$ and we have the Hölder continuity estimate

$$|F'(z) - F'(w)| \lesssim |z - w|^{\min\{1,\frac{4}{d}\}}(|z| + |w|)^{\frac{4}{d}-\min\{1,\frac{4}{d}\}} \tag{2.1}$$

for all $z, w \in \mathbb{C}$. By the fundamental theorem of calculus,

$$F(z+w) - F(z) = \int_0^1 w \cdot F'(z+\theta w) d\theta$$

and hence

$$F(z+w) = F(z) + O(|w||z|^{\frac{4}{d}}) + O(|w|^{\frac{4+d}{d}})$$

for all complex values z and w.

Let $e^{it\Delta}$ be the free Schrödinger propagator. In physical space this is given by the formula

$$e^{it\Delta} f(x) = \frac{1}{(4\pi it)^{d/2}} \int_{\mathbb{R}^d} e^{i|x-y|^2/4t} f(y) dy$$

for $t \neq 0$ (using a suitable branch cut to define $(4\pi it)^{d/2}$), while in frequency space one can write this as

$$\widehat{e^{it\Delta} f}(\xi) = e^{-4\pi^2 it|\xi|^2} \hat{f}(\xi). \tag{2.2}$$

In particular, the propagator obeys the *dispersive inequality*

$$\|e^{it\Delta} f\|_{L_x^\infty} \lesssim |t|^{-\frac{d}{2}} \|f\|_{L_x^1} \tag{2.3}$$

for all times $t \neq 0$.

We also recall *Duhamel's formula*

$$u(t) = e^{i(t-t_0)\Delta} u(t_0) - i \int_{t_0}^t e^{i(t-s)\Delta} (iu_t + \Delta u)(s) ds. \tag{2.4}$$

Definition 2.1. *A pair of exponents (q, r) is called Schrödinger-admissible if*

$$\frac{2}{q} + \frac{d}{r} = \frac{d}{2}, \quad 2 \leq q, r \leq \infty, \quad \textit{and} \quad (q, r, d) \neq (2, \infty, 2).$$

Throughout this paper we will use the following admissible pairs:

$$\left(2, \frac{2d}{d-2}\right) \quad \text{and} \quad (\gamma, \rho) := \left(\frac{2(d+2)}{d-2s}, \frac{2d(d+2)}{d^2+4s}\right) \quad \text{with } 0 < s < 1.$$

Let $\rho^* := \frac{2(d+2)}{d-2s}$. Using Hölder and Sobolev embedding, on any spacetime slab $I \times \mathbb{R}^d$ we estimate

$$\|F(u)\|_{\gamma', \rho'} \lesssim |I|^{\frac{2s}{d}} \|u\|_{\gamma,\rho} \|u\|_{\gamma,\rho^*}^{\frac{4}{d}} \lesssim |I|^{\frac{2s}{d}} \|\langle \nabla \rangle^s u\|_{\gamma,\rho}^{1+\frac{4}{d}}. \tag{2.5}$$

Let $I \times \mathbb{R}^d$ be a spacetime slab; we define the Strichartz norm

$$\|u\|_{S^0(I)} := \sup_{(q,r) \text{ admissible}} \|u\|_{L_t^q L_x^r (I \times \mathbb{R}^d)}.$$

We define the Strichartz space $S^0(I)$ to be the closure of all test functions under the Strichartz norm $\|\cdot\|_{S^0(I)}$. We use $N^0(I)$ to denote the dual space of $S^0(I)$.

We record the standard Strichartz estimates which we will invoke repeatedly throughout this paper (see [33, 17] for (q,r) admissible with $q>2$ and [24] for the endpoint $(2,\frac{2d}{d-2})$).

LEMMA 2.2. *Let I be a compact time interval, $t_0\in I$, $s\geq 0$, and let u be a solution to the forced Schrödinger equation*

$$iu_t+\Delta u=\sum_{i=1}^m F_i$$

for some functions $F_1,\ldots,F_m$. Then

$$\||\nabla|^s u\|_{S^0(I)}\lesssim\|u(t_0)\|_{\dot H^s(\mathbb{R}^d)}+\sum_{i=1}^m\||\nabla|^s F_i\|_{L_t^{q_i'}L_x^{r_i'}(I\times\mathbb{R}^d)} \tag{2.6}$$

for any admissible pairs (q_i,r_i), $1\leq i\leq m$.

We will also need some Littlewood–Paley theory. Specifically, let $\varphi(\xi)$ be a smooth bump supported in the ball $|\xi|\leq 2$ and equaling one on the ball $|\xi|\leq 1$. For each dyadic number $N\in 2^{\mathbb{Z}}$ we define the Littlewood–Paley operators

$$\begin{aligned}\widehat{P_{\leq N}f}(\xi)&:=\varphi(\xi/N)\hat f(\xi),\\ \widehat{P_{>N}f}(\xi)&:=[1-\varphi(\xi/N)]\hat f(\xi),\\ \widehat{P_N f}(\xi)&:=[\varphi(\xi/N)-\varphi(2\xi/N)]\hat f(\xi).\end{aligned}$$

Similarly we can define $P_{<N}$, $P_{\geq N}$, and $P_{M<\cdot\leq N}:=P_{\leq N}-P_{\leq M}$ whenever M and N are dyadic numbers. We will frequently write $f_{\leq N}$ for $P_{\leq N}f$ and similarly for the other operators. We recall the following standard Bernstein- and Sobolev-type inequalities.

LEMMA 2.3. *For any $1\leq p\leq q\leq\infty$ and $s>0$, we have*

$$\begin{aligned}\|P_{\geq N}f\|_{L_x^p}&\lesssim N^{-s}\||\nabla|^s P_{\geq N}f\|_{L_x^p},\\ \||\nabla|^s P_{\leq N}f\|_{L_x^p}&\lesssim N^s\|P_{\leq N}f\|_{L_x^p},\\ \||\nabla|^{\pm s}P_N f\|_{L_x^p}&\sim N^{\pm s}\|P_N f\|_{L_x^p},\\ \|P_{\leq N}f\|_{L_x^q}&\lesssim N^{\frac{d}{p}-\frac{d}{q}}\|P_{\leq N}f\|_{L_x^p},\\ \|P_N f\|_{L_x^q}&\lesssim N^{\frac{d}{p}-\frac{d}{q}}\|P_N f\|_{L_x^p}.\end{aligned}$$

For $N>1$, we define the Fourier multiplier I_N (cf. [10]) by

$$\widehat{I_N u}(\xi):=m_N(\xi)\hat u(\xi),$$

where m_N is a smooth radial decreasing function such that

$$m_N(\xi)=\begin{cases}1 & \text{if}\quad |\xi|\leq N,\\ \bigl(\frac{|\xi|}{N}\bigr)^{s-1} & \text{if}\quad |\xi|\geq 2N.\end{cases}$$

Thus, I_N is the identity operator on frequencies $|\xi|\leq N$ and behaves like a fractional integral operator of order $1-s$ on higher frequencies. In particular, I_N maps H_x^s to H_x^1; this allows us to access the theory available for H_x^1 data. We collect the basic properties of I_N into the following lemma.

LEMMA 2.4. *Let $1 < p < \infty$ and $0 \le \sigma \le s < 1$. Then*

$$\|I_N f\|_p \lesssim \|f\|_p, \tag{2.7}$$

$$\||\nabla|^\sigma P_{>N} f\|_p \lesssim N^{\sigma-1}\|\nabla I_N f\|_p, \tag{2.8}$$

$$\|f\|_{H^s_x} \lesssim \|I_N f\|_{H^1_x} \lesssim N^{1-s}\|f\|_{H^s_x}. \tag{2.9}$$

Proof. The estimate (2.7) is a direct consequence of the multiplier theorem.

To prove (2.8), we write

$$\||\nabla|^\sigma P_{>N} f\|_p = \|P_{>N}|\nabla|^\sigma(\nabla I_N)^{-1}\nabla I_N f\|_p.$$

The claim follows again from the multiplier theorem.

Now we turn to (2.9). By the definition of the operator I_N and (2.8),

$$\begin{aligned}\|f\|_{H^s_x} &\lesssim \|P_{\le N} f\|_{H^s_x} + \|P_{>N} f\|_2 + \||\nabla|^s P_{>N} f\|_2\\ &\lesssim \|P_{\le N} I_N f\|_{H^1_x} + N^{-1}\|\nabla I_N f\|_2 + N^{s-1}\|\nabla I_N f\|_2\\ &\lesssim \|I_N f\|_{H^1_x}.\end{aligned}$$

On the other hand, since the operator I_N commutes with $\langle\nabla\rangle^s$, we get

$$\|I_N f\|_{H^1_x} = \|\langle\nabla\rangle^{1-s} I_N \langle\nabla\rangle^s f\|_2 \lesssim N^{1-s}\|\langle\nabla\rangle^s f\|_2 \lesssim N^{1-s}\|f\|_{H^s_x},$$

which proves the last inequality in (2.9). Note that a similar argument also yields

$$\|I_N f\|_{\dot H^1_x} \lesssim N^{1-s}\|f\|_{\dot H^s_x}. \qquad \square \tag{2.10}$$

The estimate (2.8) shows that we can control the high frequencies of a function f in the Sobolev space $H^{\sigma,p}_x$ by the smoother function $I_N f$ in a space with a loss of derivative but a gain of negative power of N. This fact is crucial in extracting the negative power of N when estimating the increment of the modified Hamiltonian.

In one and two dimensions, one can use multilinear analysis to understand commutator expressions like $F(I_N u) - I_N F(u)$; on the Fourier side, one can expand this commutator into a product of Fourier transforms of u and $I_N u$ and carefully measure the frequency interactions to derive an estimate (see, for example, [13, 36]). However, this is not possible in dimensions $d \ge 3$. Instead, we will have to rely on the following rougher (weaker, but more robust) lemma.

LEMMA 2.5. *Let $1 < r, r_1, r_2 < \infty$ be such that $\frac{1}{r} = \frac{1}{r_1} + \frac{1}{r_2}$ and let $0 < \nu < s$. Then*

$$\|I_N(fg) - (I_N f)g\|_r \lesssim N^{-(1-s+\nu)}\|I_N f\|_{r_1}\|\langle\nabla\rangle^{1-s+\nu} g\|_{r_2}. \tag{2.11}$$

Proof. Applying a Littlewood–Paley decomposition to f and g, we write

$$\begin{aligned}I_N(fg) - (I_N f)g = I_N(f g_{\le 1}) - (I_N f) g_{\le 1} &+ \sum_{1<M\in 2^{\mathbb{Z}}}\big[I_N(f_{\lesssim M} g_M) - (I_N f_{\lesssim M}) g_M\big]\\ &+ \sum_{1<M\in 2^{\mathbb{Z}}}\big[I_N(f_{\gg M} g_M) - (I_N f_{\gg M}) g_M\big]\end{aligned}$$

$$
\begin{aligned}
&= I_N(f_{\gtrsim N} g_{\leq 1}) - (I_N f_{\gtrsim N}) g_{\leq 1} + \sum_{N \lesssim M \in 2^{\mathbb{Z}}} \big[I_N(f_{\lesssim M} g_M) - (I_N f_{\lesssim M}) g_M \big] \\
&\quad + \sum_{1 < M \in 2^{\mathbb{Z}}} \big[I_N(f_{\gg M} g_M) - (I_N f_{\gg M}) g_M \big] \\
&= I + II + III. \qquad (2.12)
\end{aligned}
$$

The second equality above follows from the fact that the operator I_N is the identity operator on frequencies $|\xi| \leq N$; thus,

$$
I_N(f_{\ll N} g_{\leq 1}) = (I_N f_{\ll N}) g_{\leq 1} \quad \text{and} \quad I_N(f_{\lesssim M} g_M) = (I_N f_{\lesssim M}) g_M \quad \forall\, M \ll N.
$$

We first consider II. Dropping the operator I_N, by Hölder and Bernstein we estimate

$$
\begin{aligned}
\| I_N(f_{\lesssim M} g_M) - (I_N f_{\lesssim M}) g_M \|_r &\lesssim \| f_{\lesssim M} \|_{r_1} \| g_M \|_{r_2} \\
&\lesssim \Big(\frac{M}{N} \Big)^{1-s} \| I_N f \|_{r_1} \| g_M \|_{r_2} \\
&\lesssim M^{-\nu} N^{-(1-s)} \| I_N f \|_{r_1} \| |\nabla|^{1-s+\nu} g \|_{r_2}.
\end{aligned}
$$

Summing over all M such that $N \lesssim M \in 2^{\mathbb{Z}}$, we get

$$
II \lesssim N^{-(1-s+\nu)} \| I_N f \|_{r_1} \| |\nabla|^{1-s+\nu} g \|_{r_2}. \qquad (2.13)
$$

We turn now towards III. Applying a Littlewood–Paley decomposition to f, we write each term in III as

$$
\begin{aligned}
I_N(f_{\gg M} g_M) - (I_N f_{\gg M}) g_M &= \sum_{1 \ll k \in \mathbb{N}} \big[I_N(f_{2^k M} g_M) - (I_N f_{2^k M}) g_M \big] \\
&= \sum_{\substack{1 \ll k \in \mathbb{N} \\ N \lesssim 2^k M}} \big[I_N(f_{2^k M} g_M) - (I_N f_{2^k M}) g_M \big].
\end{aligned}
$$

To derive the second inequality, we used again the fact that the operator I_N is the identity on frequencies $|\xi| \leq N$.

We write

$$
\big[I_N(f_{2^k M} g_M) - (I_N f_{2^k M}) g_M \big]\widehat{\ }(\xi) = \int_{\xi = \xi_1 + \xi_2} (m_N(\xi_1 + \xi_2) - m_N(\xi_1)) \widehat{f_{2^k M}}(\xi_1) \widehat{g_M}(\xi_2).
$$

For $|\xi_1| \sim 2^k M$, $k \gg 1$, and $|\xi_2| \sim M$, the fundamental theorem of calculus implies

$$
|m_N(\xi_1 + \xi_2) - m_N(\xi_1)| \lesssim 2^{-k} \Big(\frac{2^k M}{N} \Big)^{s-1}.
$$

By the Coifman–Meyer multilinear multiplier theorem [8, 9] and by Bernstein, we get

$$
\begin{aligned}
\| I_N(f_{2^k M} g_M) - (I_N f_{2^k M}) g_M \|_r &\lesssim 2^{-k} \Big(\frac{2^k M}{N} \Big)^{s-1} \| f_{2^k M} \|_{r_1} \| g_M \|_{r_2} \\
&\lesssim 2^{-k} M^{-(1-s+\nu)} \| I_N f \|_{r_1} \| |\nabla|^{1-s+\nu} g \|_{r_2}.
\end{aligned}
$$

Summing over M and k such that $N \lesssim 2^k M$, and recalling that $0 < \nu < s$, we get

$$III \lesssim N^{-(1-s+\nu)} \|I_N f\|_{r_1} \||\nabla|^{1-s+\nu} g\|_{r_2}. \tag{2.14}$$

To estimate I, we apply the same argument as for III. We get

$$\begin{aligned} I = \|I_N(f_{\gtrsim N} g_{\leq 1}) - (I_N f_{\gtrsim N}) g_{\leq 1}\|_r &\lesssim \sum_{k\in\mathbb{N}, 2^k \gtrsim N} \|I_N(f_{2^k} g_{\leq 1}) - (I_N f_{2^k}) g_{\leq 1}\|_r \\ &\lesssim \sum_{k\in\mathbb{N}, 2^k \gtrsim N} 2^{-k} \|I_N f\|_{r_1} \|g\|_{r_2} \\ &\lesssim N^{-1} \|I_N f\|_{r_1} \|g\|_{r_2}. \end{aligned} \tag{2.15}$$

Putting together (2.12) through (2.15), we derive (2.11). ◻

As an application of Lemma 2.5 we have the following commutator estimate.

LEMMA 2.6. *Let $1 < r, r_1, r_2 < \infty$ be such that $\frac{1}{r} = \frac{1}{r_1} + \frac{1}{r_2}$. Then for any $0 < \nu < s$ we have*

$$\|\nabla I_N F(u) - (I_N \nabla u) F'(u)\|_r \lesssim N^{-1+s-\nu} \|\nabla I_N u\|_{r_1} \|\langle\nabla\rangle^{1-s+\nu} F'(u)\|_{r_2}, \tag{2.16}$$

$$\|\nabla I_N F(u)\|_r \lesssim \|\nabla I_N u\|_{r_1} \|F'(u)\|_{r_2} + N^{-1+s-\nu} \|\nabla I_N u\|_{r_1} \|\langle\nabla\rangle^{1-s+\nu} F'(u)\|_{r_2}. \tag{2.17}$$

Proof. As

$$\nabla F(u) = \nabla u \cdot F'(u),$$

the estimate (2.16) follows immediately from Lemma 2.5 with $f := \nabla u$ and $g := F'(u)$. The estimate (2.17) is a consequence of (2.16) and the triangle inequality. ◻

Since we work at regularity $0 < s < 1$, we will need the following fractional chain rules to estimate our nonlinearity in H^s_x.

LEMMA 2.7 (fractional chain rule for a C^1 function). *Suppose that $F \in C^1(\mathbb{C})$, $\sigma \in (0,1)$, and $1 < r, r_1, r_2 < \infty$ such that $\frac{1}{r} = \frac{1}{r_1} + \frac{1}{r_2}$. Then*

$$\||\nabla|^\sigma F(u)\|_r \lesssim \|F'(u)\|_{r_1} \||\nabla|^\sigma u\|_{r_2}.$$

LEMMA 2.8 (fractional chain rule for a Lipschitz function). *Let F be a Lipschitz function, $\sigma \in (0,1)$, and $1 < r < \infty$. Then*

$$\||\nabla|^\sigma F(u)\|_r \lesssim \|F'\|_\infty \||\nabla|^\sigma u\|_r.$$

LEMMA 2.9 (fractional derivatives for fractional powers). *Let F be a Hölder continuous function of order $0 < \alpha < 1$. Then for every $0 < \sigma < \alpha$, $1 < r < \infty$, and $\frac{\sigma}{\alpha} < \delta < 1$ we have*

$$\big\||\nabla|^\sigma F(u)\big\|_r \lesssim \big\||u|^{\alpha - \frac{\sigma}{\delta}}\big\|_{r_1} \big\||\nabla|^\delta u\big\|^{\frac{\sigma}{\delta}}_{\frac{\sigma}{\delta} r_2}, \tag{2.18}$$

provided $\frac{1}{r} = \frac{1}{r_1} + \frac{1}{r_2}$ and $(1 - \frac{\sigma}{\alpha\delta}) r_1 > 1$.

The first two results originate in [7] and [32]; for a textbook treatment see [35]. The third result can be found in Appendix A of [38]. Using the chain rule estimates in these lemmas, we can upgrade the pointwise-in-time commutator estimate in Lemma 2.6 to a spacetime estimate.

Lemma 2.10. *Let I be a compact time interval and let $\frac{1}{1+\min\{1,\frac{4}{d}\}} < s < 1$. Then*

$$\|(I_N\nabla u)F'(u) - \nabla I_N F(u)\|_{L_t^2 L_x^{\frac{2d}{d+2}}(I\times\mathbb{R}^d)} \lesssim N^{-\min\{1,\frac{4}{d}\}s+}\|\langle\nabla\rangle I_N u\|_{S^0(I)}^{1+\frac{4}{d}}, \tag{2.19}$$

$$\|\langle\nabla\rangle I_N F(u)\|_{N^0(I)} \lesssim |I|^{\frac{2s}{d}}\|\langle\nabla\rangle I_N u\|_{S^0(I)}^{1+\frac{4}{d}} + N^{-\min\{1,\frac{4}{d}\}s+}\|\langle\nabla\rangle I_N u\|_{S^0(I)}^{1+\frac{4}{d}}. \tag{2.20}$$

Proof. Throughout the proof, all spacetime norms will be taken on the slab $I\times\mathbb{R}^d$.

As by hypothesis $s > \frac{1}{1+\min\{1,\frac{4}{d}\}}$, there exists $\varepsilon_0 > 0$ such that for any $0<\varepsilon<\varepsilon_0$ we have $s > \frac{1+\varepsilon}{1+\min\{1,\frac{4}{d}\}}$. Let $\nu := \min\{1,\frac{4}{d}\}s - (1-s) - \varepsilon$ with $0<\varepsilon<\varepsilon_0$. It is easy to check that $0<\nu<s$. Applying Lemma 2.6 with this value of ν, we get

$$\begin{aligned}\|(\nabla I_N u)F'(u) - &\nabla I_N F(u)\|_{2,\frac{2d}{d+2}}\\ &\lesssim N^{-\min\{1,\frac{4}{d}\}s+\varepsilon}\|\nabla I_N u\|_{2,\frac{2d}{d-2}}\|\langle\nabla\rangle^{\min\{1,\frac{4}{d}\}s-\varepsilon}F'(u)\|_{\infty,\frac{d}{2}}.\end{aligned}$$

The claim (2.19) will follow immediately from the estimate above, provided we show

$$\|\langle\nabla\rangle^{\min\{1,\frac{4}{d}\}s-\varepsilon}F'(u)\|_{\infty,\frac{d}{2}} \lesssim \|\langle\nabla\rangle I_N u\|_{\infty,2}^{\frac{4}{d}}. \tag{2.21}$$

We start by observing that for any (q,r) admissible pair,

$$\|\langle\nabla\rangle^s u\|_{q,r} \lesssim \|\langle\nabla\rangle I_N u\|_{q,r}. \tag{2.22}$$

Indeed, decomposing $u := u_{\leq N} + u_{>N}$ and using Lemma 2.4 and the fact that I_N is the identity operator on frequencies $|\xi|\leq N$, we get

$$\begin{aligned}\|u\|_{q,r} &\leq \|u_{\leq N}\|_{q,r} + \|u_{>N}\|_{q,r}\\ &\lesssim \|I_N u_{\leq N}\|_{q,r} + N^{-1}\|\nabla I_N u_{>N}\|_{q,r}\\ &\lesssim \|\langle\nabla\rangle I_N u\|_{q,r}.\end{aligned}$$

Similarly, we estimate

$$\begin{aligned}\||\nabla|^s u\|_{q,r} &\leq \||\nabla|^s u_{\leq N}\|_{q,r} + \||\nabla|^s u_{>N}\|_{q,r}\\ &\lesssim \||\nabla|^s I_N u_{\leq N}\|_{q,r} + N^{s-1}\|\nabla I_N u\|_{q,r}\\ &\lesssim \|\langle\nabla\rangle I_N u\|_{q,r},\end{aligned}$$

and the estimate (2.22) follows.

As $F'(u) = O(|u|^{\frac{4}{d}})$, by (2.22) we get

$$\|F'(u)\|_{\infty,\frac{d}{2}} \lesssim \|\langle\nabla\rangle I_N u\|_{\infty,2}^{\frac{4}{d}}. \tag{2.23}$$

We first prove (2.21) for $d\leq 4$. By (2.23), we estimate

$$\begin{aligned}\|\langle\nabla\rangle^{\min\{1,\frac{4}{d}\}s-\varepsilon}F'(u)\|_{\infty,\frac{d}{2}} &= \|\langle\nabla\rangle^{s-\varepsilon}F'(u)\|_{\infty,\frac{d}{2}}\\ &\lesssim \|\langle\nabla\rangle^s F'(u)\|_{\infty,\frac{d}{2}}\\ &\lesssim \|F'(u)\|_{\infty,\frac{d}{2}} + \||\nabla|^s F'(u)\|_{\infty,\frac{d}{2}}\\ &\lesssim \|\langle\nabla\rangle I_N u\|_{\infty,2}^{\frac{4}{d}} + \||\nabla|^s F'(u)\|_{\infty,\frac{d}{2}}.\end{aligned}$$

Using Lemmas 2.7 (for $d = 3$) and 2.8 (for $d = 4$) together with (2.22), we estimate

$$\||\nabla|^s F'(u)\|_{\infty,\frac{d}{2}} \lesssim \||\nabla|^s u\|_{\infty,2} \|u\|_{\infty,2}^{\frac{4}{d}-1} \lesssim \|\langle\nabla\rangle I_N u\|_{\infty,2}^{\frac{4}{d}}.$$

Thus, (2.21) holds for $d \leq 4$.

If $d > 4$, using Lemma 2.9 (with $\alpha := \frac{4}{d}$, $\sigma := \frac{4s}{d} - \varepsilon$, and $\delta := s$) and (2.22), we get

$$\begin{aligned}\||\nabla|^{\min\{1,\frac{4}{d}\}s-\varepsilon} F'(u)\|_{\infty,\frac{d}{2}} = \||\nabla|^{\frac{4s}{d}-\varepsilon} F'(u)\|_{\infty,\frac{d}{2}} &\lesssim \|u\|_{\infty,2}^{\frac{\varepsilon}{s}} \||\nabla|^s u\|_{\infty,2}^{\frac{4}{d}-\frac{\varepsilon}{s}} \\ &\lesssim \|\langle\nabla\rangle I_N u\|_{\infty,2}^{\frac{4}{d}}.\end{aligned}$$

From this and (2.23) we derive (2.21) in the case $d > 4$. The claim (2.19) follows.

We now consider (2.20). Using (2.5), (2.7), and (2.22), we obtain

$$\|I_N F(u)\|_{N^0(I)} \lesssim \|F(u)\|_{\gamma',\rho'} \lesssim |I|^{\frac{2s}{d}} \|\langle\nabla\rangle^s u\|_{\gamma,\rho}^{1+\frac{4}{d}} \lesssim |I|^{\frac{2s}{d}} \|\langle\nabla\rangle I_N u\|_{\gamma,\rho}^{1+\frac{4}{d}}.$$

Similarly, by Hölder and (2.22),

$$\begin{aligned}\|(I_N \nabla u) F'(u)\|_{N^0(I)} &\lesssim \|(I_N \nabla u) F'(u)\|_{\gamma',\rho'} \\ &\lesssim |I|^{\frac{2s}{d}} \|I_N \nabla u\|_{\gamma,\rho} \|u\|_{\gamma,\rho^*}^{\frac{4}{d}} \\ &\lesssim |I|^{\frac{2s}{d}} \|I_N \nabla u\|_{\gamma,\rho} \|\langle\nabla\rangle^s u\|_{\gamma,\rho}^{\frac{4}{d}} \\ &\lesssim |I|^{\frac{2s}{d}} \|\langle\nabla\rangle I_N u\|_{\gamma,\rho}^{1+\frac{4}{d}}.\end{aligned}$$

The estimate (2.20) follows from the estimates above, from (2.19), and from the triangle inequality. $\square$

We end this section with the following concentration-compactness lemma.

LEMMA 2.11 (concentration-compactness [20]). *Let $\{v_n\}_{n\geq 1}$ be a bounded sequence in $H^1(\mathbb{R}^d)$ such that*

$$\limsup_{n\to\infty} \|\nabla v_n\|_2 \leq M < \infty$$

and

$$\limsup_{n\to\infty} \|v_n\|_{2+\frac{4}{d}} \geq m > 0.$$

Then there exists $\{x_n\}_{n\geq 1} \subset \mathbb{R}^d$ such that, up to a subsequence,

$$v_n(\cdot + x_n) \rightharpoonup V \quad \textit{weakly in } H_x^1 \textit{ as } n \to \infty$$

with $\|V\|_2 \geq \sqrt{\frac{d}{d+2}\frac{m^{\frac{2}{d}+1}}{M}} \|Q\|_2$.

3. Local H_x^s theory. In this section, we review the local H_x^s theory for (1.1) as described in [6].

PROPOSITION 3.1 (local well-posedness in H_x^s [6]). *Let $0 < s < 1$ and $u_0 \in H^s(\mathbb{R}^d)$. Then (1.1) is well posed on $[0, T_{lwp}]$ with*

$$T_{lwp} = C_0 \|\langle\nabla\rangle^s u_0\|_2^{-\frac{2}{s}}.$$

Moreover, the unique solution u enjoys the following estimate:

$$\|\langle\nabla\rangle^s u\|_{S^0([0,T_{lwp}])} \lesssim \|\langle\nabla\rangle^s u_0\|_2.$$

Here, C_0 and the implicit constant depend only on dimension d and regularity s.

As a direct consequence of the H^s_x local well-posedness result, we have the following lower bound on the blowup rate of the H^s_x-norm.

COROLLARY 3.2 (blowup criterion [6]). *Let $0 < s < 1$ and $u_0 \in H^s_x$. Assume that the unique solution u to (1.1) blows up at time $0 < T^* < \infty$. Then there exists a constant C depending only on d and s such that*

$$\|u(t)\|_{H^s_x} \geq C(T^* - t)^{-\frac{s}{2}}.$$

As a variation of Proposition 3.1, we have the following.

PROPOSITION 3.3 ("modified" local well-posedness). *Let $\frac{\frac{4}{d}+1}{1+\min\{1,\frac{4}{d}\}+\frac{4}{d}} < s < 1$ and $u_0 \in H^s_x$. Let also*

$$N \gg \|u_0\|_{H^s_x}^{\frac{4}{\min\{4,d\}s-(4+d)(1-s)-d\varepsilon}} \quad \textit{for any } \varepsilon > 0 \textit{ sufficiently small,} \tag{3.1}$$

$$\tilde{T}_{lwp} := c_0\|\langle\nabla\rangle I_N u_0\|_2^{-\frac{2}{s}} \quad \textit{for a small constant } c_0 = c_0(d,s). \tag{3.2}$$

Then (1.1) is well posed on $[0,\tilde{T}_{lwp}]$, and moreover

$$\|\langle\nabla\rangle I_N u\|_{S^0([0,\tilde{T}_{lwp}])} \lesssim \|\langle\nabla\rangle I_N u_0\|_2. \tag{3.3}$$

Proof. As by Lemma 2.4,

$$\|u_0\|_{H^s_x} \lesssim \|\langle\nabla\rangle I_N u_0\|_2,$$

choosing c_0 sufficiently small, we get

$$\tilde{T}_{lwp} \leq T_{lwp}.$$

Thus, (1.1) is well posed in H^s_x on $[0,\tilde{T}_{lwp}]$. Let u be the unique solution; by Proposition 3.1 we have

$$\|\langle\nabla\rangle^s u\|_{S^0([0,\tilde{T}_{lwp}])} \lesssim \|u_0\|_{H^s_x}. \tag{3.4}$$

On the other hand, by Strichartz, for any $t \leq \tilde{T}_{lwp}$ we estimate

$$\|\langle\nabla\rangle I_N u\|_{S^0([0,t])} \lesssim \|\langle\nabla\rangle I_N u_0\|_2 + \|\langle\nabla\rangle I_N F(u)\|_{N^0([0,t])}.$$

As by hypothesis, $s > \frac{\frac{4}{d}+1}{1+\min\{1,\frac{4}{d}\}+\frac{4}{d}} > \frac{1}{1+\min\{1,\frac{4}{d}\}}$, by Lemma 2.10 we get

$$\begin{aligned}\|\langle\nabla\rangle I_N u\|_{S^0([0,t])} \lesssim{}& \|\langle\nabla\rangle I_N u_0\|_2 + t^{\frac{2s}{d}}\|\langle\nabla\rangle I_N u\|_{S^0([0,t])}^{1+\frac{4}{d}} \\ &+ N^{-\min\{1,\frac{4}{d}\}s+\varepsilon}\|\langle\nabla\rangle I_N u\|_{S^0([0,t])}^{1+\frac{4}{d}}\end{aligned} \tag{3.5}$$

for any $\varepsilon > 0$ sufficiently small. For N satisfying (3.1), we use Lemma 2.4 and (3.4) to estimate

$$
\begin{aligned}
N^{-\min\{1,\frac{4}{d}\}s+\varepsilon}&\|\langle\nabla\rangle I_N u\|_{S^0([0,t])}^{1+\frac{4}{d}}\\
&\lesssim N^{-\min\{1,\frac{4}{d}\}s+\varepsilon}N^{(1+\frac{4}{d})(1-s)}\|\langle\nabla\rangle^s u\|_{S^0([0,t])}^{1+\frac{4}{d}}\\
&\ll \|u_0\|_{H_x^s}^{\frac{4}{\min\{4,d\}s-(4+d)(1-s)-d\varepsilon}\times[-\min\{1,\frac{4}{d}\}s+\varepsilon+(1+\frac{4}{d})(1-s)]}\|u_0\|_{H_x^s}^{1+\frac{4}{d}}\\
&\lesssim \|u_0\|_{H_x^s}^{-\frac{4}{d}}\|u_0\|_{H_x^s}^{1+\frac{4}{d}}\lesssim \|\langle\nabla\rangle I_N u_0\|_2.
\end{aligned}
$$

Here, we have used the fact that $s > \frac{\frac{4}{d}+1}{1+\min\{1,\frac{4}{d}\}+\frac{4}{d}}$ implies that there exists $\varepsilon_1 > 0$ sufficiently small such that for any $0 < \varepsilon < \varepsilon_1$ we have $s > \frac{\frac{4}{d}+1+\varepsilon}{1+\min\{1,\frac{4}{d}\}+\frac{4}{d}}$; thus the power of N in the estimates above is negative for $0 < \varepsilon < \varepsilon_1$.

Returning to (3.5), we conclude that

$$\|\langle\nabla\rangle I_N u\|_{S^0([0,t])} \lesssim \|\langle\nabla\rangle I_N u_0\|_2 + t^{\frac{2s}{d}}\|\langle\nabla\rangle I_N u\|_{S^0([0,t])}^{1+\frac{4}{d}}.$$

Standard arguments yield (3.3), provided

$$t \le \tilde{T}_{lwp} = c_0(d,s)\|\langle\nabla\rangle I_N u_0\|_2^{-\frac{2}{s}}. \qquad \square$$

4. Modified energy increment. The main purpose of this section is to prove that the modified energy of u, $E(I_N u)$, grows much slower than the modified kinetic energy of u, $\|\nabla I_N u\|_2^2$. As will be shown later, this result is crucial in establishing the main theorems.

Before stating the result, we need to introduce more notation. We define

$$\Lambda(t) := \sup_{0\le\tau\le t}\|u(\tau)\|_{H_x^s} \quad \text{and} \quad \Sigma(t) := \sup_{0\le\tau\le t}\|I_N u(\tau)\|_{H_x^1}.$$

With this notation we have the following.

PROPOSITION 4.1 (increment of the modified energy). *Let $s_0(d) < s < 1$ and let $u_0 \in H_x^s$ such that the corresponding solution u to* (1.1) *blows up at time $0 < T^* < \infty$. Let $0 < T < T^*$. Then for*

$$N(T) := C\Lambda(T)^{\frac{p(s)}{2(1-s)}}, \tag{4.1}$$

we have

$$|E(I_{N(T)}u(T))| \lesssim \Lambda(T)^{p(s)}.$$

Here, C and the implicit constant depend only on s, T^*, and $\|u_0\|_{H_x^s}$, and $p(s)$ is given by

$$p(s) := \frac{2(2+\frac{2}{s}+\frac{8}{d})(1-s)}{\min\{1,\frac{4}{d}\}s-(\frac{2}{s}+\frac{8}{d})(1-s)-}.$$

Note that by Lemma 2.4, $\Lambda(T) \lesssim \Sigma(T)$. Thus, if the solution blows up at time T^*, the modified energy $E(I_{N(T)}u(T))$ is at most $O(\Sigma(T)^{p(s)})$, which is much smaller than the modified kinetic energy, $\|\nabla I_{N(T)}u(T)\|_2^2 = O(\Sigma(T)^2)$ for $s > s_0(d)$.

We prove Proposition 4.1 in two steps. The first step is to control the increment of the modified energy of u on intervals of local well-posedness $[0,\tilde{T}_{lwp}]$. The second step is to divide the interval $[0,T]$ into finitely many subintervals of local well-posedness, control the increment of the modified energy of u on each of these subintervals, and sum these bounds.

We start with the following.

LEMMA 4.2 (local increment of the modified energy). *Let* $\frac{1+\frac{4}{d}}{1+\frac{4}{d}+\min\{1,\frac{4}{d}\}} < s < 1$ *and let* $u_0 \in H_x^s$. *Assume that* N *and* $\tilde{T}_{lwp}$ *satisfy* (3.1) *and* (3.2), *respectively, that is,*

$$N \gg \|u_0\|_{H_x^s}^{\frac{4}{\min\{4,d\}s-(4+d)(1-s)-}},$$
$$\tilde{T}_{lwp} := c_0 \|\langle\nabla\rangle I_N u_0\|_2^{-\frac{2}{s}} \quad \text{for a small constant } c_0 = c_0(d,s).$$

Then

$$\sup_{t\in[0,\tilde{T}_{lwp}]} |E(I_N u(t))| \leq |E(I_N u_0)| + CN^{-\min\{1,\frac{4}{d}\}s+}(\|I_N\langle\nabla\rangle u_0\|_2^{2+\frac{4}{d}} + \|I_N\langle\nabla\rangle u_0\|_2^{2+\frac{8}{d}}).$$

Here, the constant C *depends on* s, T^*, *and* $\|u_0\|_{H_x^s}$.

Proof. Note that by Proposition 3.3, (1.1) is well posed on $[0,\tilde{T}_{lwp}]$. Furthermore, the unique solution u to (1.1) on $[0,\tilde{T}_{lwp}]$ satisfies

$$\|\langle\nabla\rangle I_N u\|_{S^0([0,\tilde{T}_{lwp}])} \lesssim \|\langle\nabla\rangle I_N u_0\|_2. \tag{4.2}$$

Let $0 < t \leq \tilde{T}_{lwp}$; throughout the rest of the proof, all spacetime norms will be taken on $[0,t]\times\mathbb{R}^d$. By the fundamental theorem of calculus, we can write the modified energy increment as

$$\begin{aligned} E(I_N u(t)) - E(I_N u_0) &= \int_0^t \frac{\partial}{\partial s} E(I_N u(s))\, ds \\ &= \mathrm{Re} \int_0^t \int_{\mathbb{R}^d} \overline{I_N u_t}(-\Delta I_N u + F(I_N u))\, dx\, ds. \end{aligned}$$

As $I_N u_t = i\Delta I_N u - iI_N F(u)$, we have

$$\mathrm{Re} \int_0^t \int_{\mathbb{R}^d} \overline{I_N u_t}(-\Delta I_N u + I_N F(u))\, dx\, ds = 0.$$

Thus, after an integration by parts,

$$E(I_N u(t)) - E(I_N u_0) = \mathrm{Re} \int_0^t \int_{\mathbb{R}^d} \overline{I_N u_t}[F(I_N u) - I_N F(u)]\, dx\, ds$$
$$= -\mathrm{Im} \int_0^t \int_{\mathbb{R}^d} \overline{\nabla I_N u} \cdot \nabla[F(I_N u) - I_N F(u)]\, dx\, ds \tag{4.3}$$
$$- \mathrm{Im} \int_0^t \int_{\mathbb{R}^d} \overline{I_N F(u)} \cdot [F(I_N u) - I_N F(u)]\, dx\, ds. \tag{4.4}$$

Consider the contribution from (4.3). By the triangle inequality,

$$\begin{aligned} \|\nabla[F(I_N u) - I_N F(u)]\|_{2,\frac{2d}{d+2}} &\lesssim \|(\nabla I_N u)[F'(I_N u) - F'(u)]\|_{2,\frac{2d}{d+2}} \\ &\quad + \|(\nabla I_N u)F'(u) - \nabla I_N F(u)\|_{2,\frac{2d}{d+2}}. \end{aligned}$$

By Hölder, (2.1), (2.8), and (2.22), we estimate

$$\begin{aligned}
&\|(\nabla I_N u)[F'(I_N u) - F'(u)]\|_{2,\frac{2d}{d+2}} \\
&\qquad \lesssim \|\nabla I_N u\|_{2,\frac{2d}{d-2}} \|F'(I_N u) - F'(u)\|_{\infty,\frac{d}{2}} \\
&\qquad \lesssim \|\langle\nabla\rangle I_N u\|_{S^0([0,t])} \big\| |I_N u - u|^{\min\{1,\frac{4}{d}\}} (|I_N u| + |u|)^{\frac{4}{d}-\min\{1,\frac{4}{d}\}} \big\|_{\infty,\frac{d}{2}} \\
&\qquad \lesssim \|\langle\nabla\rangle I_N u\|_{S^0([0,t])} \|P_{>N} u\|_{\infty,2}^{\min\{1,\frac{4}{d}\}} \|u\|_{\infty,2}^{\frac{4}{d}-\min\{1,\frac{4}{d}\}} \\
&\qquad \lesssim N^{-\min\{1,\frac{4}{d}\}} \|\langle\nabla\rangle I_N u\|_{S^0([0,t])}^{1+\frac{4}{d}}.
\end{aligned}$$

Combining this with (2.19), we get

$$\|\nabla[F(I_N u) - I_N F(u)]\|_{2,\frac{2d}{d+2}} \lesssim N^{-\min\{1,\frac{4}{d}\}s+} \|\langle\nabla\rangle I_N u\|_{S^0([0,t])}^{1+\frac{4}{d}}. \tag{4.5}$$

Therefore

$$\begin{aligned}
|(4.3)| &\lesssim \|\nabla I_N u\|_{2,\frac{2d}{d-2}} \|\nabla[F(I_N u) - I_N F(u)]\|_{2,\frac{2d}{d+2}} \\
&\lesssim N^{-\min\{1,\frac{4}{d}\}s+} \|\langle\nabla\rangle I_N u\|_{S^0([0,t])}^{2+\frac{4}{d}}. \qquad (4.6)
\end{aligned}$$

We turn now toward (4.4). By (4.5) and Sobolev embedding, we estimate

$$\begin{aligned}
|(4.4)| &\lesssim \|\nabla[F(I_N u) - I_N F(u)]\|_{2,\frac{2d}{d+2}} \||\nabla|^{-1} I_N F(u)\|_{2,\frac{2d}{d-2}} \\
&\lesssim N^{-\min\{1,\frac{4}{d}\}s+} \|\langle\nabla\rangle I_N u\|_{S^0([0,t])}^{1+\frac{4}{d}} \|I_N F(u)\|_{2,2}. \qquad (4.7)
\end{aligned}$$

To estimate the last factor in (4.7), we drop the operator I_N and use Sobolev embedding to obtain

$$\|I_N F(u)\|_{2,2} \lesssim \|u\|_{\frac{2(d+4)}{d},\frac{2(d+4)}{d}}^{1+\frac{4}{d}} \lesssim \||\nabla|^{\frac{d}{d+4}} u\|_{\frac{2(d+4)}{d},\frac{2(d+4)}{d+2}}^{1+\frac{4}{d}}.$$

Note that $(\frac{2(d+4)}{d}, \frac{2(d+4)}{d+2})$ is a Schrödinger-admissible pair. Decompose $u := u_{\leq N} + u_{>N}$. To estimate the low frequencies, we use the fact that I_N is the identity on frequencies $|\xi| \leq N$:

$$\||\nabla|^{\frac{d}{d+4}} u_{\leq N}\|_{\frac{2(d+4)}{d},\frac{2(d+4)}{d+2}} \lesssim \|\langle\nabla\rangle I_N u\|_{S^0([0,t])}.$$

For the high frequencies, we use Lemma 2.4 to get

$$\||\nabla|^{\frac{d}{d+4}} u_{>N}\|_{\frac{2(d+4)}{d},\frac{2(d+4)}{d+2}} \lesssim N^{-\frac{4}{d+4}} \|\nabla I_N u_{>N}\|_{S^0([0,t])},$$

provided $s > \frac{d}{d+4}$; this condition is satisfied since by assumption,

$$s > \frac{1+\frac{4}{d}}{1+\frac{4}{d}+\min\{1,\frac{4}{d}\}} > \frac{d}{d+4}.$$

Therefore,

$$\|I_N F(u)\|_{2,2} \lesssim \|\langle\nabla\rangle I_N u\|_{S^0([0,t])}^{1+\frac{4}{d}}. \tag{4.8}$$

By (4.7) and (4.8), we obtain

$$|(4.4)| \lesssim N^{-\min\{1,\frac{4}{d}\}s+} \|\langle\nabla\rangle I_N u\|_{S^0([0,t])}^{2+\frac{8}{d}}. \tag{4.9}$$

Collecting (4.2), (4.6), and (4.9), we get

$$|E(I_N u(t)) - E(I_N u_0)| \lesssim N^{-\min\{1,\frac{4}{d}\}s+} (\|\langle\nabla\rangle I_N u_0\|_2^{2+\frac{4}{d}} + \|\langle\nabla\rangle I_N u_0\|_2^{2+\frac{8}{d}}).$$

This proves Lemma 4.2. ☐

Next, we use Lemma 4.2 to prove Proposition 4.1.

Let $T < T^*$ and $\Lambda(T)$ and $\Sigma(T)$ be defined as in the beginning of this section. By Proposition 3.3, if we take

$$\begin{cases} N(T) \gg \Lambda(T)^{\frac{4}{\min\{4,d\}s-(4+d)(1-s)-}}, \\ \delta := c_0 \Sigma(T)^{-\frac{2}{s}}, \end{cases} \tag{4.10}$$

then the solution u satisfies the estimate

$$\|\langle\nabla\rangle I_{N(T)} u\|_{S^0([t,t+\delta])} \lesssim \|\langle\nabla\rangle I_{N(T)} u(t)\|_2 \lesssim \Sigma(T),$$

uniformly in t, provided $[t, t+\delta] \subset [0,T]$. Thus, splitting $[0,T]$ into $O(\frac{T}{\delta})$ subintervals and applying Lemma 4.2 on each of these subintervals, we get

$$\begin{aligned} \sup_{t\in[0,T]} |E(I_{N(T)} u(t))| &\lesssim |E(I_{N(T)} u_0)| + \frac{T}{\delta} N(T)^{-\min\{1,\frac{4}{d}\}s+} \Sigma(T)^{2+\frac{4}{d}} \\ &\quad + \frac{T}{\delta} N(T)^{-\min\{1,\frac{4}{d}\}s+} \Sigma(T)^{2+\frac{8}{d}} \\ &\lesssim |E(I_{N(T)} u_0)| + N(T)^{-\min\{1,\frac{4}{d}\}s+} \Sigma(T)^{2+\frac{4}{d}+\frac{2}{s}} \\ &\quad + N(T)^{-\min\{1,\frac{4}{d}\}s+} \Sigma(T)^{2+\frac{8}{d}+\frac{2}{s}}. \end{aligned} \tag{4.11}$$

Using interpolation, Sobolev embedding, and Lemma 2.4, we estimate

$$\begin{aligned} |E(I_{N(T)} u_0)| &\lesssim \|\nabla I_N u_0\|_2^2 + \|I_N u_0\|_{2+\frac{4}{d}}^{2+\frac{4}{d}} \\ &\lesssim N^{2(1-s)} \|u_0\|_{H_x^s}^2 + \|I_N u_0\|_2^{\frac{4}{d}} \|\nabla I_N u_0\|_2^2 \\ &\lesssim N^{2(1-s)} \big(\|u_0\|_{H_x^s}^2 + \|u_0\|_{H_x^s}^{2+\frac{4}{d}} \big) \\ &\lesssim N^{2(1-s)}. \end{aligned} \tag{4.12}$$

Moreover, by Lemma 2.4, we also have

$$\Sigma(T) \lesssim N(T)^{1-s} \Lambda(T). \tag{4.13}$$

Substituting (4.12) and (4.13) into (4.11), we obtain

$$\begin{aligned} \sup_{t\in[0,T]} |E(I_{N(T)} u(t))| &\lesssim N(T)^{2(1-s)} + N(T)^{-\min\{1,\frac{4}{d}\}s+(2+\frac{4}{d}+\frac{2}{s})(1-s)+} \Lambda(T)^{2+\frac{4}{d}+\frac{2}{s}} \\ &\quad + N(T)^{-\min\{1,\frac{4}{d}\}s+(2+\frac{8}{d}+\frac{2}{s})(1-s)+} \Lambda(T)^{2+\frac{8}{d}+\frac{2}{s}}. \end{aligned} \tag{4.14}$$

Optimizing (4.14), we observe that if

$$N(T) \sim \Lambda(T)^{\frac{2+\frac{8}{d}+\frac{2}{s}}{\min\{1,\frac{4}{d}\}s-(1-s)(\frac{8}{d}+\frac{2}{s})-}}, \tag{4.15}$$

then $N(T)$ satisfies the assumption (4.10), and moreover

$$\sup_{t\in[0,T]} |E(I_{N(T)}u(t))| \lesssim N(T)^{2(1-s)}$$
$$\lesssim \Lambda(T)^{\frac{2(2+\frac{8}{d}+\frac{2}{s})(1-s)}{\min\{1,\frac{4}{d}\}s-(1-s)(\frac{8}{d}+\frac{2}{s})-}}.$$

Let

$$p(s) := \frac{2(2+\frac{8}{d}+\frac{2}{s})(1-s)}{\min\{1,\frac{4}{d}\}s-(1-s)(\frac{8}{d}+\frac{2}{s})-}.$$

Then a little work shows that the condition $0 < p(s) < 2$ leads to the restriction

$$s > s_0(d).$$

Thus, for $N(T)$ defined in (4.15) and $s > s_0(d)$, we have $0 < p(s) < 2$ and

$$\sup_{t\in[0,T]} |E(I_{N(T)}u)(t)| \lesssim \Lambda(T)^{p(s)}.$$

This proves Proposition 4.1.

5. Proof of Theorem 1.1. In this section, we use Proposition 4.1 together with Lemma 2.11 to prove Theorem 1.1.

We choose a sequence of times $\{t_n\}_{n\geq 1}$, such that $t_n \to T^*$ as $n \to \infty$ and

$$\|u(t_n)\|_{H_x^s} = \Lambda(t_n).$$

As the solution u blows up at time T^*, we must have $\Lambda(t_n) \to \infty$ as $n \to \infty$.

Set

$$\psi_n(x) := \rho_n^{\frac{d}{2}}(I_{N(t_n)}u)(t_n, \rho_n x),$$

where $N(t_n)$ is given by (4.1) with $T := t_n$ and the parameter ρ_n is given by

$$\rho_n := \frac{\|\nabla Q\|_2}{\|\nabla I_{N(t_n)}u(t_n)\|_2}.$$

By Lemma 2.4 and Corollary 3.2, we get

$$\rho_n \lesssim \frac{1}{\|u(t_n)\|_{H_x^s}} \lesssim (T^* - t_n)^{\frac{s}{2}}.$$

Basic calculations show that $\{\psi_n\}_{n\geq 1}$ is a bounded sequence in H_x^1. Indeed,

$$\begin{aligned} \|\psi_n\|_2 &= \|I_{N(t_n)}u(t_n)\|_2 \leq \|u(t_n)\|_2 = \|u_0\|_2, \\ \|\nabla\psi_n\|_2 &= \rho_n\|I_{N(t_n)}\nabla u(t_n)\|_2 = \|\nabla Q\|_2. \end{aligned} \tag{5.1}$$

By Proposition 4.1 (with $T = t_n$), we can estimate the energy of ψ_n as follows:

$$E(\psi_n) = \rho_n^2 E(I_{N(t_n)}u(t_n)) \lesssim \rho_n^2\Lambda(t_n)^{p(s)} \lesssim \|u(t_n)\|_{H_x^s}^{p(s)-2}.$$

Thus, as $p(s) < 2$ for $s > s_0(d)$,

$$E(\psi_n) \to 0 \quad \text{as} \quad n \to \infty,$$

which by the definition of the energy and (5.1) implies

$$\|\psi_n\|_{2+\frac{4}{d}}^{2+\frac{4}{d}} \to \frac{d+2}{d}\|\nabla Q\|_2^2 \quad \text{as } n\to\infty. \tag{5.2}$$

Applying Lemma 2.11 to the sequence $\{\psi_n\}_{n\geq 1}$ (with $M := \|\nabla Q\|_2$ and $m := (\frac{d+2}{d}\|\nabla Q\|_2^2)^{\frac{d}{2d+4}}$), we derive the existence of a sequence $\{x_n\}_{n\geq 1}\subset \mathbb{R}^d$ and of a function $V\in H^1(\mathbb{R}^d)$ such that $\|V\|_2 \geq \|Q\|_2$ and, up to a subsequence,

$$\psi_n(\cdot + x_n) \rightharpoonup V \quad \text{weakly in } H_x^1 \quad \text{as } n\to\infty,$$

that is,

$$\rho_n^{\frac{d}{2}}(I_{N(t_n)}u)(t_n, \rho_n\cdot + x_n) \rightharpoonup V \quad \text{weakly in } H_x^1 \quad \text{as } n\to\infty. \tag{5.3}$$

To prove Theorem 1.1, we have to eliminate the smoothing operator $I_{N(t_n)}$ from (5.3). We do so at the expense of trading the weak convergence in H_x^1 for convergence in the sense of distributions. Indeed, for any $\sigma < s$ we have

$$\begin{aligned}
\|\rho_n^{\frac{d}{2}}\big(u(t_n) - I_{N(t_n)}u(t_n)\big)(\rho_n\cdot + x_n)\|_{\dot H_x^\sigma} &= \rho_n^\sigma \|P_{\geq N(t_n)}u(t_n)\|_{\dot H_x^\sigma}\\
&\lesssim \rho_n^\sigma N(t_n)^{\sigma - s}\|P_{\geq N(t_n)}u(t_n)\|_{\dot H_x^s}\\
&\lesssim \Lambda(t_n)^{-\sigma}\Lambda(t_n)^{\frac{(\sigma-s)p(s)}{2(1-s)}}\|P_{\geq N(t_n)}u(t_n)\|_{H_x^s}\\
&\lesssim \Lambda(t_n)^{1-\sigma+\frac{(\sigma-s)p(s)}{2(1-s)}}.
\end{aligned}\tag{5.4}$$

Plugging the explicit expression for $p(s)$ into the above computation, we find that for

$$\sigma < \tilde s := \frac{2d+8s+s^2 d(2-\min\{1,\frac{4}{d}\})}{4d+16s - s^2(d\min\{1,\frac{4}{d}\}+8)},$$

the exponent of $\Lambda(t_n)$ in (5.4) is negative. Hence,

$$\|\rho_n^{\frac{d}{2}}\big(u(t_n) - I_{N(t_n)}u(t_n)\big)(\rho_n\cdot + x_n)\|_{H_x^{\tilde s -}} \to 0 \quad \text{as } n\to\infty. \tag{5.5}$$

Combining (5.3) and (5.5) finishes the proof of Theorem 1.1.

6. Proof of Theorem 1.2. By Theorem 1.1, there exists a blowup profile $V\in H_x^1$, with $\|V\|_2\geq \|Q\|_2$, and there exist sequences $\{t_n,\rho_n,x_n\}_{n\geq1}\subset \mathbb{R}_+\times\mathbb{R}_+^*\times\mathbb{R}^d$ such that $t_n\to T^*$,

$$\frac{\rho_n}{(T^*-t_n)^{\frac{s}{2}}} \lesssim 1 \quad \forall\, n\geq 1, \tag{6.1}$$

and

$$\rho_n^{\frac{d}{2}} u(t_n, \rho_n\cdot + x_n) \rightharpoonup V \quad \text{weakly as } n\to\infty. \tag{6.2}$$

From (6.2) it follows that for any $R>0$ we have

$$\liminf_{n\to\infty} \rho_n^d \int_{|x|\leq R} |u(t_n,\rho_n x + x_n)|^2 \geq \int_{|x|\leq R} |V|^2 dx,$$

which, by a change of variables, yields

$$\liminf_{n\to\infty}\sup_{y\in\mathbb{R}^d}\int_{|x-y|\le R\rho_n}|u(t_n,x)|^2dx\ge\int_{|x|\le R}|V|^2dx.$$

As by hypothesis $\frac{(T^*-t_n)^{\frac{s}{2}}}{\alpha(t_n)}\to 0$ as $n\to\infty$, (6.1) implies that $\frac{\rho_n}{\alpha(t_n)}\to 0$ as $n\to\infty$. Therefore,

$$\liminf_{n\to\infty}\sup_{y\in\mathbb{R}^d}\int_{|x-y|\le\alpha(t_n)}|u(t_n,x)|^2dx\ge\int_{|x|\le R}|V|^2dx.$$

Letting $R\to\infty$, we obtain

$$\liminf_{n\to\infty}\sup_{y\in\mathbb{R}^d}\int_{|x-y|\le\alpha(t_n)}|u(t_n,x)|^2dx\ge\|V\|_2^2.$$

As $\|V\|_2\ge\|Q\|_2$, this implies

$$\limsup_{t\to T^*}\sup_{y\in\mathbb{R}^d}\int_{|x-y|\le\alpha(t)}|u(t,x)|^2dx\ge\|Q\|_2^2.$$

As for any fixed time t, the map $y\to\int_{|x-y|\le\alpha(t)}|u(t,x)|^2dx$ is continuous and goes to zero as $y\to\infty$, and there exists $x(t)\in\mathbb{R}^d$ such that

$$\sup_{y\in\mathbb{R}^d}\int_{|x-y|\le\alpha(t)}|u(t,x)|^2dx=\int_{|x-x(t)|\le\alpha(t)}|u(t,x)|^2dx.$$

This finally implies

$$\limsup_{t\to T^*}\int_{|x-x(t)|\le\alpha(t)}|u(t,x)|^2dx\ge\|Q\|_2^2,$$

which proves Theorem 1.2.

7. Proof of Theorem 1.3. In this section, we upgrade Theorem 1.1 to Theorem 1.3 under the additional assumption $\|u_0\|_2=\|Q\|_2$.

With the notation used in the proof of Theorem 1.1, we have

$$\|\psi_n\|_2\le\|u_0\|_2=\|Q\|_2\le\|V\|_2.$$

On the other hand, using the semicontinuity of weak convergence,

$$\|V\|_2\le\liminf_{n\to\infty}\|\psi_n\|_2\le\|Q\|_2.$$

Therefore,

$$\|V\|_2=\|Q\|_2=\lim_{n\to\infty}\|\psi_n\|_2.$$

Thus, as $\psi_n(\cdot+x_n)\rightharpoonup V$ weakly in L^2_x (up to a subsequence which we still denote by $\psi_n(\cdot+x_n)$), we conclude that

$$\psi_n(\cdot+x_n)\to V\ \text{ strongly in } L^2_x.$$

Moreover, by the Gagliardo–Nirenberg inequality and the boundedness of $\{\psi_n\}_{n\geq 1}$ in H^1_x, we have

$$\psi_n(\cdot + x_n) \to V \quad \text{in } L_x^{2+\frac{4}{d}}.$$

Combining this with (5.2) and the sharp Gagliardo–Nirenberg inequality, we obtain

$$\|\nabla Q\|_2 \leq \|\nabla V\|_2.$$

By the semicontinuity of weak convergence, we also have

$$\|\nabla V\|_2 \leq \liminf_{n\to\infty} \|\nabla \psi_n\|_2 = \|\nabla Q\|_2,$$

and so

$$\|\nabla V\|_2 = \|\nabla Q\|_2 = \lim_{n\to\infty} \|\nabla \psi_n\|_2.$$

Thus, as $\psi_n(\cdot + x_n) \rightharpoonup V$ in H^1_x, we conclude that

$$\psi_n(\cdot + x_n) \to V \quad \text{strongly in } H^1_x.$$

In particular, this implies

$$E(V) = 0.$$

Collecting the properties of V, we find

$$V \in H^1_x, \quad \|V\|_2 = \|Q\|_2, \quad \|\nabla V\|_2 = \|\nabla Q\|_2, \quad \text{and} \quad E(V) = 0.$$

The variational characterization of the ground state [39] implies that

$$V(x) = e^{i\theta} Q(x + x_0)$$

for some $(e^{i\theta}, x_0) \in (S^1 \times \mathbb{R}^d)$. Thus,

$$\rho_n^{\frac{d}{2}} (I_{N(t_n)} u)(t_n, \rho_n x + x_n) \to e^{i\theta} Q(x + x_0) \quad \text{strongly in } H^1_x \text{ as } n \to \infty. \tag{7.1}$$

Theorem 1.3 follows from (5.5) and (7.1).

Acknowledgments. The authors would like to thank the organizers of the Nonlinear Dispersive Equations program at M.S.R.I. during which this work was completed. The authors would also like to thank Jim Colliander for introducing us to the problem.

REFERENCES

[1] P. Begout and A. Vargas, *Mass Concentration Phenomena for the L^2-Critical Nonlinear Schrödinger Equation*, preprint.

[2] J. Bourgain, *Refinements of Strichartz inequality and applications to 2D-NLS with critical nonlinearity*, Internat. Math. Res. Notices, 5 (1998), pp. 253–283.

[3] J. Bourgain, *Scattering in the energy space and below for 3D NLS*, J. Anal. Math., 75 (1998), pp. 267–297.

[4] J. Bourgain, *New Global Well-Posedness Results for Non-linear Schrödinger Equations*, AMS, Providence, RI, 1999.

[5] T. Cazenave and F. B. Weissler, *Critical nonlinear Schrödinger Equation*, Nonlinear Anal., 14 (1990), pp. 807–836.

[6] T. Cazenave, *Semilinear Schrödinger Equations*, Courant Lect. Notes Math. 10, AMS, Providence, RI, 2003.

[7] M. Christ and M. Weinstein, *Dispersion of small amplitude solutions of the generalized Korteweg-de Vries equation*, J. Funct. Anal., 100 (1991), pp. 87–109.

[8] R. R. Coifman and Y. Meyer, *On commutators of singular integrals and bilinear singular integrals*, Trans. Amer. Math. Soc., 212 (1975), pp. 315–331.

[9] R. R. Coifman and Y. Meyer, *Ondelettes et opérateurs.* III, *Opérateurs multilinéaires*, Actualites Math., Hermann, Paris, 1991.

[10] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Almost conservation laws and global rough solutions to a nonlinear Schrödinger equation*, Math. Res. Lett., 9 (2002), pp. 659–682.

[11] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Polynomial upper bounds for the orbital instability of the* 1*D cubic NLS below the energy norm*, Discrete Contin. Dyn. Syst., 9 (2003), pp. 31–54.

[12] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Polynomial upper bounds for the instability of the nonlinear Schrödinger equation below the energy norm*, Commun. Pure Appl. Anal., 2 (2003), pp. 33–50.

[13] J. Colliander, S. Raynor, C. Sulem, and J. D. Wright, *Ground state mass concentration in the L^2-critical nonlinear Schrödinger equation below H^1*, Math. Res. Lett., 12 (2005), pp. 357–375.

[14] D. Fang and S. Zhong, *Cauchy problem for the L^2-critical nonlinear Schrödinger equation below H^1*, Nonlinear Anal., 62 (2005), pp. 117–130.

[15] P. Gerard, *Description du defaut de compacite de l'injection de Sobolev*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 213–233.

[16] J. Ginibre and G. Velo, *Scattering theory in the energy space for a class of nonlinear Schrödinger equations*, J. Math. Pure. Appl., 64 (1985), pp. 363–401.

[17] J. Ginibre and G. Velo, *Smoothing properties and retarded estimates for some dispersive evolution equations*, Comm. Math. Phys., 144 (1992), pp. 163–188.

[18] R. T. Glassey, *On the blowing up of solution to the Cauchy problem for nonlinear Schrödinger operators*, J. Math. Phys., 8 (1977), pp. 1794–1797.

[19] N. Hayashi and Y. Tsutsumi, *Remarks on the scattering problem for nonlinear Schrödinger equations*, in Differential Equations and Mathematical Physics (Birmingham, AL, 1986), Lecture Notes in Math. 1285, Springer, Berlin, 1987, pp. 162–168.

[20] T. Hmidi and S. Keraani, *Blowup theory for the critical nonlinear Schrödinger equations revisited*, Int. Math. Res. Not., 46 (2005), pp. 2815–2828.

[21] T. Hmidi and S. Keraani, *Remarks on the Blowup for the L^2-Critical Nonlinear Schrödinger Equations*, preprint.

[22] T. Kato, *On nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Phys. Théor., 46 (1987), pp. 113–129.

[23] T. Kato, *On nonlinear Schrödinger equations*, II. *H^s-solutions and unconditional well-posedness*, J. Anal. Math., 67 (1995), pp. 281–306.

[24] M. Keel and T. Tao, *Endpoint Strichartz estimates*, Amer. Math. J., 120 (1998), pp. 955–980.

[25] S. Keraani, *On the blow up phenomenon of the critical nonlinear Schrödinger equation*, J. Funct. Anal., 235 (2006), pp. 171–192.

[26] M. K. Kwong, *Uniqueness of positive solutions of $\Delta u - u + u^p = 0$ in $\mathbb{R}^n$*, Arch. Rational Mech. Anal., 105 (1989), pp. 243–266.

[27] K. Mcleod and J. Serrin, *Nonlinear Schrödinger equation. Uniqueness of positive solutions of $\Delta u + f(u) = 0$ in $\mathbb{R}^n$*, Arch. Rational Mech. Anal., 99 (1987), pp. 115–145.

[28] F. Merle, *Determination of blow-up solutions with minimal mass for nonlinear Schrödinger equations with critical power*, Duke Math. J., 69 (1993), pp. 427–454.

[29] F. Merle and P. Raphael, *The blow-up dynamic and upper bound on the blow-up rate for critical nonlinear Schrödinger equation*, Ann. of Math. (2), 161 (2005), pp. 157–222.

[30] F. Merle and P. Raphael, *Profiles and quantization of the blow up mass for critical nonlinear Schrödinger equation*, Comm. Math. Phys., 253 (2005), pp. 675–704.

[31] F. Merle and Y. Tsutsumi, *L^2-concentration of blow-up solutions for the nonlinear Schrödinger equation with the critical power nonlinearity*, J. Differential Equations, 84 (1990), pp. 205–214.

[32] G. Staffilani, *On the generalized Korteweg–de Vries-type equations*, Differential Integral Equations, 10 (1997), pp. 777–796.

[33] R. S. Strichartz, *Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations*, Duke Math. J., 44 (1977), pp. 705–714.

[34] T. Tao, *On the asymptotic behavior of large radial data for a focusing non-linear Schrödinger equation*, Dyn. Partial Differ. Equ., 1 (2004), pp. 1–48.

[35] M. E. Taylor, *Tools for PDE. Pseudodifferential Operators, Paradifferential Operators, and Layer Potentials*, Math. Surveys Monogr. 81, AMS, Providence, RI, 2000.

[36] N. Tzirakis, *Mass concentration phenomenon for the quintic nonlinear Schrödinger equation in one dimension*, SIAM J. Math. Anal., 37 (2006), pp. 1923–1946.

[37] M. Visan, *The Defocusing Energy-Critical Nonlinear Schrödinger Equation in Dimensions Five and Higher*, Ph.D. Thesis, University of California, Los Angeles, CA.

[38] M. Visan, *The defocusing energy-critical nonlinear Schrödinger equation in higher dimensions*, Duke Math. J., to appear.

[39] M. I. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates*, Comm. Math. Phys., 87 (1982/1983), pp. 567–576.

[40] M. I. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal., 16 (1985), pp. 472–491.

[41] M. I. Weinstein, *Lyapunov stability of ground states of nonlinear Schrödinger equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–68.

[42] V. E. Zakharov and E. A. Kuznetsov, *Quasi-classical theory for three-dimensional wave collapse*, Soviet Phys. JETP, 64 (1986), pp. 773–380.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 57–76

INVERSE SOURCE PROBLEMS IN TRANSPORT EQUATIONS*

GUILLAUME BAL† AND ALEXANDRU TAMASAN‡

Abstract. This paper proposes an iterative technique to reconstruct the source term in transport equations, which account for scattering effects, from boundary measurements. In the two-dimensional setting, the full outgoing distribution in the phase space (position and direction) needs to be measured. In three space dimensions, we show that measurements for angles that are orthogonal to a given direction are sufficient. In both cases, the derivation is based on a perturbation of the inversion of the two-dimensional attenuated Radon transform and requires that (the anisotropic part of) scattering be sufficiently small. We present an explicit iterative procedure, which converges to the source term we want to reconstruct. Applications of the inversion procedure include optical molecular imaging, an increasingly popular medical imaging modality.

Key words. attenuated Radon transform, scattering, optical molecular imaging

AMS subject classifications. 35R30, 35R15, 92C65

DOI. 10.1137/050647177

1. Introduction.

Optical molecular imaging (OMI) is being increasingly studied as a powerful detection method in medical imaging. New biochemical markers are currently being engineered to attach to specific molecules and thus be used to detect faulty genes and other molecular processes, which precede the development of certain diseases. This makes possible the detection of such diseases long before phenotypical symptoms appear. In OMI, the markers are light-emitting molecules, such as fluorophores or luminophores. Compared to other molecular imaging techniques, such as single photon emission tomography (SPECT) or positron emission tomography (PET), optical markers emit low-energy near-infrared photons that are relatively harmless to the human body. Other advantages are their high sensitivity to oxygen levels, metal ion concentrations, pH, and lipid composition, for instance; see [31, 32, 40, 41] for recent references in the biomedical literature.

The inverse problem consists of reconstructing the spatial distribution of the markers from measurements of light intensities at the boundary of the object we wish to image. Two main types of markers are used in OMI, namely, bioluminescent and fluorescent markers. In both cases, the propagation in human tissues of the photons emitted by the markers can quite satisfactorily be modeled as inverse source problems of time-harmonic and steady-state radiative transfer equations [11, 21, 42]. To simplify the presentation, we consider only the steady-state problem here, for which relatively few results exist in the mathematical literature.

Our main result consists in providing an explicit (and converging) iterative technique to reconstruct the source term from boundary measurements of the photon intensity in the phase space, i.e., as a function of position and angular direction. We consider the two-dimensional and the three-dimensional settings. In both cases,

*Received by the editors December 9, 2005; accepted for publication (in revised form) October 31, 2006; published electronically April 10, 2007.
http://www.siam.org/journals/sima/39-1/64717.html

†Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027 (gb2030@columbia.edu). This author was supported in part by NSF grant DMS-0239097 and by an Alfred P. Sloan fellowship.

‡Department of Mathematics, University of Central Florida, Orlando, FL 32816 (tamasan@math.ucf.edu).

we have to assume that the *anisotropic part* of scattering is sufficiently regular and small (in the sense that a certain operator linear in the anisotropic part of the scattering term must have a norm bounded by one in appropriate spaces). In three dimensions, we show that measurements of the photon intensity for directions orthogonal to an arbitrary given vector are sufficient. Both results are based on perturbations of the Novikov inversion formula to invert the attenuated Radon transform (see [3, 7, 9, 19, 14, 15, 20, 26, 29, 30] for some references on that problem), and we thus show that the Novikov inversion formula is stable under perturbations by a scattering operator. How small scattering has to be in terms of the absorbing and geometric properties of the domain is somewhat characterized in Corollary 3.7.

Several imaging techniques such as SPECT and PET are based on the inversion of the Radon transform or the attenuated Radon transform. Because optical markers emit low-energy light, the photons scatter before they are measured. This renders the inversion more difficult than in the higher-energy methods SPECT and PET and necessitates the use of transport equations that account for scattering effects.

For earlier works on the inverse source problem of transport equations based on different methods, we refer the reader to [2, 22, 33, 35, 37, 1]. A more general geometric setting than is presented here can be found in [35] (see also [36]), where particles may travel along the geodesics of a general Riemannian manifold. For a fixed domain Ω the main assumption in the uniqueness and stability result in [35] imposes a smallness condition on a, even in the absence of scattering. This constraint is not needed here.

In this paper, we are interested only in the source reconstruction and assume that the absorption and scattering coefficients are known; see [4, 5, 6, 12, 18, 23, 24, 38, 39] for references on the determination of these parameters.

The iterative procedure presented here will not work in the highly scattering regime (unless that scattering is fully isotropic), in which case the diffusion approximation should be used [4]. It should be mostly effective in situations where scattering needs to be accounted for to obtain a desired accuracy in the reconstruction and yet is not too strong for a method based on a perturbation of a nonscattering inversion technique to converge. Practically, we expect this situation to arise in OMI in small domains (on the order of 5–10 mean free paths), such as small animals, and in SPECT and PET, where moderate scattering is accounted for.

The rest of the paper is organized as follows. Section 2 introduces the inverse source problem in transport equations and presents our main results. The derivation of the results is postponed to section 3 for the two-dimensional case and section 4 for the three-dimensional extension.

2. An inverse source problem. The distribution of photons emitted by the markers is denoted by $f(\mathbf{x})$, where position $\mathbf{x} \in \Omega \subset \mathbb{R}^d$. Here Ω is a bounded open convex domain and $d = 2, 3$ is the space dimension. We normalize the light speed to unity and denote by $\boldsymbol{\theta} \in S^{d-1}$ the direction of the photons. Notice that $d = 3$ is the physical model, whereas $d = 2$ is not physical as photons are allowed to travel only in a two-dimensional plane.

Let $u(\mathbf{x}, \boldsymbol{\theta})$ be the density of photons at position x moving in the direction $\boldsymbol{\theta}$, and let

$$\Gamma_\pm \equiv \Gamma_\pm(\Omega) = \{(\mathbf{x}, \boldsymbol{\theta}) \in \partial\Omega \times S^{d-1}, \quad \pm\boldsymbol{\theta} \cdot \mathbf{n}(\mathbf{x}) > 0\} \tag{1}$$

denote the boundary spaces. Here $\mathbf{n}(\mathbf{x})$ is the outward normal to Ω at $\mathbf{x} \in \partial\Omega$. The density of particles satisfies the radiation transfer (transport) equation

$$\begin{aligned} &\boldsymbol{\theta} \cdot \nabla_{\mathbf{x}} u(\mathbf{x}, \boldsymbol{\theta}) + a(\mathbf{x}) u(\mathbf{x}, \boldsymbol{\theta}) = K u(\mathbf{x}, \boldsymbol{\theta}) + f(\mathbf{x}) \quad \text{in } \Omega \times S^{d-1}, \\ &u(\mathbf{x}, \boldsymbol{\theta}) = 0 \quad \text{on } \Gamma_-, \end{aligned} \tag{2}$$

where the measure $d\boldsymbol{\theta}$ is the usual surface measure on the unit sphere normalized such that $\int_{S^{d-1}} d\boldsymbol{\theta} = 1$. Photon interaction with the underlying medium is modeled by an absorption parameter $a(\mathbf{x})$ and a scattering operator

$$Ku(\mathbf{x}, \boldsymbol{\theta}) = \int_{S^{d-1}} k(\mathbf{x}, \boldsymbol{\theta} \cdot \boldsymbol{\theta}') u(\mathbf{x}, \boldsymbol{\theta}') d\boldsymbol{\theta}', \tag{3}$$

where $k(\mathbf{x}, \mu)$ is the scattering coefficient.

Throughout this paper, the absorption a and the scattering k are nonnegative smooth functions which satisfy the subcritical inequality

$$a(\mathbf{x}) - \int_{S^{d-1}} k(\mathbf{x}, \boldsymbol{\theta} \cdot \boldsymbol{\theta}') d\boldsymbol{\theta}' \geq 0, \tag{4}$$

pointwise in $\mathbf{x} \in \Omega$. We assume that the absorption coefficient $a \in C^2(\Omega)$ and it is of compact support. The smoothness of the scattering coefficient k will be stated in the theorems. The subcritical assumption (4) defines the forward operator, which maps the source f to the distribution u and is bounded from $L^2(\Omega)$ to $L^2(\Omega \times S^{d-1})$. Since $\overline{\Omega}$ is compact, it is shown in [25] that the operator norm of the forward operator is bounded independently of the scattering kernel k in (4).

Moreover, the outgoing photon distribution, defined as the trace of $u(\mathbf{x}, \boldsymbol{\theta})$ on Γ_+, is well defined and belongs to $L^2_{\boldsymbol{\theta} \cdot \mathbf{n}}(\Gamma_+)$, in the sense that $\int_{\Gamma_+} \boldsymbol{\theta} \cdot \mathbf{n}\, u^2(\mathbf{x}, \boldsymbol{\theta}) d\sigma(\mathbf{x}) d\boldsymbol{\theta} < \infty$, where $d\sigma$ is the surface measure on $\partial\Omega$. For additional references on the mathematical theory of the transport equation (2), see, for instance, [10, 13, 25].

We remark that for the inverse problem under consideration, the fact that a has compact support is not an essential restriction. Indeed, we can extend a to a larger domain $\tilde{\Omega} \supset \Omega$ by preserving its C^2 smoothness and such that it vanishes near the boundary. The scattering is extended by zero outside Ω. Now we can consider the boundary value problem in the larger domain $\tilde{\Omega}$. The operator norm of the forward operator may increase, but its bound is still independent of the scattering coefficient on Ω. Since scattering vanishes on $\tilde{\Omega} \backslash \Omega$, the values of the transport solution at the boundary of the extended domain $\tilde{\Omega}$ on $\Gamma_{\pm}(\tilde{\Omega})$ can easily be related to the values of the transport solution at the boundary of the initial domain Ω on $\Gamma_{\pm}(\Omega)$ by a bijective transformation.

It is convenient in the analysis to have unbounded spatial domains. We extended $f(\mathbf{x})$, $k(\mathbf{x}, \mu)$, and $a(\mathbf{x})$ by 0 on $\mathbb{R}^d \backslash \Omega$. The transport equation is now recast as

$$\begin{aligned} &\boldsymbol{\theta} \cdot \nabla_{\mathbf{x}} u(\mathbf{x}, \boldsymbol{\theta}) + a(\mathbf{x}) u(\mathbf{x}, \boldsymbol{\theta}) = K u(\mathbf{x}, \boldsymbol{\theta}) + f(\mathbf{x}) \quad \text{in } \mathbb{R}^d \times S^{d-1}, \\ &\lim_{t \to \infty} u(\mathbf{x} - t\boldsymbol{\theta}, \boldsymbol{\theta}) = 0 \quad \text{on } \mathbb{R}^d \times S^{d-1}. \end{aligned} \tag{5}$$

The restriction of the above solution on $\Omega \times S^{d-1}$ solves (2).

Our main results are that in dimension $d = 2$, knowledge of

$$m(s, \boldsymbol{\theta}) = \lim_{t \to \infty} u(t\boldsymbol{\theta} + s\boldsymbol{\theta}^{\perp}, \boldsymbol{\theta}) \tag{6}$$

on $\mathbb{R} \times S^1$, with u the unique solution of (5), uniquely determines $f(\mathbf{x})$ compactly supported on the bounded domain Ω provided that the scattering kernel $k(\mathbf{x}, \mu)$ is sufficiently small. Moreover, the reconstruction is explicit, in the sense that $f(\mathbf{x})$ is obtained as the limit of a converging Neumann series expansion. Note that $u(t\boldsymbol{\theta} + s\boldsymbol{\theta}^{\perp}, \boldsymbol{\theta})$ is independent of t for t sufficiently large.

In three dimensions, $d = 3$, the above result generalizes as follows. Let an arbitrary vector in $\mathbb{R}^3$ be given, which after possible rotation of Ω we denote by $\mathbf{e}_z$. For $\boldsymbol{\theta} = (\cos\theta, \sin\theta, 0)$, we define $\boldsymbol{\theta}^{\perp} = (-\sin\theta, \cos\theta, 0)$. Then knowledge of

$$m(z, s, \theta) = \lim_{t\to\infty} u(t\boldsymbol{\theta} + s\boldsymbol{\theta}^{\perp} + z\mathbf{e}_z, \boldsymbol{\theta}) \tag{7}$$

for $(z, s, \theta) \in \mathbb{R} \times \mathbb{R} \times (0, 2\pi)$ uniquely determines $f(\mathbf{x})$ compactly supported on the bounded domain Ω. This result also requires that $k(\mathbf{x}, \mu)$ be sufficiently small in an appropriate sense and the reconstruction is explicit in the sense mentioned above. This implies that the outgoing measurements are known only for directions orthogonal to $\mathbf{e}_z$. Note that in both cases, the problem is formally determined since both the measurements as well as the unknown source term are d-dimensional.

To state the regularity and smallness assumption of the scattering, we introduce the following notation. When $d = 2$, we identify $k(\mathbf{x}, \boldsymbol{\theta}\cdot\boldsymbol{\theta}') = k(\mathbf{x}, \cos(\theta - \theta')) = \tilde{k}(\mathbf{x}, \theta - \theta')$ and define the Fourier coefficients $k_n(\mathbf{x})$ by

$$k_n(\mathbf{x}) = \frac{1}{2\pi}\int_0^{2\pi} \tilde{k}(\mathbf{x}, \theta) e^{-in\theta} d\theta. \tag{8}$$

By $\hat{k}_n(\boldsymbol{\xi})$, we mean its Fourier transform $\hat{k}_n(\boldsymbol{\xi}) = \int_{\mathbb{R}^2} e^{-i\boldsymbol{\xi}\cdot\mathbf{x}} k_n(\mathbf{x}) d\mathbf{x}$.

When $d = 3$, we use the Legendre polynomials expansion in $L^2[-1, 1]$:

$$k(\mathbf{x}, t) = \sum_{n=0}^{\infty} k_n(\mathbf{x}) P_n(t). \tag{9}$$

By $\hat{k}_n(\boldsymbol{\xi}', z)$, we mean the restricted Fourier transform to the horizontal plane $\hat{k}_n(\boldsymbol{\xi}, z) = \int_{\mathbb{R}^2} e^{-i\mathbf{x}'\cdot\boldsymbol{\xi}'} k_n(\mathbf{x}', z) d\mathbf{x}'$. For $\boldsymbol{\theta} = (\cos\theta\sin\phi, \sin\theta\sin\phi, \cos\phi) \in S^2$, $0 \le \theta < 2\pi$, $0 \le \phi < \pi$, let

$$Y_{nm}(\boldsymbol{\theta}) = C_{nm}^{1/2} e^{im\theta} P_n^m(\cos\phi) \tag{10}$$

denote the spherical harmonics on the sphere S^2. Here, P_n^m are the associated Legendre polynomials and $C_{nm} = (2n+1)(n-m)!/(n+m)!$; see [17] for details. We need only to consider horizontal directions

$$\boldsymbol{\theta} \in S_H^2 = \{\boldsymbol{\theta} \in S^2 : \boldsymbol{\theta}\cdot e_z = 0\}. \tag{11}$$

We are ready to formulate our main results, whose proof is postponed to the following sections. The decay uses the usual notation $\langle n\rangle = (1 + |n|^2)^{1/2}$.

THEOREM 2.1 (two-dimensional case). *Let $f(\mathbf{x}) \in L^2(\mathbb{R}^2)$ be a source term of compact support in Ω and $a \in C_0^2(\Omega)$ an absorption coefficient of compact support. Then there exists $\epsilon > 0$ depending on the size of the support and on the smoothness of a such that, for a scattering coefficient k with*

$$\max_{n\in\mathbb{Z}} \langle n\rangle^{\alpha} \|\hat{k}_n\|^2_{L^1(\mathbb{R}^2)} < \epsilon \tag{12}$$

for some $\alpha > 1$, the measurements $m(s, \theta)$ in (6) uniquely determine the source term $f(\mathbf{x})$. Moreover, the source term $f(\mathbf{x})$ can be obtained as the limit of the explicit *convergent Neumann series in* (44) *below; see also Remark* 3.12. *A more explicit expression for* ϵ *in* (12) *can be found in Corollary* 3.7 *below.*

THEOREM 2.2 (three-dimensional case). *Let* $f(\mathbf{x}) \in L^2(\mathbb{R}^3)$ *be a source term of compact support and* $a \in C_0^2(\Omega)$ *an absorption coefficient of compact support. Then there exists an* $\epsilon > 0$ *depending on the size of the support and on the smoothness of* a *such that, for a scattering kernel* k *with*

$$\max_{n\in\mathbb{N}} \left(\langle n \rangle^{\alpha-1} \max_{|m|\le n} \max_{\boldsymbol{\theta}\in S_H^2} |Y_{nm}(\boldsymbol{\theta})|^2 \int_{\mathbb{R}} \|\hat{k}_n(\cdot,z)\|^2_{L^1(\mathbb{R}^2)} dz \right) \le \epsilon \tag{13}$$

for some $\alpha > 1$, *the measurements* $m(z,s,\theta)$ *in* (7) *uniquely determine the source term* $f(\mathbf{x})$. *Moreover, the source term* $f(\mathbf{x})$ *can be obtained as the limit of the convergent Neumann series expansion in* (66) *below.*

Note that each theorem requires smallness as well as (weak) smoothness on the scattering kernel k. That k is not arbitrary is already apparent in the existence theory for the forward problem, where we have assumed (4). The L^1-norm in the Fourier variables implies continuity of the scattering in the horizontal plane. The Sobolev-type decay property implies smoothness in the angular variable. For instance, $\alpha > 2$ in (12) already implies continuity of k in θ. The size of ϵ depends on the (operator) norm of the operator N_K introduced below and is not explicit. This restriction is the price to pay to obtain a reconstruction as a perturbation of the inversion of the attenuated Radon transform, where there is no scattering.

Let us conclude this section by noting, as was mentioned in the introduction, that only the *anisotropic* part of the scattering need be small.

COROLLARY 2.3. *The results stated in Theorems* 2.1 *and* 2.2 *are still valid when* (12) *and* (13), *respectively, hold only for* $n \neq 0$.

Proof. Indeed, let us decompose $K = K_0 + K_1$, where $K_0u(\mathbf{x}) = k_0(\mathbf{x}) \int_{S^{d-1}} u(\mathbf{x}, \boldsymbol{\theta}')$ $\cdot d\boldsymbol{\theta}'$ is the isotropic part of K, and $K_1 = K - K_0$ the anisotropic part. We define the source term

$$F(\mathbf{x}) = f(\mathbf{x}) + K_0u(\mathbf{x}), \tag{14}$$

where $u(\mathbf{x}, \boldsymbol{\theta})$ is the solution to (5). We then verify that the following equation also holds:

$$\begin{aligned} &\boldsymbol{\theta} \cdot \nabla_{\mathbf{x}} u(\mathbf{x}, \boldsymbol{\theta}) + a(\mathbf{x}) u(\mathbf{x}, \boldsymbol{\theta}) = K_1 u(\mathbf{x}, \boldsymbol{\theta}) + F(\mathbf{x}) \quad \text{in } \mathbb{R}^d \times S^{d-1}, \\ &\lim_{t\to\infty} u(\mathbf{x} - t\boldsymbol{\theta}, \boldsymbol{\theta}) = 0 \quad \text{on } \mathbb{R}^d \times S^{d-1}. \end{aligned} \tag{15}$$

We can then apply Theorem 2.1 or 2.2, depending on dimension d, based on the smallness assumptions on K_1, and conclude that $F(\mathbf{x})$ can be reconstructed from the boundary measurements. Once $F(\mathbf{x})$ is known, we can solve for u in (15) and thus calculate $K_0u(\mathbf{x})$. It remains to identify $f(\mathbf{x}) = F(\mathbf{x}) - K_0u(\mathbf{x})$ to conclude the proof of the corollary. ∎

3. Derivation in two space dimensions. This section is devoted to the derivation of the inversion procedure in two space dimensions and on the proof of Theorem 2.1. For any unit vector $\boldsymbol{\theta} \in S^1$, we introduce the representation $\boldsymbol{\theta} = (\cos\theta, \sin\theta)$ for $0 \le \theta < 2\pi$ and identify any function $f(\boldsymbol{\theta}) \equiv f(\theta)$.

We define the classical beam transform S and the symmetrized beam transform

D (independently of the spatial dimension d) as

$$Sa(\mathbf{x}, \boldsymbol{\theta}) = \int_{-\infty}^{0} a(\mathbf{x} + s\boldsymbol{\theta}) ds, \tag{16}$$

$$Da(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{2}\Big(\int_{-\infty}^{0} a(\mathbf{x} + s\boldsymbol{\theta}) ds - \int_{0}^{\infty} a(\mathbf{x} + s\boldsymbol{\theta}) ds \Big). \tag{17}$$

Since $a(\mathbf{x}) \in C_0^2(\Omega)$ then $Da(\mathbf{x}, \theta)$ and $(e^{Da})(\mathbf{x}, \theta)$ are well defined C^2-smooth functions. Notice that e^{Da} is an integrating factor for (5) (since $(\boldsymbol{\theta} \cdot \nabla) e^{Da} = a e^{Da}$) and so it is more convenient to consider the equation for

$$w(\mathbf{x}, \theta) = (e^{Da} u)(\mathbf{x}, \theta). \tag{18}$$

Next we multiply (5) by the integrating factor e^{Da} and integrate in the direction of $\boldsymbol{\theta}$ to see that w solves the equivalent integral equation [13, 25]

$$w(\mathbf{x}, \theta) = S e^{Da} K e^{-Da} w(\mathbf{x}, \theta) + S e^{Da} f(\mathbf{x}, \theta). \tag{19}$$

Under the subcritical assumption (4) the equation above is uniquely solvable [25] and the operator

$$T := [I - S e^{Da} K e^{-Da}]^{-1} S e^{Da} \tag{20}$$

is bounded from $L^2(\Omega)$ to $L^2(\mathbb{R}^2 \times S^1)$. Moreover, the operator norm can be bounded independently of the scattering operator K provided that (4) holds; see [25] for the details.

We have then

$$w(\mathbf{x}, \theta) = Tf(\mathbf{x}, \theta) = S e^{Da} f + S e^{Da} K e^{-Da} T f. \tag{21}$$

Let us introduce the operator L acting on functions $w(\mathbf{x}, \theta)$ as

$$Lw(s, \theta) = \lim_{t \to \infty} w(t\boldsymbol{\theta} + s\boldsymbol{\theta}^{\perp}, \theta). \tag{22}$$

The product LS is the usual Radon transform

$$Rf(s, \theta) = LSf(s, \theta). \tag{23}$$

It is convenient to work with slightly modified measurements. Let us introduce

$$g(s, \theta) = Lw(s, \theta) = \lim_{t \to \infty} (e^{Da} u)(t\boldsymbol{\theta} + s\boldsymbol{\theta}^{\perp}, \theta) = e^{\frac{1}{2} Ra}(s, \theta) m(s, \theta), \tag{24}$$

where $m(s, \theta)$ was defined in (6). Since a is known, then so are the new "measurements" $g(s, \theta)$.

Let us finally introduce the attenuated X-ray transform operator

$$R_a f(s, \theta) = LS e^{Da} f(s, \theta). \tag{25}$$

Applying L to (19), we deduce that the measurements $g(s, \theta)$ are given by

$$g(s, \theta) = R_a f(s, \theta) + R e^{Da} K e^{-Da} T f(s, \theta). \tag{26}$$

An inversion for R_a was recently obtained in [29]; see also [3, 7, 9, 19, 26, 30] for recent works on the attenuated X-ray transform. We define the inversion operator N, acting on functions $g(s,\theta)$ defined on $\mathbb{R}\times S^1$, by

$$Ng(\mathbf{x}) = \frac{1}{4\pi}\int_0^{2\pi} \boldsymbol{\theta}^{\perp}\cdot\nabla_{\mathbf{x}}(R^*_{-a,\theta}H_a g)(\mathbf{x},\theta)d\theta, \tag{27}$$

where

$$\begin{aligned} R^*_{a,\theta}g(\mathbf{x}) &= e^{D_\theta a(\mathbf{x})}g(\mathbf{x}\cdot\boldsymbol{\theta}^{\perp}), \\ H_a &= C_cHC_c + C_sHC_s, & Hu(t) &= \frac{1}{\pi}\int_{\mathbb{R}}\frac{u(s)}{t-s}ds, \\ C_cg(s,\theta) &= g(s,\theta)\cos\left(\frac{HRa(s,\theta)}{2}\right), & C_sg(s,\theta) &= g(s,\theta)\sin\left(\frac{HRa(s,\theta)}{2}\right). \end{aligned} \tag{28}$$

The integral in the Hilbert transform H, which acts in C_c and in C_s on the s variable, has to be understood in the principal value sense. Note that $H_a = H$ in the absence of absorption ($a \equiv 0$) and that the above formula then becomes the usual inversion of the Radon transform [27].

We thus formally apply the operator N to (26) and obtain the equation for $f(\mathbf{x})$:

$$Ng(\mathbf{x}) = f(\mathbf{x}) + NRe^{Da}Ke^{-Da}Tf(\mathbf{x}). \tag{29}$$

Let $\chi(\mathbf{x})$ be a cut-off function supported on Ω and such that $\chi\equiv 1$ on the support of f. The equation above is recast as

$$\chi(\mathbf{x})Ng(\mathbf{x}) = f(\mathbf{x}) + \chi NRe^{Da}Ke^{-Da}Tf(\mathbf{x}) =: (I-N_K)f(\mathbf{x}), \tag{30}$$

where we have introduced the operator $N_K = -\chi NRe^{Da}Ke^{-Da}T$. This equation is of Fredholm type since K has a smoothing property which makes $Ke^{-Da}T : L^2(\Omega)\hookrightarrow L^2(\Omega\times S^{d-1})$ compact; see [25]. We do not use the compactness property in Theorems 2.1 and 2.2. Our reconstruction results are based on a smallness assumption on N_K. Note, however, that, more generally, our results show that the source term $f(\mathbf{x})$ can be reconstructed from the measured data provided that 1 is not an eigenvalue of the compact operator N_K.

The proof of Theorem 2.1 is based on the following result.

THEOREM 3.1. *The operator N_K defined above is bounded from $L^2(\Omega)$ to $L^2(\Omega)$.*

We study first the mapping properties of the scattering operator K. For this we introduce the functional spaces

$$L^{\hat{2}}(\mathbb{R}^2;C^0(S^1)) = \Big\{u(\mathbf{x},\boldsymbol{\theta}) \text{ s.t. } \hat{u}(\boldsymbol{\xi},\boldsymbol{\theta})\in L^2(\mathbb{R}^2;C^0(S^1))\Big\}, \tag{31}$$

$$L^{\hat{2}}(\Omega;C^0(S^1)) = \Big\{u\in L^{\hat{2}}(\mathbb{R}^2;C^0(S^1)) \text{ s.t. supp } u(\cdot,\theta)\subseteq\Omega\Big\}, \tag{32}$$

endowed with the norm

$$\|u\|^2_{\hat{2},\infty} = \|\hat{u}\|^2_{L^2(\mathbb{R}^2;C^0(S^1))} = \int_{\mathbb{R}^2}\max_{\boldsymbol{\theta}\in S^1}|\hat{u}(\boldsymbol{\xi},\boldsymbol{\theta})|^2 d\boldsymbol{\xi}.$$

Since K is a convolution in the angular variable, it is decomposed as

$$Ku(\mathbf{x},\boldsymbol{\theta}) = \sum_{n=-\infty}^{\infty} k_n(\mathbf{x})u_n(\mathbf{x})e^{in\theta}. \tag{33}$$

LEMMA 3.2. *Assume that* supp $k(\cdot,\theta)\subset\Omega$ *and that* $\max_{n\in\mathbb{Z}}\langle n\rangle^\alpha\|k_n\|^2_{L^1(\mathbb{R}^2)}<C$ *for some* $\alpha>1$. *Then the operator* K *maps* $L^2(\mathbb{R}^2\times S^1)$ *to* $L^{\hat{2}}(\Omega;C^0(S^1))$.

Proof. Taking the Fourier transform in the space variable in (33), we get

$$|\widehat{Ku}(\xi,\boldsymbol{\theta})|^2=\left|\sum_{-\infty}^{\infty}(\widehat{k_n}*\widehat{u_n})(\xi)e^{in\theta}\right|^2\le\left(\sum_{-\infty}^{\infty}\left|\widehat{k}_n*\widehat{u}_n\right|(\boldsymbol{\xi})\right)^2$$
$$\le\left(\sum_{-\infty}^{\infty}\frac{1}{\langle n\rangle^\alpha}\right)\left(\sum_{-\infty}^{\infty}\langle n\rangle^\alpha|\hat{k}_n*\hat{u}_n|^2(\boldsymbol{\xi})\right).$$

Now we take the maximum over $\boldsymbol{\theta}\in S^1$ on both sides and integrate in $\xi\in\mathbb{R}^2$ to get

$$\|Ku\|^2_{L^{\hat{2}}(\mathbb{R}^2;C^0(S^1))}\le\left(\sum_{-\infty}^{\infty}\frac{1}{\langle n\rangle^\alpha}\right)\sum_{-\infty}^{\infty}\langle n\rangle^\alpha\|\hat{k}_n*\hat{u}_n\|^2_{L^2(\mathbb{R}^2)}$$
$$\le\left(\sum_{-\infty}^{\infty}\frac{1}{\langle n\rangle^\alpha}\right)\sum_{-\infty}^{\infty}\langle n\rangle^\alpha\|\hat{k}_n\|^2_{L^1(\mathbb{R}^2)}\|\hat{u}_n\|^2_{L^2(\mathbb{R}^2)}$$
$$\le\left(\sum_{-\infty}^{\infty}\frac{1}{\langle n\rangle^\alpha}\right)\max_{n\in\mathbb{N}}\langle n\rangle^\alpha\|\hat{k}_n\|^2_{L^1(\mathbb{R}^2)}\sum_{-\infty}^{\infty}\|\hat{u}_n\|^2_{L^2(\mathbb{R}^2)}$$
$$\le C\left(\sum_{-\infty}^{\infty}\frac{1}{\langle n\rangle^\alpha}\right)\|u\|^2_{L^2(\mathbb{R}^2\times S^1)}.\qquad\square$$

LEMMA 3.3. *Let* $h:\mathbb{R}^2\times S^1\to\mathbb{R}$ *be a smooth map such that*

$$\int_{\mathbb{R}^2}\max_{\boldsymbol{\theta}}|\hat{h}(\xi,\boldsymbol{\theta})|d\xi<\infty.$$

Then the operator M_h *of multiplication by* h *is bounded from* $L^{\hat{2}}(\Omega;C^0(S^1))$ *to itself.*

Proof.

$$\|M_hf\|^2_{\hat{2},\infty}=\|hf\|^2_{\hat{2},\infty}=\int_{\mathbb{R}^2}\max_{\boldsymbol{\theta}\in S^1}|\widehat{hf}|^2(\xi,\boldsymbol{\theta})d\xi=\int_{\mathbb{R}^2}\max_{\theta\in S^1}|\hat{h}*_\xi\hat{f}|^2(\xi,\theta)d\xi$$
$$\le\left\|\max_{\boldsymbol{\theta}\in S^1}|\hat{h}|*\max_{\boldsymbol{\theta}\in S^1}|\hat{f}|\right\|^2_{L^2}\le\left\|\max_{\boldsymbol{\theta}\in S^1}|\hat{h}|\right\|^2_{L^1}\|f\|^2_{\hat{2},\infty}.\qquad\square\tag{34}$$

Recall that, when acting on maps $f(x,\boldsymbol{\theta})$, R denotes the Radon transform in $x\in\mathbb{R}^2$. The following smoothing property holds.

LEMMA 3.4. *The operator* $R:L^{\hat{2},\infty}(\Omega\times S^1)\to H^{1/2}(\mathbb{R};L^2(S^1))$ *is bounded; more precisely,*

$$\int_{S^1}\int_{\mathbb{R}}\left|\widehat{Rg}(\rho,\theta)\right|^2(1+|\rho|)d\rho d\theta\le(4\pi|\Omega|^2+3)\|g\|^2_{\hat{2},\infty},$$

where $|\Omega|$ *denotes the volume of* Ω.

Proof. Notice first that the Fourier slice theorem $\widehat{Rg}(\rho,\theta)=\hat{g}(\rho\boldsymbol{\theta}^\perp,\theta)$ holds. On the right-hand side the Fourier transform is taken with respect to the space variable only. The following inequalities hold:

$$\begin{aligned}\int_0^{2\pi}\int_0^\infty \left|\widehat{Rg}(\rho,\theta)\right|^2 \rho d\rho d\theta &= \int_0^{2\pi}\int_0^\infty \left|\hat{g}(\rho\boldsymbol{\theta}^\perp,\theta)\right|^2 \rho d\rho d\theta\\ &\le \int_0^{2\pi}\int_0^\infty \max_{\nu\in S^1}\left|\hat{g}(\rho\boldsymbol{\theta}^\perp,\nu)\right|^2 \rho d\rho d\theta = \|g\|^2_{\hat{2},\infty},\\ \int_0^{2\pi}\int_{-\infty}^0 \left|\widehat{Rg}(\rho,\theta)\right|^2 |\rho| d\rho d\theta &= \int_0^{2\pi}\int_0^\infty \left|\hat{g}(-\rho\boldsymbol{\theta}^\perp,\theta)\right|^2 \rho d\rho d\theta\\ &\le \int_0^{2\pi}\int_0^\infty \max_{\nu\in S^1}\left|\hat{g}(-\rho\boldsymbol{\theta}^\perp,\nu)\right|^2 \rho d\rho d\theta = \|g\|^2_{\hat{2},\infty}.\end{aligned}$$

We also have

$$\int_0^{2\pi}\int_0^\infty |\widehat{Rg}(\rho,\theta)|^2 d\rho d\theta \le \int_0^{2\pi}\int_0^1 |\hat{g}(\rho\boldsymbol{\theta}^\perp,\theta)|^2 d\rho d\theta + \int_0^{2\pi}\int_0^\infty |\widehat{Rg}(\rho,\theta)|^2 \rho d\rho d\theta$$

and

$$\int_0^{2\pi}\int_0^1 |\hat{g}(\rho\boldsymbol{\theta}^\perp,\theta)|^2 d\rho d\theta \le \int_0^{2\pi}\int_0^1 \max_{\nu\in S^1}|\hat{g}(\rho\boldsymbol{\theta}^\perp,\nu)|^2 d\rho d\theta \le 2\pi \max_{|\boldsymbol{\xi}|\le 1}|\hat{g}(\boldsymbol{\xi},\nu_0)|^2,$$

where to simplify notation we have defined

$$|\hat{g}(\boldsymbol{\xi},\nu_0)| = \max_{\nu\in S^1}|\hat{g}(\boldsymbol{\xi},\nu)|.$$

Let $\{\chi_n(\mathbf{x})\}_{n\ge 1}$ be a sequence of smooth cut-off functions equal to 1 on Ω, the support of g, and equal to 0 at $\mathbf{x}$ such that $d(\mathbf{x},\Omega) > n^{-1}$; and let $\chi_n(\mathbf{x};\boldsymbol{\xi}) = e^{i\mathbf{x}\cdot\boldsymbol{\xi}}\chi_n(\mathbf{x})$. Then we verify that

$$\hat{g}(\boldsymbol{\xi},\nu_0) = \int_{\mathbb{R}^2} e^{-i\mathbf{x}\cdot\boldsymbol{\xi}}\chi_n(\mathbf{x})g(\mathbf{x},\nu_0)d\mathbf{x} = \int_{\mathbb{R}^2}\overline{\chi_n(\mathbf{x};\boldsymbol{\xi})}g(\mathbf{x},\nu_0)d\mathbf{x} = \int_{\mathbb{R}^2}\overline{\hat{\chi_n}(\boldsymbol{\eta};\boldsymbol{\xi})}\hat{g}(\boldsymbol{\eta},\nu_0)d\boldsymbol{\eta},$$

from which we deduce the following bound:

$$|\hat{g}(\boldsymbol{\xi},\nu_0)| \le \inf_n ||\hat{\chi_n}(\cdot;\xi)||_{L^2}||\hat{g}(\cdot,\nu_0)||_{L^2} = \inf_n ||\chi_n||_{L^2}\|g\|_{\hat{2},\infty} \le |\Omega|\|g\|_{\hat{2},\infty}. \tag{35}$$

Similarly, we have

$$\int_0^{2\pi}\int_{-\infty}^0 |\widehat{Rg}(\rho,\theta)|^2 d\rho d\theta \le 2\pi|\Omega|^2\|g\|^2_{\hat{2},\infty} + \int_0^{2\pi}\int_{-\infty}^0 |\widehat{Rg}(\rho,\theta)|^2|\rho| d\rho d\theta.$$

This concludes the proof of the lemma. ∎

LEMMA 3.5. *Let $f\in H^{1/2}(\mathbb{R};L^2(S^1))$ and $\phi(\mathbf{x},\theta)$ be a smooth function such that*

$$\int_{\mathbb{R}^2}\max_{\nu\in S^1}|\hat{\phi}(\boldsymbol{\xi};\nu)|^2(1+|\boldsymbol{\xi}|^2)d\boldsymbol{\xi} < \infty. \tag{36}$$

Then the map $(\mathbf{x},\theta)\to\phi(\mathbf{x},\theta)f(\mathbf{x}\cdot\boldsymbol{\theta}^\perp,\boldsymbol{\theta})$ is in $H^{1/2}(\mathbb{R}^2;L^2(S^1))$.

Proof. We have the following sequence of inequalities:

$$
\begin{aligned}
&\int_0^{2\pi}\int_{\mathbb{R}^2}(1+|\boldsymbol{\xi}|^2)^{\frac12}|\widehat{\phi f}(\boldsymbol{\xi},\theta)|^2 d\boldsymbol{\xi} d\theta\\
&= \int_0^{2\pi}\int_{\mathbb{R}^2}(1+|\boldsymbol{\xi}|^2)^{\frac12}\left|\int_{\mathbb{R}}\hat\phi(\boldsymbol{\xi}\cdot\boldsymbol{\theta},t;\theta)\hat f(\boldsymbol{\xi}\cdot\boldsymbol{\theta}^\perp-t,\theta)dt\right|^2 d\boldsymbol{\xi} d\theta\\
&\le \int_0^{2\pi}\left\{\int_{\mathbb{R}}\left[\int_{\mathbb{R}^2}(1+|\boldsymbol{\xi}|^2)^{\frac12}|\hat\phi(\xi_1,t;\theta)|^2|\hat f(\xi_2-t;\theta)|^2 d\xi_1 d\xi_2\right]^{\frac12}dt\right\}^2 d\theta\\
&\le \int_0^{2\pi}\left\{\int_{\mathbb{R}}\left[\int_{\mathbb{R}^2}(1+|\xi_1|^2)^{\frac12}(1+|\xi_2|^2)^{\frac12}|\hat\phi(\xi_1,t;\theta)|^2|\hat f(\xi_2-t;\theta)|^2 d\xi_1 d\xi_2\right]^{\frac12}dt\right\}^2 d\theta\\
&= \int_0^{2\pi}\left\{\int_{\mathbb{R}}(1+t^2)^{\frac14}\Big[\int_{\mathbb{R}^2}(1+|\xi_1|^2)^{\frac12}|\hat\phi(\xi_1,t;\theta)|^2\left(\frac{1+|\xi_2+t|^2}{1+|t|^2}\right)^{\frac12}\right.\\
&\qquad\qquad\left.\cdot|\hat f(\xi_2;\theta)|^2 d\xi_1 d\xi_2\Big]^{\frac12}dt\right\}^2 d\theta\\
&\le \int_0^{2\pi}\left\{\int_{\mathbb{R}}(1+t^2)^{\frac14}\Big[\int_{\mathbb{R}^2}(1+|\xi_1|^2)^{\frac12}|\hat\phi(\xi_1,t;\theta)|^2(1+|\xi_2|^2)^{\frac12}\right.\\
&\qquad\qquad\left.\cdot|\hat f(\xi_2;\theta)|^2 d\xi_1 d\xi_2\Big]^{\frac12}dt\right\}^2 d\theta\\
&= \int_0^{2\pi}\int_{\mathbb{R}}(1+|\xi_2|^2)^{\frac12}|\hat f(\xi_2;\theta)|^2 d\xi_2\left\{\int_{\mathbb{R}}(1+t^2)^{\frac14}\right.\\
&\qquad\qquad\left.\cdot\left[\int_{\mathbb{R}}(1+|\xi_1|^2)^{\frac12}|\hat\phi(\xi_1,t;\theta)|^2 d\xi_1\right]^{\frac12}dt\right\}^2 d\theta\\
&\le \int_0^{2\pi}\left(\int_{\mathbb{R}}(1+|\xi_2|^2)^{\frac12}|\hat f(\xi_2;\theta)|^2 d\xi_2\int_{\mathbb{R}^2}(1+t^2)^{\frac12}(1+|\xi_1|^2)^{\frac12}\max_{\nu\in S^1}|\hat\phi(\xi_1,t;\nu)|^2 dt\xi_1\right)d\theta\\
&\le \int_0^{2\pi}\int_{\mathbb{R}}(1+|\xi_2|^2)^{\frac12}|\hat f(\xi_2;\theta)|^2 d\xi_2 d\theta\int_{\mathbb{R}^2}(1+|\boldsymbol{\xi}|^2)\max_{\nu\in S^1}|\hat\phi(\boldsymbol{\xi};\nu)|^2 d\boldsymbol{\xi},
\end{aligned}
$$

where we have used the Minkowsky and Cauchy inequalities. From the second line onward, we have used the θ dependent coordinates $\xi_1=\boldsymbol{\xi}\cdot\boldsymbol{\theta}$ and $\xi_2=\boldsymbol{\xi}\cdot\boldsymbol{\theta}^\perp$. □

Recall that $\chi(\mathbf{x})$ defined before (29) is a smooth cut-off function supported in Ω. To simplify notation, let

$$
\begin{aligned}
f_1(\mathbf{x},\theta)&=e^{Da}(\mathbf{x},\theta)\mathrm{trig}(HRa(\mathbf{x}\cdot\boldsymbol{\theta}^\perp,\theta)/2),\\
f_2(s,\theta)&=\mathrm{trig}(HRa(s,\theta)/2),\\
f_3(\mathbf{x},\theta)&=\chi(\mathbf{x})(\boldsymbol{\theta}^\perp\cdot\nabla_{\mathbf{x}})f_1(\mathbf{x},\theta)
\end{aligned}
$$

be smooth functions depending on the attenuation a only, where trig stands for either sin or cos. The composition operator χNR becomes

$$
\chi NRw(\mathbf{x})=\frac{\chi(\mathbf{x})}{4\pi}\int_0^{2\pi}\boldsymbol{\theta}^\perp\cdot\nabla\Big(f_1(\mathbf{x},\theta)H[f_2(s,\theta)R[w](s,\theta)](\mathbf{x}\cdot\boldsymbol{\theta}^\perp,\theta)\Big)d\theta,
$$

where the above is understood as a sum over the values that the functions trig can take in f_1 and f_2. We now follow ideas from [6], though we need to derive some of the estimates in the Fourier domain to characterize the norm of the operator. Let $\phi_n(\mathbf{x})$, $n = 1, 2, \dots$, be an orthonormal basis of $L^2(\Omega)$ and let

$$\chi(\mathbf{x})f_1(\mathbf{x},\theta) = \sum_{n=1}^{\infty} \alpha_n(\theta)\phi_n(\mathbf{x}), \qquad \alpha_n(\theta) = \int_\Omega \chi(\mathbf{x})f_1(\mathbf{x},\theta)\phi_n(\mathbf{x})d\mathbf{x}. \tag{37}$$

PROPOSITION 3.6. *The composition operator χNR maps $L^{\hat{2}}(\Omega; C^0(S^1))$ to $L^2(\Omega)$. Moreover, we have the more explicit characterization*

$$\begin{aligned}\|\chi NRw\|_{L^2} \le \|w\|_{\hat{2},\infty} \int_{\mathbb{R}^2} \max_{\boldsymbol{\theta}} |\widehat{e^{Da}}(\boldsymbol{\xi},\boldsymbol{\theta})|d\boldsymbol{\xi} \Bigg[\int_{\mathbb{R}^2} \max_{\nu\in S^1} |\hat{f}_3(\boldsymbol{\xi};\nu)|^2(1+|\boldsymbol{\xi}|^2)d\boldsymbol{\xi} \\ +2\left(\sum_{n=1}^{\infty} \max_{\nu\in S^1} |\alpha_n(\nu)|^2\right)^{\frac{1}{2}} \int_{-\infty}^{\infty} |\hat{f}_2(s,\nu_0)|(2+4\pi|\Omega|s^2)ds\Bigg].\end{aligned} \tag{38}$$

Proof. Using Lemma 3.4 we obtain that $Rw \in H^{1/2}(\mathbb{R}; L^2(S^1))$. Since $a \in C_0^2(\Omega)$ it is easy to check that $\chi(\mathbf{x})(\boldsymbol{\theta}^\perp \cdot \nabla_x f_1(\mathbf{x},\theta))$ satisfies the smoothness condition (36). From Lemma 3.5 we get that the map

$$(\mathbf{x},\theta) \to \chi(\mathbf{x}) \left(\boldsymbol{\theta}^\perp \cdot \nabla_x f_1(\mathbf{x},\theta)\right) H[f_2(s,\theta)R[w](s,\theta)](\mathbf{x}\cdot\boldsymbol{\theta}^\perp,\theta)$$

is in $H^{1/2}(\mathbb{R}^2; L^2(S^1))$.

Next we show that the operator M defined by

$$Mw(\mathbf{x}) = \int_0^{2\pi} \chi(\mathbf{x})f_1(\mathbf{x},\theta) \left(\boldsymbol{\theta}^\perp \cdot \nabla_x\right) H[f_2(s,\theta)R(w)(s,\theta)](\mathbf{x}\cdot\boldsymbol{\theta}^\perp,\theta)d\theta$$

is bounded from $L^{\hat{2}}(\Omega; C^0(S^1))$ in $L^2(\Omega)$. We have

$$\begin{aligned}&\left\| \int_0^{2\pi} \sum_{n=1}^{\infty} \alpha_n(\theta)\phi_n(\mathbf{x})(\boldsymbol{\theta}^\perp \cdot \nabla)H(f_2Rw(s,\theta))(\mathbf{x}\cdot\boldsymbol{\theta}^\perp,\theta)d\theta \right\|_{L_x^2}^2 \\ &= \sum_{n=1}^{\infty} \left\langle \phi_n; \int_0^{2\pi} \alpha_n(\theta)(\boldsymbol{\theta}^\perp \cdot \nabla)H(f_2Rw(s,\theta)(\mathbf{x}\cdot\boldsymbol{\theta}^\perp,\theta))d\theta \right\rangle_{L_x^2}^2 \\ &\le \sum_{n=1}^{\infty} \left\| \int_0^{2\pi} \alpha_n(\theta)(\boldsymbol{\theta}^\perp \cdot \nabla)H(f_2Rw(s,\theta))(\mathbf{x}\cdot\boldsymbol{\theta}^\perp,\theta)d\theta \right\|_{L_x^2}^2 \\ &\le \sum_{n=1}^{\infty} \left\| F_{\mathbf{x}\to\boldsymbol{\xi}} \left\{ \int_0^{2\pi} \alpha_n(\theta)(\boldsymbol{\theta}^\perp \cdot \nabla)H(f_2Rw(s,\theta))(\mathbf{x}\cdot\boldsymbol{\theta}^\perp,\theta)d\theta \right\}(\boldsymbol{\xi}) \right\|_{L_\xi^2}^2 \\ &= \sum_{n=1}^{\infty} \left\| \alpha_n(\xi_-)\widehat{f_2Rw}(-|\boldsymbol{\xi}|,\xi_-) + \alpha_n(\xi_+)\widehat{f_2Rw}(|\boldsymbol{\xi}|,\xi_+) \right\|_{L_\xi^2}^2 \\ &\le \left(\sum_{n=1}^{\infty} \max_{\nu\in S^1} |\alpha_n(\nu)|^2\right) \left(\|\widehat{f_2Rw}(-|\boldsymbol{\xi}|,\xi_-)\|_{L_\xi^2}^2 + \|\widehat{f_2Rw}(|\boldsymbol{\xi}|,\xi_+)\|_{L_\xi^2}^2 \right).\end{aligned}$$

In the above expressions, $F_{\mathbf{x}\to\boldsymbol{\xi}}$ represents the two-dimensional Fourier transform in the $\mathbf{x}$ variable, while $\hat{\cdot}$ represents the one-dimensional Fourier transform. The angles

$\xi_\pm$ are defined by $\boldsymbol{\xi}\cdot\boldsymbol{\theta}=0$. The estimate above uses the fact that, for any $f\in H^{1/2}(\mathbb{R})$, $F_{\mathbf{x}\to\boldsymbol{\xi}}[\boldsymbol{\theta}^\perp\cdot\nabla Hf(\mathbf{x}\cdot\boldsymbol{\theta}^\perp)]=\hat f(\boldsymbol{\xi}\cdot\boldsymbol{\theta}^\perp)\delta(\boldsymbol{\theta}\cdot\frac{\boldsymbol{\xi}}{|\boldsymbol{\xi}|})$; see [7, pp. 413, 415] for the details. By the Fourier slice theorem we verify that

$$\widehat{f_2Rw}(|\boldsymbol{\xi}|,\xi_+)=\int_{\mathbb{R}}\hat f_2(|\boldsymbol{\xi}|-s,\xi_+)\hat w\left(s\frac{\xi}{|\xi|},\xi_+\right)ds.$$

Let us now define ν_0 and ν_1 as the values of the angles where $\max_{\nu\in S^1}|\hat f_2(\rho,\nu)|$ and $\max_{\nu\in S^1}|\hat w(\rho,\nu)|$ are achieved, respectively. We compute

$$\begin{aligned}
(39)\qquad &\left\{\int_{\mathbb{R}^2}|\widehat{f_2Rw}(|\boldsymbol{\xi}|,\xi_+)|^2d\boldsymbol{\xi}\right\}^{\frac12}=\left\{\int_0^{2\pi}\int_0^\infty|\widehat{f_2Rw}(r,-\boldsymbol{\theta}^\perp)|^2rdrd\theta\right\}^{\frac12}\\
&=\left\{\int_0^{2\pi}\int_0^\infty\left|\int_{\mathbb{R}}\hat f_2(s,-\boldsymbol{\theta}^\perp)\hat w((r-s)\boldsymbol{\theta},-\boldsymbol{\theta}^\perp)|ds\right|^2rdrd\theta\right\}^{\frac12}\\
&\le\left\{\int_0^{2\pi}\int_0^\infty\left(\int_{\mathbb{R}}\max_{\nu\in S^1}|\hat f_2(s,\nu)|\max_{\nu\in S^1}|\hat w((r-s)\boldsymbol{\theta},\nu)|ds\right)^2rdrd\theta\right\}^{\frac12}\\
&=\left\{\int_0^{2\pi}\int_0^\infty\left(\int_{\mathbb{R}}|\hat f_2(s,\nu_0)||\hat w((r-s)\boldsymbol{\theta},\nu_1)|ds\right)^2rdrd\theta\right\}^{\frac12}\\
&\le\int_{\mathbb{R}}\left\{\int_0^{2\pi}\int_0^\infty|\hat f_2(s,\nu_0)|^2|\hat w((r-s)\boldsymbol{\theta},\nu_1)|^2rdrd\theta\right\}^{\frac12}ds\\
&=\int_{\mathbb{R}}|\hat f_2(s,\nu_0)|\left\{\int_0^{2\pi}\int_0^\infty|\hat w((r-s)\boldsymbol{\theta},\nu_1)|^2rdrd\theta\right\}^{\frac12}ds.
\end{aligned}$$

We evaluate the last term by splitting the integral $\int_0^\infty(\dots)ds+\int_{-\infty}^0(\dots)ds$. To estimate

$$\int_0^\infty|\hat f_2(s,\nu_0)|\left\{\int_0^{2\pi}\int_0^\infty|\hat w((r-s)\boldsymbol{\theta},\nu_1)|^2rdrd\theta\right\}^{\frac12}ds,$$

we further split the inner integral into $\int_0^{2s}(\dots)drd\theta+\int_{2s}^\infty(\dots)drd\theta$. We obtain that

$$\begin{aligned}
\int_0^{2\pi}\int_{2s}^\infty|\hat w((r-s)\boldsymbol{\theta},\nu_1)|^2rdrd\theta&\le2\int_0^{2\pi}\int_{2s}^\infty|\hat w((r-s)\boldsymbol{\theta},\nu_1)|^2(r-s)drd\theta\\
&=2\int_{|\boldsymbol{\xi}|\ge s}|\hat w(\boldsymbol{\xi},\nu_1)|^2d\boldsymbol{\xi},
\end{aligned}$$

and

$$\begin{aligned}
\int_0^{2\pi}\int_0^{2s}|\hat w((r-s)\boldsymbol{\theta},\nu_1)|^2rdrd\theta&=\int_0^{2\pi}\int_{-s}^s|\hat w(t\boldsymbol{\theta},\nu_1)|^2(t+s)dtd\theta\\
&=2\int_{|\boldsymbol{\xi}|\le s}|\hat w(\boldsymbol{\xi},\nu_1)|^2d\boldsymbol{\xi}+2s\int_0^{2\pi}\int_0^s|\hat w(t\boldsymbol{\theta},\nu_1)|^2dtd\theta\\
&\le2\int_{|\boldsymbol{\xi}|\le s}|\hat w(\boldsymbol{\xi},\nu_1)|^2d\boldsymbol{\xi}+4\pi s^2\max_{|\boldsymbol{\xi}|\le s}|\hat w(\boldsymbol{\xi},\nu_1)|^2\\
&\le2\int_{|\boldsymbol{\xi}|\le s}|\hat w(\boldsymbol{\xi},\nu_1)|^2d\boldsymbol{\xi}+4\pi s^2|\Omega|\|\hat w(\boldsymbol{\xi},\nu_1)\|^2_{L^2_\xi}.
\end{aligned}$$

The last inequality uses in a crucial way the estimate (35) and the fact that w is compactly supported. We have obtained so far that

$$\int_0^\infty |\hat{f}_2(s,\nu_0)| \left\{ \int_0^{2\pi} \int_0^\infty |\hat{w}((r-s)\boldsymbol{\theta},\nu_1)|^2 r dr d\theta \right\}^{\frac{1}{2}} ds \\ \leq \|w\|_{\hat{2},\infty} \int_0^\infty |\hat{f}_2(s,\nu_0)(2+4\pi|\Omega|s^2) ds.$$

The other contribution is handled similarly:

$$\begin{aligned}
&\int_{-\infty}^0 |\hat{f}_2(s,\nu_0)| \left\{ \int_0^{2\pi} \int_0^\infty |\hat{w}((r-s)\boldsymbol{\theta},\nu_1)|^2 r dr d\theta \right\}^{\frac{1}{2}} ds \\
&\quad = \int_0^\infty |\hat{f}_2(-s,\nu_0)| \left\{ \int_0^{2\pi} \int_0^\infty |\hat{w}((r+s)\boldsymbol{\theta},\nu_1)|^2 r dr d\theta \right\}^{\frac{1}{2}} ds \\
&\quad \leq \int_0^\infty |\hat{f}_2(-s,\nu_0)| \left\{ \int_0^{2\pi} \int_0^\infty |\hat{w}((r+s)\boldsymbol{\theta},\nu_1)|^2 (r+s) dr d\theta \right\}^{\frac{1}{2}} ds \\
&\quad = \|w\|_{\hat{2},\infty} \int_0^\infty |\hat{f}_2(-s,\nu_0)| ds = \|w\|_{\hat{2},\infty} \int_{-\infty}^0 |\hat{f}_2(s,\nu_0)| ds.
\end{aligned}$$

Combined with the estimate in (39), we have obtained that

$$\left\{ \int_{\mathbb{R}^2} |\widehat{f_2 Rw}|(|\boldsymbol{\xi}|,\xi_+)|^2 d\boldsymbol{\xi} \right\}^{\frac{1}{2}} \leq \|w\|_{\hat{2},\infty} \int_{-\infty}^\infty |\hat{f}_2(s,\nu_0)|(2+4\pi|\Omega|s^2) ds. \tag{40}$$

A similar calculation shows that

$$\left\{ \int_{\mathbb{R}^2} |\widehat{f_2 Rw}(-|\boldsymbol{\xi}|,\xi_-)|^2 d\boldsymbol{\xi} \right\}^{\frac{1}{2}} \leq \|w\|_{\hat{2},\infty} \int_{-\infty}^\infty |\hat{f}_2(s,\nu_0)|(2+4\pi|\Omega|s^2) ds. \tag{41}$$

In summary, we have obtained the following estimate for the operator M:

$$\|Mw\|_{L^2} \leq 2\|w\|_{\hat{2},\infty} \left(\sum_{n=1}^\infty \max_{\nu\in S^1} |\alpha_n(\nu)|^2 \right)^{\frac{1}{2}} \int_{-\infty}^\infty |\hat{f}_2(s,\nu_0)|(2+4\pi|\Omega|s^2) ds. \tag{42}$$

This concludes the proof of the proposition. ◻

Note that N has range in $L^2_{\text{loc}}(\mathbb{R}^2)$ and not necessarily in $L^2(\mathbb{R}^2)$. This is where the assumption of the compactness of the support of the source term $f(\mathbf{x})$, which is natural in practice, comes into play. The above estimate shows the role played by the size of the support of the source term.

The proof of Theorem 3.1 follows from the preceding lemmas and proposition. As we have seen, T maps $L^2(\mathbb{R}^2)$ to $L^2(\mathbb{R}^2)$, and let $|||T|||_{L^2\to L^2}$ denote its operator norm. The preceding calculations allow us to obtain the following more explicit version of Theorem 3.1.

Corollary 3.7. *Assume that K is such that $\max_n \langle n\rangle^{\alpha/2}\|\hat{k}_n\|_{L^1(\mathbb{R}^2)} \leq C_\alpha$ for some $\alpha > 1$. Then the operator norm $|||N_K|||_{L^2\to L^2}$ is bounded by the following* *expression:*

$$\begin{aligned}
(43)\qquad &\Bigg[\int_{\mathbb{R}^2}\max_{\nu\in S^1}|\hat f_3(\boldsymbol{\xi};\nu)|^2(1+|\boldsymbol{\xi}|^2)d\xi \\
&+2\left(\sum_{n=1}^{\infty}\max_{\nu\in S^1}|\alpha_n(\nu)|^2\right)^{\frac12}\int_{-\infty}^{\infty}|\hat f_2(s,\nu_0)|(2+4\pi|\Omega|s^2)ds\Bigg] \\
&\times C_\alpha\left(\sum_{-\infty}^{\infty}\frac{1}{\langle n\rangle^\alpha}\right)^{\frac12}\left(\int_{\mathbb{R}^2}\max_{\boldsymbol{\theta}}|\widehat{e^{Da}}(\xi,\boldsymbol{\theta})|d\xi\right)\|e^{-Da}\|_{L^\infty(\mathbb{R}^2\times S^1)}|||T|||_{L^2\to L^2}.
\end{aligned}$$

From [25] we know that the operator norm $|||T|||$ can be bounded independently of the scattering k. Since all the other terms in (43) are independent of k, for C_α small enough, the operator norm of N_K is bounded by a constant less than one. This proves the first part of Theorem 2.1. Note that the constraint on the norm of N_K is only sufficient to solve (30) and by no means necessary. Reconstructions based on (30) thus have a larger domain of validity than what we consider in Theorem 2.1. The reconstruction procedure is based on the fact that the Neumann series

$$f(\mathbf{x})=\sum_{n=0}^{\infty}N_K^nNg(\mathbf{x}) \tag{44}$$

converges in $L^2(\Omega)$ to the solution $f(\mathbf{x})$. This provides us with an explicit reconstruction formula to recover $f(\mathbf{x})$ from the measurements

$$m(s,\theta)=e^{-\frac12 Ra}(s,\theta)g(s,\theta)$$

and concludes the proof of Theorem 2.1. Let us conclude this section with a few remarks.

Remark 3.8. The measurements $m(s,\theta)$ for $s\in\mathbb{R}$ and $0\le\theta<2\pi$ are redundant. Indeed in the case $a\equiv 0$ and $k\equiv 0$, the measurements satisfy $m(s,\theta)=m(-s,\theta+\pi)$ so that the source term can be reconstructed from knowledge of $m(s,\theta)$ on $Z=\mathbb{R}\times(0,\pi)$. When $a\neq 0$, such a redundancy still exists, although it is harder to characterize. Under certain smallness assumptions on $a(\mathbf{x})$, an explicit procedure to reconstruct the source term from m on Z when $k=0$ was proposed in [7] and implemented in [8]. That measurements on Z suffice to determine the source term was recently obtained in [34]; see also [28] in the case of constant absorption. The explicit procedure proposed in [7] can be extended to the case of scattering kernels so that, provided that k is sufficiently small, the source term is uniquely determined by $m(s,\theta)$ on Z.

Remark 3.9. We could have considered more general scattering kernels of the form $k(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\theta}')$ so long as the smoothing of the scattering kernel K imposed in Lemma 3.2 still holds. The description of this smoothing effect in terms of the scattering coefficients is simplified for kernels of the form $k(\mathbf{x},\boldsymbol{\theta}\cdot\boldsymbol{\theta}')$. However, this is the only place where the specific structure of the kernel has been used (except for the subcriticality condition (4), which should hold with $k(\mathbf{x},\boldsymbol{\theta}\cdot\boldsymbol{\theta}')$ replaced by both $k(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\theta}')$ and $k(\mathbf{x},\boldsymbol{\theta}',\boldsymbol{\theta})$; see [13]).

Remark 3.10. The smoothing effect of the scattering kernel described in Lemma 3.2 is rendered necessary (at least some sort of smoothing is) by the behavior of the Radon transform and the inversion operator N. Although NR maps functions in $L^2(\Omega)$ to functions in $L^2(\Omega)$ (since the operator NR is then identity), this is no longer the case for functions in $L^2(\Omega\times S^1)$ that depend nontrivially on θ. We need to map functions from the smaller space $L^{\hat{2}}(\mathbb{R}^2; C^0(S^1))$, which is made possible by the regularizing effect of K.

Remark 3.11. Under appropriate assumptions on the scattering kernel K, (29) is indeed of Fredholm type as the operator N_K can be shown to be compact. Indeed NR is a bounded operator, whereas the operator KT (as well as $Ke^{-Da}T$ for smooth absorption $a(\mathbf{x})$) can be shown to be compact under general assumptions. We refer the reader to [25] for such results and to [16] for connected results on averaging lemmas.

Remark 3.12. The reconstruction of the source term can be obtained by the following iterative scheme. We consider the setting of Corollary 2.3. Let $g(s,\theta) = e^{Ra/2}m(s,\theta)$ be the measurements. We initialize the algorithm as

$$F^{(0)}(\mathbf{x}) = Ng(\mathbf{x}). \tag{45}$$

Provided that $F^{(k)}(\mathbf{x})$ is known, we solve for $u^{(k)}$ in

$$\begin{aligned} &\boldsymbol{\theta}\cdot\nabla_{\mathbf{x}}u^{(k)}(\mathbf{x},\theta) + a(\mathbf{x})u^{(k)}(\mathbf{x},\theta) = K_1u^{(k)}(\mathbf{x},\theta) + F^{(k)}(\mathbf{x}) \quad \text{in } \mathbb{R}^2\times S^1,\\ &\lim_{t\to\infty} u^{(k)}(\mathbf{x}-t\boldsymbol{\theta},\theta) = 0 \quad \text{on } \mathbb{R}^2\times S^1. \end{aligned} \tag{46}$$

We next solve for $v^{(k)}(\mathbf{x},\theta)$ in

$$\begin{aligned} &\boldsymbol{\theta}\cdot\nabla_{\mathbf{x}}v^{(k)}(\mathbf{x},\theta) + a(\mathbf{x})v^{(k)}(\mathbf{x},\theta) = K_1u^{(k)}(\mathbf{x},\theta) \quad \text{in } \mathbb{R}^2\times S^1,\\ &\lim_{t\to\infty} v^{(k)}(\mathbf{x}-t\boldsymbol{\theta},\theta) = 0 \quad \text{on } \mathbb{R}^2\times S^1. \end{aligned} \tag{47}$$

We then compute the new data

$$g^{(k)}(s,\theta) = e^{Ra/2}R_av^{(k)}(s,\theta). \tag{48}$$

Finally, we set the new source term

$$F^{(k+1)}(\mathbf{x}) = N(g-g^{(k)})(\mathbf{x}). \tag{49}$$

We verify that $F^{(k)}(\mathbf{x})$ converges to $F(\mathbf{x}) = K_0u(\mathbf{x}) + f(\mathbf{x})$ in $L^2(\Omega)$ as the above algorithm is equivalent to the Neumann series expansion (44). We then solve for $u(\mathbf{x},\theta)$ and reconstruct the source term $f(\mathbf{x}) = F(\mathbf{x}) - K_0u(\mathbf{x})$.

4. Derivation in three space dimensions. The derivation in the three-dimensional case is very similar to that of the preceding section. The main observation is that the inversion of the X-ray transform can be performed "slice by slice," i.e., "z by z," using outgoing information for angles perpendicular to $\mathbf{e}_z$ only. The inversion with scattering coefficient is again considered as a perturbation of the inversion of the X-ray transform. Mathematically, the main novelty compared to the two-dimensional case is that we need to control the number of photons scattered into the directions orthogonal to $\mathbf{e}_z$.

Upon defining $w(\mathbf{x},\boldsymbol{\theta}) = (e^{Da}u)(\mathbf{x},\boldsymbol{\theta})$, we still obtain that

$$w(\mathbf{x},\boldsymbol{\theta}) = Se^{Da}Ke^{-Da}w(\mathbf{x},\boldsymbol{\theta}) + Se^{Da}f(\mathbf{x},\boldsymbol{\theta}). \tag{50}$$

We define now the trace operator P onto the horizontal directions S^2_H defined in (11). More precisely, P takes functions on $\Omega\times S^2$ onto a function on $\Omega\times S^2_H$ as follows. For $\boldsymbol{\theta} = (\cos\theta\sin\phi, \sin\theta\sin\phi, \cos\phi)$,

$$P[w(\mathbf{x},\boldsymbol{\theta})] = w(\mathbf{x},(\cos\theta,\sin\theta,0)). \tag{51}$$

For $\boldsymbol{\theta} \in S^2_H$, we define the orthogonal vector $\boldsymbol{\theta}^\perp = (-\sin\theta, \cos\theta, 0)$ and the transversal X-ray transform

$$Rf(z, s, \theta) = LSf(z, s, \theta), \tag{52}$$

where the trace operator at infinity is defined by

$$Lw(z, s, \theta) = \lim_{t\to\infty} w(t\boldsymbol{\theta} + s\boldsymbol{\theta}^\perp + z\mathbf{e}_z, \boldsymbol{\theta}). \tag{53}$$

Finally, the transversal attenuated X-ray transform is defined by

$$R_a f(z, s, \theta) = Re^{Da} f(z, s, \theta). \tag{54}$$

Note that

$$R = LS = LPS = LSP = RP, \quad \text{ and } \quad Pe^{Da} = (Pe^{Da})P,$$

so that the rescaled measurements are given by

$$g(z, s, \theta) = e^{\frac{1}{2}Ra}(z, s, \theta)m(z, s, \theta) = R_a f(z, s, \theta) + Re^{Da}PKe^{-Da}Tf(z, s, \theta). \tag{55}$$

Now the operator $R_a f(z, s, \theta)$ can be inverted at each fixed z by using the Novikov formula. Namely, for $\mathbf{x} = (\mathbf{x}', z)$, we define

$$N_3 g(\mathbf{x}) = N[g(z, \cdot, \cdot)](\mathbf{x}'), \tag{56}$$

by applying the two-dimensional operator N to $(s, \theta) \to g(s, \theta, z)$ for each $z \in \mathbb{R}$. We verify that $N_3 R_a = Id$ on functions of $\mathbf{x} \in \mathbb{R}^3$. As in the two-dimensional case, however, $N_3 R_a$ is no longer identity when applied to functions that depend on the variable $\boldsymbol{\theta}$. Thus formally applying the operator N_3 to (55), we obtain that

$$N_3 g(\mathbf{x}) = f(\mathbf{x}) + N_3 Re^{Da}PKe^{-Da}Tf(\mathbf{x}) = (I - N_K)f(\mathbf{x}), \tag{57}$$

where now $N_K = -N_3 Re^{Da}PKe^{-Da}T$. The results of section 3 extend as follows.

PROPOSITION 4.1. *The operator N_K defined above is bounded from $L^2(\Omega)$ to $L^2(\Omega)$.*

Similarly to the planar case, let $\hat{u}(\boldsymbol{\xi}', z, \theta) = \int_{\mathbb{R}^2} e^{-i\mathbf{x}'\cdot\boldsymbol{\xi}'} u(\mathbf{x}', z, \theta) d\mathbf{x}'$ denote the Fourier transform in the first two components of the spatial variable only. We work with the functional space

$$L^{\hat{2}}(\mathbb{R}^2_{\mathbf{x}'} \times \mathbb{R}_z; C^0(S^1)) = \Big\{ u(\mathbf{x}', z, \theta) \text{ s.t. } \hat{u}(\boldsymbol{\xi}', z, \theta) \in L^2(\mathbb{R}^2_{\boldsymbol{\xi}'} \times \mathbb{R}_z; C^0(S^1)) \Big\}, \tag{58}$$

where $L^2(\mathbb{R}^2_{\boldsymbol{\xi}'} \times \mathbb{R}_z; C^0(S^1))$ is endowed with the norm

$$\|\hat{u}\|^2_{L^2(\mathbb{R}^2_{\boldsymbol{\xi}'} \times \mathbb{R}_z; C^0(S^1))} = \int_{\mathbb{R}^3} \max_{\boldsymbol{\theta}\in S^1} |\hat{u}(\boldsymbol{\xi}', z, \theta)|^2 d\boldsymbol{\xi}' dz.$$

The proposition is based on the following lemmas.

LEMMA 4.2. *Consider the decomposition of $k(\mathbf{x}, \cdot) \in L^2[-1, 1]$ in Legendre polynomials*

$$k(\mathbf{x}, t) = \sum_{n=0}^{\infty} k_n(\mathbf{x}) P_n(t), \tag{59}$$

and assume that, for some $\alpha > 1$,

$$\max_{n\in\mathbb{N}}\left(\langle n\rangle^{\alpha-1}\max_{|m|\le n}\max_{\boldsymbol{\theta}\in S_z^2}|Y_{nm}(\boldsymbol{\theta})|^2\int_{\mathbb{R}}\|\hat{k}_n(\cdot,z)\|^2_{L^1(\mathbb{R}^2)}dz\right)\le C. \tag{60}$$

Then the operator PK maps $L^2(\mathbb{R}^3\times S^2)$ to $L^{\hat{2}}(\mathbb{R}^2_{\mathbf{x}'}\times\mathbb{R}_z;C^0(S^1))$.

Proof. Using the summation formula $P_n(\boldsymbol{\theta}\cdot\boldsymbol{\theta}')=\frac{1}{2n+1}\sum_{m=-n}^n Y_{nm}(\boldsymbol{\theta})Y^*_{nm}(\boldsymbol{\theta}')$ (see [17], for instance), we get the following decomposition of the scattering operator:

$$Ku(\mathbf{x},\boldsymbol{\theta})=\int_{S^2}k(\mathbf{x},\boldsymbol{\theta}\cdot\boldsymbol{\theta}')u(\mathbf{x},\boldsymbol{\theta}')d\boldsymbol{\theta}'=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{1}{2n+1}k_n(\mathbf{x})u_{nm}(\mathbf{x})Y_{nm}(\boldsymbol{\theta}), \tag{61}$$

where

$$u_{nm}(\mathbf{x})=\int_{S^2}u(\mathbf{x},\boldsymbol{\theta}')\overline{Y_{nm}(\boldsymbol{\theta}')}d\boldsymbol{\theta}'. \tag{62}$$

The Plancherel identity for the spherical harmonics gives

$$\|u\|^2_{L^2(\mathbb{R}^3\times S^2)}=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\|u_{nm}\|^2_{L^2(\mathbb{R}^3)}. \tag{63}$$

In what follows we consider $\boldsymbol{\theta}\in S^2_H$, i.e., only horizontal directions. To simplify the notation, we denote

$$\beta_n=\max_{|m|\le n}\max_{\boldsymbol{\theta}\in S_H^2}|Y_{nm}(\boldsymbol{\theta})|. \tag{64}$$

Taking the Fourier transform with respect to the horizontal variables in (61), we obtain

$$\begin{aligned}|\widehat{Ku}(\boldsymbol{\xi},z,\boldsymbol{\theta})|^2&=\left|\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{1}{2n+1}(\hat{k}_n*_{\boldsymbol{\xi}}\hat{u}_{nm})(\boldsymbol{\xi},z)Y_{nm}(\boldsymbol{\theta})\right|^2\\&\le\left(\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{\beta_n}{2n+1}\left|(\hat{k}_n*\hat{u}_{nm})(\boldsymbol{\xi},z)\right|\right)^2\\&\le\left(\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{1}{\langle n\rangle^{\alpha}(2n+1)}\right)\left(\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{\langle n\rangle^{\alpha}\beta_n^2}{2n+1}\left|(\hat{k}_n*\hat{u}_{nm})(\boldsymbol{\xi},z)\right|^2\right).\end{aligned}$$

We now take the maximum in $\boldsymbol{\theta}\in S_z^2$ and then integrate in $\boldsymbol{\xi}\in\mathbb{R}^2$. We deduce that

$$\begin{aligned}\int_{\mathbb{R}^2}\max_{\boldsymbol{\theta}\in S_z^2}|\widehat{Ku}(\boldsymbol{\xi},z,\boldsymbol{\theta})|^2d\boldsymbol{\xi}&\le\left(\sum_{n=0}^{\infty}\frac{1}{\langle n\rangle^{\alpha}}\right)\left(\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{\beta_n^2\langle n\rangle^{\alpha}}{2n+1}\|\hat{k}_n*\hat{u}_{nm}(\cdot,z)\|^2_{L^2(\mathbb{R}^2)}\right)\\&\le\left(\sum_{n=0}^{\infty}\frac{1}{\langle n\rangle^{\alpha}}\right)\left(\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{\beta_n^2\langle n\rangle^{\alpha}}{2n+1}\|\hat{k}_n(\cdot,z)\|^2_{L^1(\mathbb{R}^2)}\|\hat{u}_{nm}(\cdot,z)\|^2_{L^2(\mathbb{R}^2)}\right)\\&\le\left(\max_{n\in\mathbb{N}}\frac{\beta_n^2\langle n\rangle^{\alpha}}{2n+1}\|\hat{k}_n(\cdot,z)\|^2_{L^1(\mathbb{R}^2)}\right)\left(\sum_{n=0}^{\infty}\frac{1}{\langle n\rangle^{\alpha}}\right)\left(\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\|\hat{u}_{nm}(\cdot,z)\|^2_{L^2(\mathbb{R}^2)}\right).\end{aligned}$$

It remains to integrate in $z \in \mathbb{R}$ to obtain that

$$||Ku||^2_{L^{\hat{2}}(\mathbb{R}^2_{\mathbf{x}'}\times\mathbb{R}_z;C^0(S^1))} \leq \left(\max_{n\in\mathbb{N}} \frac{\beta_n^2\langle n\rangle^\alpha}{2n+1}\int_{\mathbb{R}}\|\hat{k}_n(\cdot,z)\|^2_{L^1(\mathbb{R}^2)}dz\right) \cdot \left(\sum_{n=0}^\infty \frac{1}{\langle n\rangle^\alpha}\right)||u||^2_{L^2(\mathbb{R}^3\times S^2)}.$$

This concludes the proof of the lemma. □

LEMMA 4.3. *The operator* N_3R *maps* $L^{\hat{2}}(\mathbb{R}^2_{\mathbf{x}'}\times\mathbb{R}_z;C^0(S^1))$ *to* $L^2(\Omega)$.

Proof. This is a direct consequence of Lemma 3.6:

$$\|N_3Rf\|^2_{L^2(\Omega)} = \int_{\mathbb{R}}\int_{\mathbb{R}^2}|N_3Rf(\mathbf{x}',z)|^2\,d\mathbf{x}'dz = \int_{\mathbb{R}}\int_{\mathbb{R}^2}|[NRf(\cdot,\cdot,z)](\mathbf{x}')|^2\,d\mathbf{x}'dz$$

$$\leq C\int_{\mathbb{R}}dz\left\{\int_{\mathbb{R}^2}\max_{\boldsymbol{\theta}\in S_z^2}\left|\hat{f}(\boldsymbol{\xi},\boldsymbol{\theta},z)\right|^2 d\boldsymbol{\xi}\right\} = \|f\|^2_{L^{\hat{2}}(\mathbb{R}^2_{\mathbf{x}'}\times\mathbb{R}_z;C^0(S^1))}. \tag{65}$$

□

The rest of the proof of Theorem 2.2 is similar to that of Theorem 2.1. Provided that scattering is sufficiently small, the Neumann series expansion

$$f(\mathbf{x}) = \sum_{n=0}^\infty N_K^n N_3 g(\mathbf{x}) \tag{66}$$

converges in $L^2(\Omega)$ strongly to the solution $f(\mathbf{x})$.

The remarks at the end of section 3 still hold in the three-dimensional setting. The main difference between the two-dimensional and three-dimensional theories is that the scattering operator is required to be more regularizing in three dimensions than in two dimensions. This is so because the three-dimensional reconstruction is based on measurements of the outgoing distribution for directions that are orthogonal to the vertical axis $\mathbf{e}_z$. The influence of the geometry on the norm of N_K could be characterized in the three-dimensional setting as we have done for the two-dimensional setting in Corollary 3.7, although we shall not do so here.

Acknowledgments. We thank the referees for valuable comments and suggestions. This work was completed in part while both authors visited the Institute for Pure and Applied Mathematics at UCLA during the program on Inverse Problems and Gunther Uhlmann at the University of Washington. We would like to thank both organizations for their hospitality.

REFERENCES

[1] D. S. Anikonov, A. E. Kovtanyuk, and I. V. Prokhorov, *Transport Equation and Tomography*, Inverse Ill-posed Probl. Ser. 30, VSP, Utrecht, The Netherlands, 2002.

[2] V. S. Antyufeev and A. N. Bondarenko, *X-ray tomography in scattering media*, SIAM J. Appl. Math., 56 (1996), pp. 573–587.

[3] E. V. Arbuzov, A. L. Bukhgeim, and S. G. Kazantsev, *Two-dimensional tomography problems and the theory of A-analytic functions*, Siberian Adv. Math., 8 (1998), pp. 1–20.

[4] S. R. Arridge, *Optical tomography in medical imaging*, Inverse Problems, 15 (1999), pp. R41–R93.

[5] H. Babovsky, *Identification of scattering media from reflected flows*, SIAM J. Appl. Math., 51 (1991), pp. 1676–1704.

[6] G. Bal, *Inverse problems for homogeneous transport equations. Part* II: *Multidimensional case*, Inverse Problems, 16 (2000), pp. 1013–1028.

[7] G. Bal, *On the attenuated Radon transform with full and partial measurements*, Inverse Problems, 20 (2004), pp. 399–419.

[8] G. Bal and P. Moireau, *Fast numerical inversion of the attenuated Radon transform with full and partial measurements*, Inverse Problems, 20 (2004), pp. 1137–1164.

[9] J. Boman and J. O. Strömberg, *Novikov's inversion formula for the attenuated Radon transform—a new approach*, J. Geom. Anal., 14 (2004), pp. 185–198.

[10] K. M. Case and P. F. Zweifel, *Linear Transport Theory*, Addison–Wesley Series in Nuclear Engineering, Addison–Wesley, Reading, MA, 1967.

[11] J. Chang, R. L. Barbour, H. Graber, and R. Aronson, *Fluorescence Optical Tomography*, in Experimental and Numerical Methods for Solving Ill-Posed Inverse Problems: Medical and Nonmedical Applications, R. L. Barbour, M. J. Carvlin, and M. A. Fiddy, eds., Proc. SPIE 2570, SPIE, Bellingham, WA, 1995, pp. 59–72.

[12] M. Choulli and P. Stefanov, *Reconstruction of the coefficients of the stationary transport equation from boundary measurements*, Inverse Problems, 12 (1996), pp. L19–L23.

[13] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology, Vol.* 6, Springer-Verlag, Berlin, 1993.

[14] D. V. Finch, *Uniqueness for the attenuated X-ray transform in the physical range*, Inverse Problems, 2 (1986), pp. 197–203.

[15] D. V. Finch, *The attenuated X-ray transform: Recent developments*, in Inside Out, Inverse Problems, and Applications, G. Uhlmann, ed., MSRI Publications 47, Cambridge University Press, Cambridge, UK, 2004, pp. 47–66.

[16] F. Golse, P.-L. Lions, B. Perthame, and R. Sentis, *Regularity of the moments of the solution of a transport equation*, J. Funct. Anal., 76 (1988), pp. 110–125.

[17] H. Hochstadt, *The Functions of Mathematical Physics*, Dover, New York, 1986.

[18] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer-Verlag, New York, 1998.

[19] S. G. Kazantsev and A. A. Bukhgeim, *Singular value decomposition for the* 2*D Fan-beam Radon transform of tensor fields*, J. Inverse Ill-Posed Probl., 12 (2004), pp. 245–278.

[20] L. A. Kunyansky, *A new SPECT reconstruction algorithm based on the Novikov's explicit inversion formula*, Inverse Problems, 17 (2001), pp. 293–306.

[21] J. Lakowicz, *Principles of Fluorescence Spectroscopy*, Plenum Press, New York, 1983.

[22] E. W. Larsen, *The inverse source problem in radiative transfer*, J. Quant. Spect. Radiat. Transfer, 15 (1975), pp. 1–5.

[23] E. W. Larsen, *Solution of three-dimensional inverse transport problems*, Transport Theory Statist. Phys., 17 (1988), pp. 147–167.

[24] N. J. McCormick, *Inverse radiative transfer problems: A review*, Nucl. Sci. Eng., 112 (1992), pp. 185–198.

[25] M. Mokhtar-Kharroubi, *Mathematical Topics in Neutron Transport Theory*, World Scientific, Singapore, 1997.

[26] F. Natterer, *Inversion of the attenuated Radon transform*, Inverse Problems, 17 (2001), pp. 113–119.

[27] F. Natterer and F. Wübbeling, *Mathematical Methods in Image Reconstruction*, SIAM Monogr. Math. Model. Comput. 5, SIAM, Philadelphia, 2001.

[28] F. Noo and J.-M. Wagner, *Image reconstruction in* 2*D SPECT with* 180 ° *acquisition*, Inverse Problems, 17 (2001), pp. 1357–1371.

[29] R. G. Novikov, *An inversion formula for the attenuated X-ray transformation*, Ark. Mat., 40 (2002), pp. 145–167.

[30] R. G. Novikov, *On the range characterization for the two-dimensional attenuated X-ray transformation*, Inverse Problems, 18 (2002), pp. 677–700.

[31] V. Ntziachristos and R. Weissleder, *Experimental three-dimensional fluorescence reconstruction of diffuse media by use of a normalized Born approximation*, Opt. Lett., 26 (2001), pp. 893–895.

[32] V. Ntziachristos, A. G. Yodh, M. Schnall, and B. Chance, *Concurrent MRI and diffuse optical tomography of breast after indocyanine green enhancement*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 2767–2772.

[33] A. N. Panchenko, *Inverse source problem of radiative transfer: A special case of the attenuated Radon transform*, Inverse Problems, 9 (1993), pp. 321–337.

[34] H. Rullgård, *Stability of the inverse problem for the attenuated Radon transform with* 180^0 *data*, Inverse Problems, 20 (2004), pp. 781–797.

[35] V. A. Sharafutdinov, *Inverse problem of determining a source in the stationary transport equation on a Riemannian manifold*, Mat. Vopr. Teor. Rasprostr. Voln., 26 (1997), pp. 236–242 (in Russian); translation in J. Math. Sci. (New York), 96 (1999), pp. 3430–3433.

[36] V. A. Sharafutdinov, *Ray Transform on Riemannian Manifolds*, Lecture Notes at the University of Washington, Seattle, WA, 1999.

[37] C. E. Siewert, *An inverse source problem in radiative transfer*, J. Quant. Spect. Radiat. Transfer, 50 (1993), pp. 603–609.

[38] P. Stefanov and G. Uhlmann, *Optical tomography in two dimensions*, Methods Appl. Anal., 10 (2003), pp. 1–9.

[39] A. Tamasan, *An inverse boundary value problem in two-dimensional transport*, Inverse Problems, 18 (2002), pp. 209–219.

[40] R. Weissleder and U. Mahmood, *Molecular imaging*, Radiology, 219 (2001), pp. 316–333.

[41] R. Weissleder, C.-H. Tung, U. Mahmood, and A. Bogdanov, *In vivo imaging of tumors with protease-activated near-infrared fluorescent probes*, Nature Biotechnology, 17 (1999), pp. 375–378.

[42] H. C. Yi, R. Sanchez, and N. J. McCormick, *Bioluminescence estimation from ocean in situ irradiances*, Applied Optics, 31 (1992), pp. 822–830.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 77–102

INTERACTION OF A BULK AND A SURFACE ENERGY WITH A GEOMETRICAL CONSTRAINT*

ANTONIN CHAMBOLLE† AND MARGHERITA SOLCI‡

Abstract. This study is an attempt to generalize in dimension higher than two the mathematical results in [E. Bonnetier and A. Chambolle, *SIAM J. Appl. Math.*, 62 (2002), pp. 1093–1121]. It is the study of a physical system whose equilibrium is the result of a competition between an elastic energy inside a domain and a surface tension, proportional to the perimeter of the domain. The domain is constrained to remain a subgraph. It is shown by Bonnetier and Chambolle that several phenomena appear at various scales as a result of this competition. In this paper, we focus on establishing a sound mathematical framework for this problem in a higher dimension. We also provide an approximation, based on a phase-field representation of the domain.

Key words. epitaxial growth, surface tension, phase-field approximation, diffuse interface, Γ-convergence

AMS subject classifications. 49J45, 74N20

DOI. 10.1137/060649173

1. Introduction.

In this paper, we seek to extend to a higher dimension the results of Bonnetier and Chambolle in [9]. There, the authors modelize the physical system which consists of a thin film of atoms deposited on a substrate, made of a different crystal. Such systems are common in the engineering of devices such as electronic chips, which are obtained by growing epitaxial films on flat surfaces.

In such a situation, the misfit between the crystalline lattices of the substrate and the film induces strains in the film. To release the elastic energy due to these strains, the atoms of the free surface of the film may diffuse and a reorganization occurs in the film. The result of this mechanism is a competition between the surface energy of the crystal and the bulk elastic energy. The former is roughly proportional to the free surface of the crystal and therefore favors flat configurations. The bulk energy, on the contrary, is best released if oscillatory patterns develop. We refer to [9] and the former study [11] for a more complete explanation of the phenomenon and for references on "stress driven rearrangement instabilities" and epitaxial growth.

Here, we restrict our study to the mathematical model which is proposed in [9] in dimension two. We extend to a higher dimension the relaxation result (implicitly contained in Lemma 2.1 and Theorem 2.2 in [9]) and show the correctness of the phase-field approximation, extending [9, Thm. 3.1]. Observe, however, that in that paper the bulk energy is a linearized elasticity energy that involves the symmetrized gradient of the displacement. It seems that up to now, the theory of "special bounded deformation" functions [6, 8] is not developed well enough to make possible the generalization of our results to that case so that we only work with $W^{1,p}$-coercive bulk energies. Alternatively, we could have decided to impose an additional (artificial) L^∞ constraint to the displacements, in which case the extension to linearized elasticity energies would have been relatively easy (see, for instance, [16]).

*Received by the editors January 5, 2006; accepted for publication (in revised form) October 30, 2006; published electronically April 13, 2007.
http://www.siam.org/journals/sima/39-1/64917.html

†CMAP, Ecole Polytechnique, CNRS, 91128 Palaiseau, France (antonin.chambolle@polytechnique.fr).
‡DAP, Università di Sassari, Palazzo Pou Salit, 07041 Alghero, Italy (margherita@uniss.it).

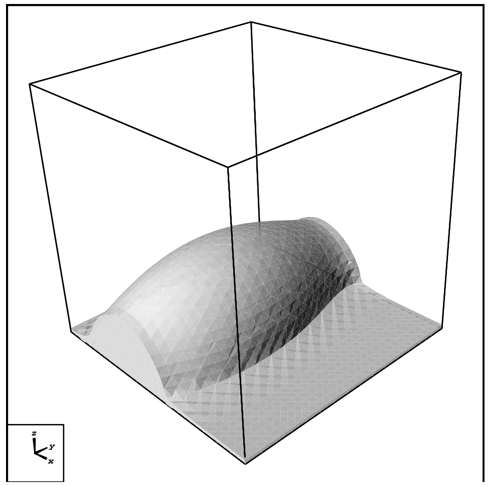


FIG. 1. *Example of an "island."*

Numerical experiments conducted by Jouve and Bonnetier [10] show that the phase-field energy introduced in section 5, in dimension three, yields results similar to the two-dimensional (2D) plots in [9]. See Figure 1 which shows how an island is formed, as a result of the competition between the surface energy and the strains in the material. Here the stretch (the lattice misfit) along the x-direction is stronger than in the y-direction, explaining the shape of the island. (In this example, the bulk energy is a linearized elasticity energy.)

To be precise, we consider in this paper a displacement in a material domain which is the subgraph of an unknown nonnegative function h. Assuming h is defined on an open Lipschitz set $\omega \subset \mathbb{R}^{N-1}$, the displacement u will be defined on the subgraph $\Omega_h := \{x = (x', x_N) \in \omega \times (0, +\infty) \,:\, x_N < h(x')\}$ of h. We will consider energies of the form

$$F(u,h) = \int_{\Omega_h} W(\nabla u)\,dx + \int_\omega \sqrt{1+|\nabla h|^2}\,dx',$$

where u sastisfies a prescribed boundary condition on the boundary $\omega \times \{0\}$. In this paper, ω will be the $(N-1)$-dimensional torus and the boundary condition of u on "$\partial\omega$" will be of periodic type, as in [9] (however, adaption to other situations will not be difficult as long as $\partial\omega$ is Lipschitz).

The goal of our paper is to show that the relaxed functional of F can be written

$$\overline{F}(u,h) = \int_{\Omega_h} W(\nabla u)\,dx + \mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(\Sigma),$$

where Σ, the "internal" discontinuity set of u "inside" the subgraph Ω_h of h (which is now a function of bounded variation (BV)), will be a "vertical" rectifiable set so that $\Omega_h \cup \Sigma$ can be viewed as a generalized subgraph.

In an article written almost simultaneously by Braides and the authors of the present paper [13], a similar problem is studied, without the constraint that the domain is the subgraph of a function. Although this may seem more general, showing that "recovery" sequences can be built, so that $\overline{F}$ is not only a lower bound but also an upper bound for the l.s.c. envelope of F, is considerably more difficult in our setting, since the sequence which is found must satisfy the constraint, and therefore has to be built in a constructive way (and not using some general existence result). This construction follows the discretization/reinterpolation technique introduced in [15, 16]. On the other hand, the lower bound in this work is almost a straightforward consequence of [13].

Finally, the last section in this paper deals with the phase-field approximation of $\overline{F}$, using the same approach as in [9].

2. Setting of the problem and statement of the result.

2.1. Functions of bounded variation. We start by recalling some definitions and results, useful in this paper, concerning functions of bounded variation; for this topic, we refer essentially to [7].

Let Ω be an open subset of $\mathbb{R}^N$. Given $u \in L^1(\Omega)$, its total variation is defined as

$$\sup \left\{ \int_\Omega u \operatorname{div} \psi \, dx \ : \ \psi \in C_c^\infty(\Omega; \mathbb{R}^N), \ |\psi(x)| \le 1 \ \forall x \in \Omega \right\} .$$

One may check that it is finite if and only if the distributional derivative Du of u is a bounded Radon measure in Ω. In this case, the total variation of u is equal to the total variation of the measure Du and is classically denoted by $|Du|(\Omega)$.

At each $x \in \Omega$, one can define upper and lower values of u as follows: The upper value is

$$u_+x(u_+(x)) \ = \ \inf \left\{ t \in [-\infty, +\infty] \ : \ \limsup_{\rho \to 0} \frac{|\{y \in \Omega \ : \ u(y) > t\}| \cap B_\rho(x)}{|B_\rho(x)|} = 0 \right\} ,$$

where $B_\rho(x)$ is the ball of radius ρ centered at x. The lower value is simply $-(-u)_+$. Defining the "jump set" of u as $S_u := \{x \in \Omega : u_-(x) < u_+(x)\}$, one can show that if $u \in BV(\Omega)$, S_u is a $(\mathcal{H}^{N-1}, N-1)$-rectifiable set (in the sense of Federer [19]) so that it admits a normal $\nu_u(x)$ at $\mathcal{H}^{N-1}$-a.e. $x \in S_u$, and Du decomposes as

$$Du \ = \ \nabla u(x) \, dx \ + \ (u_+(x) - u_-(x)) \nu_u(x) \, d\mathcal{H}^{N-1} \llcorner S_u(x) \ + \ D^c u,$$

where $D^c u$, the "Cantor part," is singular with respect to the Lebesgue measure and vanishes on any set with finite $(N-1)$-dimensional Hausdorff measure. The Radon–Nikodym derivative of Du with respect to the Lebesgue measure dx, denoted by $\nabla u(x)$, is a.e. the "approximate gradient" of u at x; see [7]. Of course, if $u \in W^{1,1}(\Omega)$, it coincides with the weak gradient.

Up to now, we have considered real-valued functions. If $u \colon \Omega \to \mathbb{R}^d$ is vector-valued, S_u will be the union of the jump sets of the d components of u. One shows, then, that when two of these jump sets intersect, the corresponding normals coincide $\mathcal{H}^{N-1}$-everywhere in the intersection up to a change of sign. The jump part of the derivative Du is given by $(u_+ - u_-) \otimes \nu_u \, d\mathcal{H}^{N-1} \llcorner S_u$, where now u_+ and u_- are not the "upper" and "lower" values (since there is no natural order in $\mathbb{R}^d$) but the orientation depends on the choice of the direction of the normal ν_u (the triple (u_-, u_+, ν_u) being equivalent to $(u_+, u_-, -\nu_u)$).

The space $SBV(\Omega)$ is defined as the subset of $BV(\Omega)$ of functions u such that $D^c u = 0$, that is, Du is absolutely continuous with respect to $dx + \mathcal{H}^{N-1} \llcorner S_u$. Then, for $p > 1$, we say that a function $u \colon \Omega \to \mathbb{R}$ belongs to the space $SBV_p(\Omega)$ if $u \in SBV(\Omega)$, $\nabla u \in L^p(\Omega; \mathbb{R}^N)$, and $\mathcal{H}^{N-1}(S_u) < +\infty$.

We say that a function $u \in L^1(\Omega)$ is a *generalized function of bounded variation* ($u \in GBV(\Omega)$) if $u^T := (-T) \vee u \wedge T$ belongs to $BV(\Omega)$ for every $T \geq 0$. If $u \in GBV(\Omega)$, setting $S_u = \bigcup_{T>0} S_{u^T}$, a truncation argument allows to define the traces $u_-(x)$ and $u_+(x)$ for a.e. $x \in S_u$. Defining, for $u \in GBV(\Omega)$, the Cantor part of the derivative as $|D^c u| = \sup_{T>0} |D^c u^T|$, we say that a function u in $GBV(\Omega)$ belongs to $GSBV(\Omega)$ if $|D^c u| = 0$, and moreover u in $GSBV(\Omega)$ belongs to $GSBV_p(\Omega)$ for $p > 1$ if $\nabla u \in L^p(\Omega; \mathbb{R}^N)$ and $\mathcal{H}^{N-1}(S_u) < +\infty$.

The following compactness result for SBV is proven in [3, 5] (see also [7, Thm. 4.8]).

THEOREM 2.1 (compactness in SBV). *Let $(u_n)_n \subset SBV(\Omega)$ satisfy*

$$\sup_n \left\{ \int_\Omega |\nabla u_n|^p \, dx + \mathcal{H}^{N-1}(S_{u_n}) \right\} < +\infty,$$

with u_n uniformly bounded in $L^\infty(\Omega)$. Then, there exist a subsequence $(u_{n_k})_k$ and $u \in SBV_p(\Omega)$ such that $u_{n_k} \to u$ a.e. in Ω, $\nabla u_{n_k} \rightharpoonup \nabla u$ in $L^p(\Omega; \mathbb{R}^N)$, and

$$\mathcal{H}^{N-1}(S_u) \leq \liminf_{k\to\infty} \mathcal{H}^{N-1}(S_{u_{n_k}}).$$

If u_n is bounded only in $L^1(\Omega)$, one shows easily by truncation that the results still hold, with $u \in GSBV_p(\Omega)$.

2.2. Subgraphs of finite perimeter. In this paper, to simplify, ω is the torus $(\mathbb{R}/\mathbb{Z})^{N-1}$; however, the extension of our results to the case of a Lipschitz bounded open subset of $\mathbb{R}^{N-1}$ does not raise any difficulties. A generic point $x \in \omega \times \mathbb{R}$ will be denoted by (x', x_N), $x' = (x_1, \ldots, x_{N-1}) \in \omega$, $x_N \in \mathbb{R}$. For $h : \omega \to \mathbb{R}_+$ measurable, we consider

$$\Omega_h = \{x \in \omega \times (-1, +\infty) \;:\; x_N < h(x')\} \quad \text{and}$$

$$\Omega_h^+ = \{x \in \omega \times (0, +\infty) \;:\; x_N < h(x')\} = \Omega_h \cap (\omega \times (0, +\infty)).$$

If $h \in BV(\omega; \mathbb{R}_+)$, the set Ω_h has a finite perimeter in the sense of Caccioppoli in $\omega \times (-1, +\infty)$ (that is, $|D\chi_{\Omega_h}|(\omega \times (-1, +\infty)) \leq |\omega| + |Dh|(\omega) < +\infty$, so that $\chi_{\Omega_h} \in BV(\omega \times (-1, +\infty))$). At each point $\xi \in \omega$ one can define the upper and lower values $h_+(\xi)$ and $h_-(\xi)$ as in the previous section. As before, it is known that $h_+ = h_-$ a.e. in ω and the set of points where $h_- < h_+$, called the jump set of h, is denoted by S_h. Then, if $x = (x', x_N) \in \omega \times (-1, +\infty)$, $x_N < h_-(x') \Rightarrow x \in \Omega_h^1$ (the set of points where Ω_h has Lebesgue density 1), $x_N > h_+(x') \Rightarrow x \in \Omega_h^0$ (the set of points where it has density 0), and $\partial_* \Omega_h = \omega \times (-1, +\infty) \setminus (\Omega_h^0 \cup \Omega_h^1)$, the measure-theoretical boundary is a subset of (and $\mathcal{H}^{N-1}$-a.e. equal to) $\bigcup_{\xi \in \omega} \{\xi\} \times [h_-(\xi), h_+(\xi)]$. It is known that the measure-theoretical boundary is $\mathcal{H}^{N-1}$-a.e. equal to a subset $\partial^* \Omega_h$ called the "reduced boundary" of De Giorgi, which contains only points x where the blowups $(\Omega_h - x)/\rho$ converge as $\rho \to 0$ (in $L^1_{\mathrm{loc}}(\mathbb{R}^N)$) to a half-space of outer normal $\nu_{\Omega_h}(x)$ (hence, Ω_h has density exactly $1/2$ at x).

Let us emphasize the fact that the boundaries $\partial \Omega_f$, $\partial_* \Omega_h$ will always, in this paper, be intended as boundaries *inside* $\omega \times (-1, +\infty)$, that is, they do not contain $\omega \times \{-1\}$.

2.3. The relaxation result. Let $W \colon M^{d\times N} \to [0, +\infty)$, with $d \geq 1$, be a continuous and quasi-convex function satisfying a p-growth condition. Let $u^0 \in W^{1,p}(\omega \times (-1, 0); \mathbb{R}^d)$.

For $h \in C^1(\omega;[0,+\infty))$ and $u \in W^{1,p}(\Omega_h^+;\mathbb{R}^d)$, with $u = u^0$ in $\omega \times \{0\}$, we set

$$F(u,h) = \int_{\Omega_h^+} W(\nabla u)\,dx + \int_\omega \sqrt{1+|\nabla h|^2}\,dx';$$

clearly, the same definition can be given for $u \in L^1(\omega \times (0,+\infty);\mathbb{R}^d)$ such that the restriction to Ω_h^+ satisfies the previous properties. Moreover, we define $F(u,h) = +\infty$ otherwise in $L^1(\omega \times (0,+\infty);\mathbb{R}^d) \times BV(\omega;[0,+\infty))$.

It is clear that equivalently one can write that $u \in W^{1,p}(\Omega_h;\mathbb{R}^d)$, with $u = u^0$ in $\omega \times (-1,0)$.

The main result of this paper is the proof of the following relaxation result for the functional F, here written in the case $d = 1$ (for the general case, see the fourth remark in section 2.4).

THEOREM 2.2. *The l.s.c. envelope of the functional F, with respect to the $L^1(\omega \times (0,+\infty)) \times L^1(\omega)$ topology, is the functional $\overline{F}\colon L^1(\omega \times (0,+\infty)) \times L^1(\omega) \to [0,+\infty]$ defined as*

$$\overline{F}(u,h) = \begin{cases} \displaystyle\int_{\Omega_h^+} W(\nabla u)\,dx + \mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(S'_u \cap \Omega_h^1) & \\ \qquad \text{if } \ h \in BV(\omega;[0,+\infty)) \ \text{ and } \ u\chi_{\Omega_h^+} \in GSBV(\omega \times (0,+\infty)), & \\ +\infty \qquad \text{otherwise,} & \end{cases}$$

where

$$S'_u \;=\; \{(x', x_N + t) \,:\, x \in S_u\,, t \geq 0\}\;.$$

Observe that, denoting $\Sigma = S'_u \cap \Omega_h^1$, Σ is a "vertical" rectifiable set, and we will sometimes write $\Gamma = \partial_*\Omega_h \cup \Sigma$, the "generalized" interface.

The proof of Theorem 2.2 will be given by showing a lower and an upper bound, respectively, in section 3 (Proposition 3.1) and in section 4 (Proposition 4.1); the thesis of Theorem 2.2 immediately follows from these results.

2.4. Some remarks.

1. In [13], a similar result is shown with mainly two differences, which both follow from the constraint that the set where u is defined is a subgraph: In the lim inf inequality, we have to keep track of the vertical parts of the boundary (S'_u) that might not be in the jump set of u (that is, one might have $(S'_u \setminus S_u) \cap \Omega_h^1 \neq \emptyset$). In the lim sup inequality, one needs to build a recovery sequence which remains a subgraph, leading to a much more complex proof than in [13].

2. In [9], one also considers the case where the surface tension for the substrate (of boundary $\omega \times \{0\}$), σ_S, can be different from the surface tension σ_C of the crystal (of boundary $\partial\Omega_h \cap (\omega \times (0,+\infty))$ if h is smooth). In this case, two different phenomena occur, depending on the fact that $\sigma_S \leq \sigma_C$ or $\sigma_C < \sigma_S$. In the latter case, it is always energetically convenient to cover (or "wet") all the surface of the substrate with an infinitesimal layer of crystal, so that the global surface tension in the relaxed energy is σ_C. In case σ_S is less than σ_C, then parts of the substrate might remain uncovered by the crystal, and the surface energy in the relaxed functional will be given by

$$\sigma_C(\mathcal{H}^{N-1}(\partial_*\Omega_h \cap (\omega \times (0,+\infty))) \,+\, 2\mathcal{H}^{N-1}(S'_u \cap \Omega_h^1)) \\ \qquad\qquad + \,\sigma_S\mathcal{H}^{N-1}(\{x' \in \omega \,:\, h(x') = 0\})\,.$$

We do not prove this result here; we fear it would make the paper harder to read, mostly because of the notation. See also Remark 4.4.

3. Still in [9], the (2D) functional F is minimized with an additional volume constraint ($\int_\omega h\,dx = 1$). It is easy to show that the relaxed functional $\overline{F}$ does not change under this constraint—see Remark 4.2 below.

4. In what follows, we will assume that $d = 1$, u is scalar, and hence W is convex. In the vectorial case, one has to assume that W is a continuous, quasi-convex function of ∇u with growth p (that is, bounded from below and above by functions of the form $a + b|\nabla u|^p$, with $b > 0$). Then, the lower bound (Proposition 3.1) remains the same thanks to results of semicontinuity for quasi-convex integrands, due to Ambrosio [4] in the SBV case (see also [7]), and to Kristensen [23] in the general case). The proof of the upper bound (Proposition 4.1, in which W does not appear) can be written with a scalar or vectorial u without any change. Then, its generalization to the Lagrangian W follows from the continuity and p-growth assumptions, as in the scalar case.

5. In [9] and the problem mentioned in the introduction, it is not u but $u - x_1$ which is 1-periodic in the first variable. Here, to simplify, everything is written with $u \in GSBV_p(\omega \times (-1, +\infty))$; that is, u is periodic in the $(N-1)$ first directions (we recall ω is the $(N-1)$-dimensional torus). Adapting the results to extend them to the case where, for instance, $u - \alpha(x_1, 0, \dots, 0) \in GSBV_p(\omega \times (-1, +\infty))$, $\alpha > 0$, would not be difficult.

3. A lower bound for the relaxed envelope of *F*. In this section we obtain a lower bound for the relaxed functional $\overline{F}$ by proving the following proposition.

PROPOSITION 3.1. *For every sequence $(u_n, h_n) \in W^{1,p}(\Omega_{h_n}) \times C^1(\omega; [0, +\infty))$, with $u_n = u_0$ in $\omega \times (-1, 0)$, such that*

$$\sup_n F(u_n, h_n) < +\infty,$$

there exist $h \in BV(\omega; [0, +\infty))$ and $u \in GSBV(\omega \times (0, +\infty))$ (with $u = 0$ out of Ω_h) such that $\chi_{\Omega_{h_n}} u_n \to u$ in $L^1(\omega \times (0, +\infty))$, $h_n \to h$ in $L^1(\omega)$,

$$\int_{\Omega_h^+} |\nabla u(x)|^p\,dx \le \liminf_{n\to\infty} \int_{\Omega_{h_n}^+} |\nabla u_n(x)|^p\,dx\,, \tag{1}$$

and

$$\mathcal{H}^{N-1}(\partial_* \Omega_h) + 2\mathcal{H}^{N-1}(S'_u \cap \Omega_h^1) \le \liminf_{n\to\infty} \int_\omega \sqrt{1 + |\nabla h_n(x')|^2}\,dx'\,. \tag{2}$$

This proposition implies immediately the lower bound for the relaxed envelope of F, that is, the first part of the proof of Theorem 2.2. Indeed, we obtain in the proof that the sequence $(u_n)_n$ converges in fact weakly in the $W^{1,p}$-topology, and since the function W is l.s.c. and convex, with growth p, the functional $G(u) = \int_{\Omega_h^+} W(\nabla u)\,dx$ is weakly l.s.c. in $W^{1,p}$; then, in the same hypothesis, we get the inequality

$$\begin{aligned} &\int_{\Omega_h^+} W(\nabla u(x))\,dx + \mathcal{H}^{N-1}(\partial_* \Omega_h) + 2\mathcal{H}^{N-1}(S'_u \cap \Omega_h^1) \\ &\qquad \le \liminf_{n\to\infty} \int_{\Omega_{h_n}^+} W(\nabla u_n(x))\,dx + \int_\omega \sqrt{1 + |\nabla h_n(x')|^2}\,dx'\,. \end{aligned} \tag{3}$$

Let us consider a sequence (u_n, h_n) such that

$$\sup_{n\ge 1} F(u_n, h_n) < +\infty\,;$$

we show that, up to a subsequence, $u_n \to u$ in $L^1(\omega \times (0, +\infty))$ and $h_n \to h$ in $L^1(\omega)$, with

$$\overline{F}(u, h) \le \liminf_{n\to\infty} F(u_n, h_n). \tag{4}$$

To prove the lower inequality, it is sufficient to consider sequences (u_n, h_n) with $h_n \in C^\infty(\omega; [0, +\infty))$ and $u_n \in W^{1,p}(\Omega^+_{h_n})$, and $u_n = u^0$ on $\omega \times \{0\}$; however, this compactness property, as well as inequality (4), will still hold if we just assume that $h_n \in W^{1,1}(\omega)$ and $u_n \in SBV_p(\omega \times (-1, +\infty))$ with $u_n = u^0$ in $\omega \times (-1, 0)$, $u_n(x) = 0$ a.e. in $\{x_N > h_n(x')\}$, and $S_u \tilde{\subset} \partial_* \Omega_{h_n}$ (where $A \tilde{\subset} B$ means $\mathcal{H}^{N-1}(A \setminus B) = 0$).

Let us consider first the compactness and l.s.c. of the jump term, and for this we will use a special notion of convergence for a jump set of SBV_p functions.

3.1. Jump set convegence. The following notion of jump set convergence is introduced by Dal Maso, Francfort, and Toader [18, Def. 4.1] and [17, Def. 3.1]. It is called "σ^p-convergence." A variant, which is independent on the exponent $p > 1$, has been introduced more recently by Giacomini and Ponsiglione; see [21].

In what follows, we denote equality and inclusion up to a $\mathcal{H}^{N-1}$-negligible set by the symbols $\tilde{=}$ and $\tilde{\subset}$, respectively.

Definition 3.2. *Let Ω be an open set in $\mathbb{R}^N$, and let $p \in (1, +\infty)$. We say that a sequence $(\Gamma_n)_{n\in\mathbb{N}}$ of subsets of Ω σ^p-converges to Γ if and only if $\sup_{n\in\mathbb{N}} \mathcal{H}^{N-1}(\Gamma_n) < +\infty$ and*

(i) *for any sequence $(v_n)_n$ of functions in $SBV_p(\Omega)$, with $S_{v_n} \tilde{\subset} \Gamma_n$, if the subsequence v_{n_k} goes to v weakly in $SBV_p(\Omega)$ as $k \to \infty$ then $S_v \tilde{\subset} \Gamma$;*

(ii) *there exists a function $v \in SBV_p(\Omega)$ and sequence $(v_n)_n$ of functions in $SBV_p(\Omega)$ converging to v such that $S_{v_n} \tilde{\subset} \Gamma_n$ for each n and $S_v = \Gamma$.*

The following compactness theorem is proven in [18, Thm. 4.7]

Theorem 3.3. *Every sequence $\Gamma_n \subset \Omega$, with $\mathcal{H}^{N-1}(\Gamma_n)$ uniformly bounded, has a σ^p-convergent subsequence.*

The proof of this theorem is based on the following lemma (cf. [18, Lem. 4.5]).

Lemma 3.4. *Let $(v_i)_{i=1}^\infty$ be a sequence in $SBV_p(\Omega) \cap L^\infty(\Omega)$, and let us assume $\mathcal{H}^{N-1}(\bigcup_{i=1}^\infty S_{v_i}) < +\infty$. Then there exist real numbers $c_i > 0$ with $\sum_{i=1}^\infty c_i < +\infty$ such that $v := \sum_{i=1}^\infty c_i v_i \in SBV_p(\Omega) \cap L^\infty(\Omega)$ and $S_v \tilde{=} \bigcup_{i=1}^\infty S_{v_i}$.*

Let us mention the following variant of the proof of Theorem 3.3, still based on Lemma 3.4: Given $\Gamma \subset \Omega$, we introduce

$$X(\Gamma) \;=\; \left\{ v \in SBV_p(\Omega; [-1, 1]) \,:\, S_v \tilde{\subset} \Gamma\,,\; \int_\Omega |\nabla v|^p \, dx \le 1 \right\}.$$

Then, if $\mathcal{H}^{N-1}(\Gamma) < +\infty$, by Ambrosio's compactness theorem, Theorem 2.1, $X(\Gamma)$ is compact in $L^1_{\mathrm{loc}}(\Omega)$ (which is metrizable). If $(\Gamma_n)_n$ is a sequence of jump sets with $L = \sup_n \mathcal{H}^{N-1}(\Gamma_n) < +\infty$, then the sets $X(\Gamma_n)$ all belong to

$$X_L \;=\; \left\{ v \in SBV_p(\Omega; [-1, 1]) \,:\, \mathcal{H}^{N-1}(S_v) \le L,\; \int_\Omega |\nabla v|^p \, dx \le 1 \right\}$$

which is also compact in $L^1_{\mathrm{loc}}(\Omega)$. Hence, a subsequence $(X(\Gamma_{n_k}))_k$ converges in the Hausdorff sense (with the Hausdorff distance in $L^1_{\mathrm{loc}}(\Omega)$ induced by a distance in $L^1_{\mathrm{loc}}(\Omega)$) to a compact $K \subset X_L$. We show that $K \subseteq X(\Gamma)$ for some Γ.

Let $(v_i)_{i=1}^\infty$ be a dense sequence in the compact set K. We first observe that, since K is convex, given any v, v' in K there exists w (given by $\theta v + (1-\theta)v'$ for an appropriate choice of θ, see, for instance, [20]) such that $S_w \tilde{=} S_v \cup S_{v'}$; hence, $\mathcal{H}^{N-1}(S_v \cup S_{v'}) \le L$. In particular, we deduce that $\mathcal{H}^{N-1}(\bigcup_{i=1}^k S_{v_i}) \le L$ for any $k \ge 1$ and, passing to the limit, that $\mathcal{H}^{N-1}(\Gamma) \le L < +\infty$, where we have let $\Gamma = \bigcup_{i=1}^\infty S_{v_i}$. Using Lemma 3.4, we deduce that there exists $v \in K$ with $\Gamma \tilde{=} S_v$. Hence Γ satisfies axiom (ii) in Definition 3.2. On the other hand, any $v \in K$ is the limit of an appropriate subsequence $v_{i(k)}$, $k \ge 1$, with $S_{v_{i(k)}} \tilde{\subset} \Gamma$, and a consequence of Ambrosio's compactness theorem is that $S_v \tilde{\subset} \Gamma$, so that axiom (i) in Definition 3.2 is also satisfied. Hence Γ_{n_k} σ^p-converges to Γ. ∎

We observe that an obvious consequence of Ambrosio's theorem is that, if Γ_n σ^p-converges to Γ,

$$\mathcal{H}^{N-1}(\Gamma) \le \liminf_{n\to\infty} \mathcal{H}^{N-1}(\Gamma_n)\,. \tag{5}$$

3.2. Proof of the lower inequality. Let $\Gamma_n = \partial\Omega_{h_n} = \{x \in \omega \times (-1,+\infty) : x_N = h_n(x')\}$ be the graph of the function h_n. Up to a subsequence, we know by Theorem 3.3 that Γ_n σ^p-converges to some Γ as $n \to \infty$. Since h_n is uniformly bounded in $W^{1,1}(\omega)$, possibly extracting another subsequence, $h_n \to h$ in $L^1(\omega)$. Equivalently, the sets Ω_{h_n} converge to Ω_h in the $L^1(\omega \times (0,+\infty))$ topology for the characteristic functions.

Clearly, $\partial_*\Omega_h \subseteq \Gamma$; indeed, if we take in Definition 3.2 the sequence $v_n = \chi_{\Omega_{h_n}}$, we find that $v_n \to \chi_{\Omega_h}$, whose jump set is $\partial_*\Omega_h$.

Let us decompose Γ in the three parts $\partial_*\Omega_h$, $\Sigma = \Gamma \cap \Omega_h^1$, and $\Sigma^0 = \Gamma \cap \Omega_h^0$. The part Σ^0 is irrelevant in our study since the functions u, limits of converging subsequences of (u_n), will all vanish outside of Ω_h.

We show that Σ is vertical: That is, for any $x = (x', x_N) \in \Sigma$, $(x', x_N + t) \in \Sigma \cup (\mathbb{R}^N \setminus \Omega_h^1)$ for any $t \ge 0$. Indeed, let $v \in SBV_p(\omega \times (-1,+\infty))$ be such that $S_v \tilde{=} \Gamma$, and let v_n be a sequence weakly converging to v in $SBV_p(\omega \times (-1,+\infty))$ with $S_{v_n} \tilde{\subset} \Gamma_n$. Consider the functions $x \mapsto v_n(x', x_N - t)\chi_{\Omega_{h_n}}(x)$, with $t < 1$, extended in an appropriate way in $\omega \times (-1, -1+t)$. These functions will converge to $x \mapsto v(x', x_N - t)\chi_{\Omega_h}(x)$, showing that $(S_v + te_N) \cap \Omega_h^1 \subset \Gamma$, which shows our claim. In particular, we deduce that $\mathcal{H}^{N-1}$-a.e. in Σ, $\nu_\Sigma \cdot e_N = 0$.

By (5), we have $\mathcal{H}^{N-1}(\partial_*\Omega_h) + \mathcal{H}^{N-1}(\Sigma) \le \liminf_{n\to\infty} \mathcal{H}^{N-1}(\Gamma_n)$. We claim that, in addition,

$$\mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(\Sigma) \le \liminf_{n\to\infty} \mathcal{H}^{N-1}(\Gamma_n).$$

This follows from [13] and the definition of σ^p-convergence. Indeed, it is a consequence of the lim inf-inequality in [13], applied to a sequence $(v_n)_{n\ge1}$ with $S_{v_n} \tilde{\subset} \Gamma_n$, weakly converging in $SBV_p(\omega \times (-1,+\infty))$ to a v such that $\Sigma \tilde{\subset} S_v$.

Let us now conclude. If $F(u_n, h_n)$ is uniformly bounded, then by integration along vertical segments we easily check that (u_n) is uniformly bounded in $L^p_{\mathrm{loc}}(\omega \times (-1,+\infty))$. Then, it is a consequence of Ambrosio's theorem, Theorem 2.1, that there exists $u \in GSBV_p(\omega \times (-1,+\infty))$ such that $u_n(x) \to u(x)$ a.e., and $\nabla u_n \rightharpoonup \nabla u$ in $L^p(\omega \times (-1,+\infty); \mathbb{R}^N)$ so that the inequality (1) holds. Clearly, u vanishes out of Ω_h. By point (i) in Definition 3.2, which is easily generalized to $GSBV_p$ functions (see [18, Prop. 4.6]), we have that $S_u \tilde{\subset} \Sigma \cup \partial_*\Omega_h$. In particular, since Σ is vertical, $S'_u \cap \Omega_h^1 \subset \Sigma$. We deduce (2). Clearly, the inequality (4) follows from (1) and (2). ∎

4. An upper bound for the relaxed envelope of *F*. We now get the upper bound for the relaxed envelope of the functional F by proving the following proposition.

Proposition 4.1. *For any u, h, with $\overline{F}(u,h) < +\infty$, there exist (u_n, h_n) with $h_n \in C^1(\omega; [0,+\infty))$, $u_n \in W^{1,p}(\Omega_{h_n})$, and $u_n = u^0$ in $\omega \times (-1, 0)$ such that $h_n \to h$ in $L^1(\omega)$, $u_n \chi_{\Omega^+_{h_n}} \to u\chi_{\Omega^+_h}$ in $L^1(\omega \times (0,+\infty))$,*

$$\limsup_{n\to\infty} \int_{\Omega^+_{h_n}} |\nabla u_n(x)|^p\,dx = \int_{\Omega^+_h} |\nabla u(x)|^p\,dx, \tag{6}$$

and

$$\limsup_{n\to\infty} \int_\omega \sqrt{1+|\nabla h_n(x')|^2}\,dx' \le \mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(S'_u \cap \Omega^1_h). \tag{7}$$

We note that the proposition completes the proof of Theorem 2.2. Indeed, if we find a sequence $(u_n)_n$ satisfying (6), we can deduce the strong convergence $\nabla u_n \chi_{\Omega^+_{h_n}} \to \nabla u \chi_{\Omega^+_h}$ in L^p; the continuity of W (together with (6) and the growth condition of W) gives the general result

$$\begin{aligned} &\limsup_{n\to\infty} \int_{\Omega^+_{h_n}} W(\nabla u_n(x))\,dx + \int_\omega \sqrt{1+|\nabla h_n(x')|^2}\,dx' \\ &\quad\le \int_{\Omega^+_h} W(\nabla u(x))\,dx + \mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(S'_u\cap\Omega^1_h), \end{aligned} \tag{8}$$

which is the $\limsup$ inequality for the functional F.

Remark 4.2. In case one adds in the definition of functional F a volume constraint (that is, $F(u,h) = +\infty$ if $\int_\omega h\,dx \neq V$, where $V > 0$ is a fixed volume), then it is easy to show that Proposition 4.1 still holds, with the sequence (h_n) satisfying the same volume constraint as the limit h. Indeed, given the sequence (h_n) provided by the proposition (without volume constraint), one clearly has $r_n = \int_\omega h_n\,dx / \int_\omega h\,dx \to 1$ as $n \to \infty$, and an appropriate scaling (of the form $x \mapsto (x', x_N/r_n)$) of the functions and the domain will provide new sequences (u_n, h_n) with $\int_\omega h_n\,dx = \int_\omega h\,dx$, still satisfying (6) and (7).

We first state the following lemma, which shows that any BV, nonnegative subgraph with an *essentially closed* boundary can be approximated from below by the subgraph of a smooth, nonnegative function.

Lemma 4.3. *Let $g \in BV(\omega; \mathbb{R}_+)$, and assume $\partial_*\Omega_g$ is essentially closed, that is, $\mathcal{H}^{N-1}(\overline{\partial_*\Omega_g} \setminus \partial_*\Omega_g) = 0$. Then, for any $\varepsilon > 0$, there exists $f \in C^\infty(\omega; \mathbb{R}_+)$ such that $0 \le f \le g$ a.e. in ω, $\|f - g\|_{L^1(\omega)} \le \varepsilon$ and*

$$\left| \int_\omega \sqrt{1+|\nabla f|^2}\,dx \;-\; \mathcal{H}^{N-1}(\partial_*\Omega_g) \right| \;\le\; \varepsilon\,.$$

Proof. Consider first the distance function $d(x) = \mathrm{dist}(x, \partial_*\Omega_g)$ in $\omega \times (-1, +\infty)$. By results on the Minkowski contents [7, 19], we have (because of our assumption of essential closedness)

$$\lim_{\varepsilon\to 0} \frac{|\{x \in \omega\times(-1,+\infty) \,:\, d(x) \le \varepsilon\}|}{2\varepsilon} \;=\; \mathcal{H}^{N-1}(\partial_*\Omega_g)\,.$$

From the BV-coarea formula (and since $|\nabla d| = 1$ a.e.),

$$\frac{|\{x \,:\, d(x)\le\varepsilon\}|}{2\varepsilon} \;=\; \frac{1}{2\varepsilon}\int_{\{d\le\varepsilon\}} |\nabla d(x)|\,dx \;=\; \frac{1}{2\varepsilon}\int_0^\varepsilon \mathcal{H}^{N-1}(\partial\{d\le s\})\,ds\,.$$

We deduce the convergence of the average values,

$$\lim_{\varepsilon\to 0} \fint_0^{\varepsilon} \frac{1}{2}\mathcal{H}^{N-1}(\partial\{d\le s\})\,ds \;=\; \mathcal{H}^{N-1}(\partial_*\Omega_g)\,,$$

so that we can find a sequence $(\varepsilon_k)_{k\ge 1}$ with $\varepsilon_k\downarrow 0$ such that

$$\lim_{k\to\infty} \frac{1}{2}\mathcal{H}^{N-1}(\partial\{d\le \varepsilon_k\}) \;=\; \mathcal{H}^{N-1}(\partial_*\Omega_g) \tag{9}$$

(without loss of generality we also may assume that the boundary $\partial\{d\le\varepsilon_k\}$ is Lipschitz).

Now, the boundary of $\{d\le\varepsilon_k\}$ is the disjoint union of the boundaries of $\{x\in\omega\times(-1,+\infty) : \mathrm{dist}(x,\Omega_g)\le\varepsilon_k\}$ and $\{x\in\Omega_g : d(x)>\varepsilon_k\}$, and both of these sets converge (in $L^1(\omega\times(-1,+\infty))$) to Ω_g so that the lim inf of their perimeter is greater or equal to $\mathcal{H}^{N-1}(\partial_*\Omega_g)$. Together with (9) it shows that these perimeters go to $\mathcal{H}^{N-1}(\partial_*\Omega_g)$, in particular,

$$\lim_{k\to\infty} \mathcal{H}^{N-1}(\partial\{x\in\Omega_g\,:\,d(x)>\varepsilon_k\}) \;=\; \mathcal{H}^{N-1}(\partial_*\Omega_g)\,.$$

This set $\{x\in\Omega_g \,:\, d(x)>\varepsilon_k\}$ is the subgraph in $\omega\times(-1,+\infty)$ of a function $g_k\in BV(\omega;[-\varepsilon_k,+\infty))$, with $g_k\le g-\varepsilon_k$ a.e. in ω. We consider $g_k^+ = g_k\vee 0$: We have $0\le g_k^+$ and

$$\mathcal{H}^{N-1}(\partial\Omega_{g_k^+}) \;\le\; \mathcal{H}^{N-1}(\partial\Omega_{g_k})$$

(since $\partial\Omega_{g_k^+}\cap(\omega\times\{0\})$ is the orthogonal projection onto $\omega\times\{0\}$ of $\partial\Omega_{g_k}\cap(\omega\times(-1,0]))$. Hence, we still have

$$\lim_{k\to\infty}\mathcal{H}^{N-1}(\partial\Omega_{g_k^+}) \;=\; \mathcal{H}^{N-1}(\partial_*\Omega_g)\,.$$

By convolution, we can build from g_k^+ a sequence of smooth functions f_k that are still nonnegative, that go to g in $L^1(\omega)$, and such that

$$\lim_{k\to\infty}\int_\omega \sqrt{1+|\nabla f_k(x')|^2}\,dx' \;=\; \mathcal{H}^{N-1}(\partial_*\Omega_g)\,.$$

Let $x'\in\omega$. By construction of g_k we have $g_k(y')\le g_+(x')$ a.e. in $\{y'\in\omega : |y'-x'|\le\varepsilon_k\}$ (where $g_+=g$ a.e. is the precise representative defined in section 2.1). Since $g\ge 0$ a.e., we also have $g_k^+(y')\le g_+(x')$ a.e. in $\{y'\in\omega : |y'-x'|\le\varepsilon_k\}$. This shows that if for each k the size of the support of the convolution kernel is chosen small enough (for instance, of diameter less than $\varepsilon_k/2$), we also have $f_k\le g_+$. This proves Lemma 4.3. □

Proof of Proposition 4.1. Let us consider, now, u and h such that $\overline{F}(u,h)<+\infty$. We divide this proof into two steps.

Step 1 (approximation of (most of) the graph). We show that we can approximate a "generalized graph" $(\partial_*\Omega_h,\Sigma)$, where $\Sigma\subset\Omega_h^1\cap(\omega\times(0,+\infty))$ is vertical in the sense that $x\in\Sigma\Rightarrow(x',x_N+t)\in\Sigma$ for any $t\ge 0$ as long as $(x',x_N+t)\in\Omega_h^1$, with the graph of a smooth function $f:\omega\to\mathbb{R}_+$, with $\Omega_f\subset\Omega_h\setminus\Sigma$ up to a small part, and with a good approximation of the total surface energy $\mathcal{H}^{N-1}(\partial_*\Omega_h)+2\mathcal{H}^{N-1}(\Sigma)$ (by the surface of the smooth graph $\int_\omega\sqrt{1+|\nabla f|^2}\,dx$).

Let us first assume that $\Sigma = \emptyset$: We claim that for any $h \in BV(\omega;\mathbb{R}_+)$ and $\varepsilon > 0$, there exists $f \in C^\infty(\omega;\mathbb{R}_+)$ such that

$$\|f-h\|_{L^1(\omega)} \;+\; \mathcal{H}^{N-1}(\partial_*\Omega_h \cap \Omega_f) \;\le\; \varepsilon \tag{10}$$

and

$$\left| \int_\omega \sqrt{1+|\nabla f(x)|^2}\,dx \;-\; \mathcal{H}^{N-1}(\partial_*\Omega_h) \right| \;\le\; \varepsilon\,. \tag{11}$$

We fix $\varepsilon > 0$. Let us consider a mollifying kernel $\rho \in C_c^\infty(\mathbb{R}^N)$, with support in the unit ball, and for any $\eta > 0$ let $\rho_\eta(x) = (1/\eta)^N \rho(x/\eta)$. For $n \ge 1$ we consider the function $w_n = \rho_{1/n} * \chi_{\Omega_h} : \omega \times \mathbb{R} \to [0,1]$. It is well known not only that $w_n \to \chi_{\Omega_h}$ strongly in L^1 but also that $\int_{\omega\times(-1,+\infty)} |\nabla w_n(x)|\,dx \;\to\; |D\chi_{\Omega_h}|(\omega\times(-1,+\infty)) \;=\; \mathcal{H}^{N-1}(\partial_*\Omega_h)$ as $n \to +\infty$.

One has, for every $x \in \Omega_h^1 \cup \partial^*\Omega_h \cup \Omega_h^0$ (hence, $\mathcal{H}^{N-1}$-a.e. $x \in \omega\times(-1,+\infty)$),

$$\lim_{n\to\infty} w_n(x) \;=\; \begin{cases} 1 & \text{if } x \in \Omega_h^1\,,\\ \frac{1}{2} & \text{if } x \in \partial^*\Omega_h\,,\\ 0 & \text{if } x \in \Omega_h^0\,. \end{cases} \tag{12}$$

The same properties are true for the sequence of (l.s.c.) functions $(\tilde{w}_n)_{n\ge1}$ defined by

$$\tilde{w}_n(x) \;=\; \begin{cases} w_n(x) & \text{if } x \in \omega\times[0,+\infty)\,,\\ 1 & \text{if } x \in \omega\times(-1,0)\,. \end{cases}$$

Indeed, using the coarea formula, one sees that

$$\begin{aligned} |D\tilde{w}_n|(\omega\times(-1,+\infty)) \;&=\; \int_0^1 \mathcal{H}^{N-1}(\partial\{\tilde{w}_n > s\})\,ds \\ &\le\; \int_0^1 \mathcal{H}^{N-1}(\partial\{w_n > s\})\,ds \;=\; \int_{\omega\times(0,+\infty)} |\nabla w_n(x)|\,dx \end{aligned}$$

since $\mathcal{H}^{N-1}(\partial\{w_n > s\} \cap (\omega\times(-1,0))) \ge \mathcal{H}^{N-1}(\{x' \in \omega : w_n(x',0) \le s\}) = \mathcal{H}^{N-1}(\partial\{\tilde{w}_n > s\} \cap (\omega\times(-1,0)))$, the second set being the projection onto $\omega\times\{0\}$ of the first one. We deduce that $\limsup_{n\to\infty} |D\tilde{w}_n|(\omega\times(-1,+\infty)) \le \mathcal{H}^{N-1}(\partial_*\Omega_h)$, but since $\tilde{w}_n \to \chi_{\Omega_h}$, it yields $\lim_{n\to\infty} |D\tilde{w}_n|(\omega\times(-1,+\infty)) = \mathcal{H}^{N-1}(\partial_*\Omega_h)$. Clearly, (12) is also true for $\tilde{w}$ since $\Omega_h^1 \supset \omega\times(-1,0)$. We drop the tilde in the sequel and just write w_n instead of $\tilde{w}_n$.

For a.e. $s \in (0,1)$, one also checks that $\lim_{n\to\infty} |\{w_n > s\}\triangle\Omega_h| = 0$, and using Fatou's lemma and the coarea formula, that for a.e. $s \in (0,1)$, $\{w_n > s\}$ is an open set such that $\liminf_{n\to\infty} \mathcal{H}^{N-1}(\partial\{w_n > s\}) = \mathcal{H}^{N-1}(\partial_*\Omega_h)$. Thus, up to a subsequence (possibly depending on s), we may assume $\lim_{n\to\infty} \mathcal{H}^{N-1}(\partial\{w_n > s\}) = \mathcal{H}^{N-1}(\partial_*\Omega_h)$. Let us consider $s^* \in (2/3, 3/4)$ and an appropriate subsequence such that this property is true, and let us consider the corresponding sequence of sets $\{x \in \omega\times(-1,+\infty) : w_n(x) > s^*\}$. We have that $\mathcal{H}^{N-1}(\partial_*\Omega_h \cap \{w_n > s^*\}) = \int_{\partial_*\Omega_h} \chi_{\{w_n > s^*\}}(x)\,d\mathcal{H}^{N-1}(x)$, and since by (12), $\chi_{\{w_n > s^*\}}(x) \to 0$ $\mathcal{H}^{N-1}$-a.e. in $\partial_*\Omega_h$, we find $\mathcal{H}^{N-1}(\partial_*\Omega_h \cap \{w_n > s^*\}) \to 0$ as $n \to \infty$. We fix n large such that

$$\begin{gathered} |\{w_n > s^*\}\triangle\Omega_h| \;+\; \mathcal{H}^{N-1}(\partial_*\Omega_h \cap \{w_n > s^*\}) \;\le\; \frac{\varepsilon}{2}\,, \\ \left|\mathcal{H}^{N-1}(\partial\{w_n > s^*\}) \;-\; \mathcal{H}^{N-1}(\partial_*\Omega_h)\right| \;\le\; \frac{\varepsilon}{2}\,. \end{gathered}$$

It is clear that there exists $g : \omega \to [0, +\infty)$ a BV function such that $\{w_n > s^*\} = \{x_N < g(x')\}$. By Lemma 4.3 applied to g, we find a smooth function $f \le g$, $f \ge 0$, satisfying both (10) and (11).

Now, assume $\Sigma \neq \emptyset$. First, possibly replacing h by $h \wedge (M-1) = \min\{h, M-1\}$, $M > 1$ large, we may assume without loss of generality that h is bounded by $M-1$. Let us then define Σ' by $\Sigma' = \bigcup_{x\in\Sigma}\{x'\}\times[x_N, M]$ and recall that by assumption $\Sigma' \cap \Omega_h^1 = \Sigma$. We may also assume without loss of generality that $\mathcal{H}^{N-1}(\Sigma' \cap (\omega \times [0, M])) < +\infty$, possibly replacing (in a preliminary step) h with $h_\delta = (h-\delta)^+$, $\delta > 0$ small, and Σ with $\Sigma_\delta = \Sigma \cap \Omega_{h_\delta}$: Indeed, one will have that $\Sigma'_\delta \cap \{h_\delta(x') \le x_N \le h_\delta(x') + \delta\} \subseteq \Sigma$ so that $\mathcal{H}^{N-1}(\Sigma'_\delta \cap (\omega \times [0, M])) \le (M/\delta)\mathcal{H}^{N-1}(\Sigma) < +\infty$. Now, let $K \subseteq \Sigma'$ be a compact set such that $\mathcal{H}^{N-1}(\Sigma' \setminus K) \le \varepsilon/10$. Observe that, if K' is defined as Σ', also $\mathcal{H}^{N-1}(\Sigma' \setminus K') \le \varepsilon/10$, and K' is compact.

Let us build the sequence of l.s.c. functions $(w_n)_{n\ge1}$ and find a level $s^* \in (2/3, 3/4)$, as previously. By (12), we have that $\chi_{\{w_n > s^*\}}$ converges to 1 in Ω_h^1, while it tends to 0 $\mathcal{H}^{N-1}$-a.e. outside. In particular, $\mathcal{H}^{N-1}(K' \cap \{w_n > s^*\}) \to \mathcal{H}^{N-1}(K' \cap \Omega_h^1)$ as $n \to \infty$, and this limit satifies $\mathcal{H}^{N-1}(\Sigma) - \varepsilon/10 \le \mathcal{H}^{N-1}(K' \cap \Omega_h^1) \le \mathcal{H}^{N-1}(\Sigma)$. We can hence choose n such that

$$|\{w_n > s^*\} \triangle \Omega_h| \ + \ \mathcal{H}^{N-1}(\partial_* \Omega_h \cap \{w_n \ge s^*\}) \ \le \ \frac{\varepsilon}{4}\,,$$

$$\left|\mathcal{H}^{N-1}(\partial\{w_n > s^*\}) \ - \ \mathcal{H}^{N-1}(\partial_*\Omega_h)\right| \ \le \ \frac{\varepsilon}{4}\,,$$

and

$$\left|\mathcal{H}^{N-1}(K' \cap \{w_n > s^*\}) \ - \ \mathcal{H}^{N-1}(\Sigma)\right| \ \le \ \frac{\varepsilon}{8}\,.$$

Observe now that since the set K' is compact, its Minkowski content $|\{\mathrm{dist}(\cdot, K') < s\}|/(2s)$ converges to $\mathcal{H}^{N-1}(K')$ as $s \to 0$ (see [7, 19]). As in the proof of Lemma 4.3, we deduce that there exists a sequence $(s_k)_{k\ge1}$ such that

$$\lim_{k\to\infty} \mathcal{H}^{N-1}(\partial\{\mathrm{dist}(\cdot, K') > s_k\}) \ = \ 2\mathcal{H}^{N-1}(K')\,. \tag{13}$$

We introduce the measures $\mu_k = \mathcal{H}^{N-1} \llcorner \partial\{\mathrm{dist}(\cdot, K') > s_k\}$. Up to a subsequence, we may assume that they converge (weakly-$*$) to a measure μ supported on K'. A consequence of the lim inf inequality in [13] is that $\mu(A) \ge 2\mathcal{H}^{N-1}(K' \cap A)$ for any open set $A \subset \omega \times (0, +\infty)$. (In this simple case, it can be shown directly by a slicing argument; see, for instance, [12, Lem. 2]). Together with (13), it shows that $\mu = 2\mathcal{H}^{N-1} \llcorner K'$. In particular, if k is large enough and provided we have chosen s^* such that $\mathcal{H}^{N-1}(K' \cap \partial\{w_n > s^*\}) = 0$ (almost any choice suits, since $\mathcal{H}^{N-1}(K' \cap (\omega \times \{0\})) = 0$—otherwise $\mathcal{H}^{N-1}(\Sigma')$ would be infinite), we have

$$\left|\mathcal{H}^{N-1}(\partial\{\mathrm{dist}(\cdot, K') > s_k\} \cap \{w_n > s^*\}) \ - \ 2\mathcal{H}^{N-1}(\Sigma)\right| \ \le \ \frac{\varepsilon}{2}\,,$$

while $|\{\mathrm{dist}(\cdot, K') \le s_k\}| \le \varepsilon/4$ and $\mathcal{H}^{N-1}(\partial\{w_n > s^*\} \cap \{\mathrm{dist}(\cdot, K') \le s_k\}) \le \varepsilon/8$.

For such values of k, the open set $\{\mathrm{dist}(\cdot, K') > s_k\} \cap \{w_n > s^*\} \cap (\omega \times (-1, +\infty))$ (with piecewise Lipschitz boundary if s_k was properly chosen) is the subgraph Ω_g of a nonnegative BV function g with $\|g - h\|_{L^1(\omega)} \le \varepsilon/2$, $\mathcal{H}^{N-1}(\partial\Omega_g \setminus \partial_*\Omega_g) = 0$,

$$\mathcal{H}^{N-1}((\partial_*\Omega_h \cup \Sigma) \cap \Omega_g) \ \le \ \frac{\varepsilon}{2},$$

and $\partial\Omega_g = (\partial\{\mathrm{dist}(\cdot, K') > s_k\} \cap \{w_n > s^*\}) \cup (\partial\{w_n > s^*\} \cap \{\mathrm{dist}(\cdot, K') > s_k\})$, so that

$$\left|\mathcal{H}^{N-1}(\partial\Omega_g) - (\mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(\Sigma))\right| \le \frac{3\varepsilon}{4}.$$

Then, invoking again Lemma 4.3, we find a smooth function $f \le g$, $f \ge 0$, with $\|f - h\|_{L^1(\omega)} \le \varepsilon$,

$$\mathcal{H}^{N-1}((\partial_*\Omega_h \cup \Sigma) \cap \Omega_f) \le \varepsilon \tag{14}$$

and

$$\left|\int_\omega \sqrt{1 + |\nabla f(x)|^2}\, dx - (\mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(\Sigma))\right| \le \varepsilon\,. \tag{15}$$

Remark 4.4. We have, in addition,

$$\lim_{\varepsilon\to 0} \mathcal{H}^{N-1}(\{x' \in \omega : f_\varepsilon(x') = 0\}) = \mathcal{H}^{N-1}(\{x' \in \omega : h(x') = 0\})$$

(f_ε denoting the f obtained for a particular $\varepsilon > 0$). Indeed, for $\eta > 0$, there exist $k > 1$ such that $\mathcal{H}^{N-1}(\{h < 1/k\}) \le \mathcal{H}^{N-1}(\{h = 0\}) + \eta$ and $K \subset \omega$ with $\mathcal{H}^{N-1}(K) \le \eta$ such that $f_\varepsilon \to h$ uniformly in $\omega \setminus K$. Then, if ε is small enough, $h(x') \ge 1/k$ and $x' \notin K$ will yield $f_\varepsilon(x') \ge 1/(2k)$; hence, $\{f_\varepsilon = 0\} \subset K \cup \{h < 1/k\}$ so that $\mathcal{H}^{N-1}(\{f_\varepsilon = 0\}) \le \mathcal{H}^{N-1}(\{h = 0\}) + 2\eta$. We deduce that $\limsup_{\varepsilon\to 0} \mathcal{H}^{N-1}(\{f_\varepsilon = 0\}) \le \mathcal{H}^{N-1}(\{h = 0\})$. On the other hand, since $\mathcal{H}^{N-1}(\partial_*\Omega_h \cap \Omega_{f_\varepsilon}) \to 0$, we see that $\mathcal{H}^{N-1}(\{h = 0\} \cap \{f_\varepsilon > 0\}) \to 0$ so that $\mathcal{H}^{N-1}(\{h = 0\} \cap \{f_\varepsilon = 0\}) \to \mathcal{H}^{N-1}(\{h = 0\})$; hence, $\mathcal{H}^{N-1}(\{h = 0\}) \le \liminf_{\varepsilon\to 0} \mathcal{H}^{N-1}(\{f_\varepsilon = 0\})$.

A consequence is that in case (as in [9]) the "substrate" $\{x_N \le 0\}$ has a superficial tension σ_S less than the superficial tension σ_C of the crystal, that is, the surface energy of $(\partial_*\Omega_h, \Sigma)$ is

$$\sigma_S \mathcal{H}^{N-1}(\{h = 0\}) + \sigma_C(\mathcal{H}^{N-1}(\partial_*\Omega_h \cap (\omega \times (0, +\infty))) + 2\mathcal{H}^{N-1}(\Sigma))\,,$$

then f can fulfill the additional requirement

$$\begin{aligned}\Big|\sigma_S \mathcal{H}^{N-1}(\{f = 0\}) + \sigma_C \int_{\{f>0\}} \sqrt{1 + |\nabla f|^2}\, dx \\ - \left(\sigma_S \mathcal{H}^{N-1}(\{h = 0\}) + \sigma_C(\mathcal{H}^{N-1}(\partial_*\Omega_h \cap (\omega \times (0, +\infty))) + 2\mathcal{H}^{N-1}(\Sigma))\right)\Big| \le \varepsilon.\end{aligned}$$

If on the other hand $\sigma_C < \sigma_S$, this is not optimal (in terms of relaxation, approximating (h, Σ) with $(h + \delta, \Sigma + \delta e_N)$, δ small, will reduce the energy).

Step 2. (approximation of both the graph and the displacement). We now show that if $u \in GSBV_p(\omega \times (-1, +\infty))$ is given, with $S_u \subseteq \partial_*\Omega_h \cup \Sigma$, $u = 0$ out of Ω_h, and $u = u^0$ on $\omega \times (-1, 0)$ (where $u^0 \in W^{1,p}(\omega \times (-1, 0))$, $\Sigma \subset \Omega_h^1 \cap (\omega \times (0, +\infty))$ vertical), then there exists $(u_n, h_n)_{n\ge 1}$, with $h_n \in C^\infty(\omega; \mathbb{R}_+)$, $u_n \in W^{1,p}(\Omega_{h_n})$, and $u_n = u^0$ in $\omega \times (-1, 0)$, such that as $n \to \infty$, $h_n \to h$ in $L^1(\omega)$, and (extending both u_n and ∇u_n with zero out of Ω_{h_n}) $u_n \to u$ in $L^1(\omega \times (-1, +\infty))$, $\nabla u_n \to \nabla u$ strongly in $L^p(\omega \times (-1, +\infty); \mathbb{R}^N)$,

$$\lim_{n\to\infty} \int_\omega \sqrt{1 + |\nabla h_n(x)|^2}\, dx = \mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(\Sigma)\,.$$

Before entering the proof of this second step, which is very technical, let us give a rough idea of how it goes. It follows a discretization/reinterpolation argument introduced in [15, 16]. We first discretize the function u in $\omega \times (-1, +\infty)$ on a regular square grid (of the form $\eta\mathbb{Z}^N$ with η small). Then, we reinterpolate this discretization into a piecewise continuous function u^η. All this is done in a way that ensures that some suitable volume energy of u^η is controlled by the same energy of u and converges in the limit $\eta \to 0$. On the other hand, the jump set of this approximation u^η is "close" in some sense to the jump set of u, but a drawback of this technique is that we can control only its total surface by $C \times (\mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(\Sigma))$ (with C a large constant depending on the dimension N).

To overcome this difficulty, we have to use the approximation f provided by the previous step: Instead of doing the construction in the whole set $\omega \times (-1, +\infty)$, we work in the smooth set Ω_f, where f is such that the jump of u in Ω_f has a surface of order ε (and satisfies (15)). In this way, the jump set of u^η in Ω_f is now controlled by $C\varepsilon$, and after extending u^η to $\omega \times (-1, +\infty)$ by zero above the graph of f, we get a couple (u^η, h^η) with total surface energy controlled by $\mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(\Sigma) + (C+1)\varepsilon$, as required.

Let us now enter the details. We fix $\varepsilon > 0$. By Step 1, there exists $f \in C^\infty(\omega)$ with $\|f - h\|_{L^1(\omega)} \le \varepsilon$, such that both (14) and (15) hold. (In particular, (14) states that f is "almost" below h (and Σ) in the sense that very little of $\partial_*\Omega_h \cup \Sigma$ lies below f.) We denote by v the $GSBV_p$ function that is equal to u in Ω_f, to 0 in $(\omega \times (0, +\infty)) \setminus \Omega_f$, and to u^0 in $\omega \times (-1, 0)$. Possibly choosing f closer to h, we may assume, also, that $\|v - u\|_{L^1(\omega \times (-1,+\infty))} \le \varepsilon$. Eventually, we also extend v (by symmetry) slightly below $\omega \times \{-1\}$ to the set $\omega \times (-1-\delta, -1)$, $0 < \delta < 1$.

Let us define, for $\xi \in \mathbb{R}^N$, the anisotropic potential

$$W_p(\xi) \ := \ \sum_{i=1}^N |\xi_i|^p \,.$$

Clearly, $v \in GSBV_p(\omega \times (-1-\delta, +\infty))$, and one has, if δ is small enough,

$$\begin{aligned} \int_{\Omega_f^\delta} W_p(\nabla v(x))\,dx \ &= \ \int_{\omega\times(-1-\delta,+\infty)} W_p(\nabla v(x))\,dx \\ &\le \ \int_{\omega\times(-1,+\infty)} W_p(\nabla u(x))\,dx + \varepsilon, \end{aligned} \tag{16}$$

where $\Omega_f^\delta = \{x \in \omega \times (-1-\delta, +\infty) \,:\, x_N < f(x')\}$. The jump set of v satisfies $S_v \subset \partial\Omega_f \cup ((\partial_*\Omega_h \cup \Sigma) \cap \Omega_f)$; its surface energy is estimated by (14) and (15).

For $n \ge 1$, let $\eta = 1/n$ be a discretization step. Given $y \in (0,1)^N$, we introduce a discretization of v by setting $v_k^{y,\eta} = v(y\eta + k\eta)$, with $(k_1, \dots, k_{N-1}) \in (\mathbb{Z}/n\mathbb{Z})^{N-1}$ and $k_N \in \mathbb{Z} \cap [-(1+\delta)/\eta - y_N, +\infty)$ (so that only points in $\omega \times (-1-\delta, +\infty)$ are considered).

We also define a "discrete jump" of $v^{y,\eta}$: We let, for $i = 1, \dots, N$, and y, k as above, $l_k^{i,y,\eta} = 0$ if $(\partial_*\Omega_h \cup \Sigma) \cap [y\eta + k\eta, y\eta + (k+e_i)\eta] = \emptyset$ and $l_k^{i,y,\eta} = 1$ otherwise. We have that $l^{i,y,\eta} = \chi_{S_\eta^i}(y\eta + k\eta)$, where the set S_η^i is given by

$$S_\eta^i \ = \ (\partial_*\Omega_h \cup \Sigma) \, + \, [-\eta e_i, 0] \,.$$

Here $(e_1, \dots, e_N)$ is the canonical basis of $\mathbb{R}^N$, and as usual the sum of two sets A, B is $A + B = \{a + b \,:\, a \in A, b \in B\}$.

The discrete energy of $(v_k^{y,\eta},(l_k^{i,y,\eta})_{i=1}^N)_k$ is defined by

$$D_\eta^y \;=\; \sum_{i=1}^N D_\eta^{i,y} \;\text{ with }\; D_\eta^{i,y}=\eta^N\sum_k (1-l_k^{i,y,\eta})\frac{|v_{k+e_i}^{y,\eta}-v_k^{y,\eta}|^p}{\eta^p}+\alpha\frac{l_k^{i,y,\eta}}{\eta}\,,$$

where the sum is taken on all k such that the segment $[y\eta+k\eta, y\eta+(k+e_i)\eta]$ lies inside the subgraph Ω_f^δ. The parameter $\alpha>0$ will be fixed later on.

Let us compute the average $\int_{y\in(0,1)^N} D_\eta^y\,dy$. For each i, one has (using the change of variable $(y,k)\mapsto x=(y+k)\eta$)

$$\text{(17)}\qquad \int_{(0,1)^N} D_\eta^{i,y}\,dy \;=\; \int_{\mathcal{O}_\eta^i}(1-\chi_{S_\eta^i})(x)\frac{|v(x+\eta e_i)-v(x)|^p}{\eta^p}\;+\;\alpha\frac{\chi_{S_\eta^i}(x)}{\eta}\,dx\,,$$

where the domain of integration is

$$\mathcal{O}_\eta^i=\left\{x\in\omega\times(-1-\delta,+\infty)\;:\;x_N<\min_{0\le t\le 1} f(x'+t\eta e_i)\right\}\quad\text{if } i\le N-1,\ \text{and}$$
$$\mathcal{O}_\eta^N=\{x\in\omega\times(-1-\delta,+\infty)\;:\;x_N<f(x')-\eta\}\;.$$

We now use a slicing technique introduced by Gobbino [22] (based on the slicing properties of $GSBV$ functions [7] and applied in a similar setting in [2, 15, 16]). The second integral in (17) is decomposed into an integral on $e_i^\perp$ and an integral along the direction e_i, as follows:

$$\int_{(0,1)^N} D_\eta^{i,y}\,dy \;=\; \int_{e_i^\perp} d\mathcal{H}^{N-1}(z)\int_{\{s:z+se_i\in\mathcal{O}_\eta^i\}}(1-\chi_{S_\eta^i})(z+se_i)\frac{|v(z+(s+\eta)e_i)-v(z+se_i)|^p}{\eta^p}\;+\;\alpha\frac{\chi_{S_\eta^i}(z+se_i)}{\eta}\,ds\,.$$

For $\mathcal{H}^{N-1}$-a.e. $z\in e_i^\perp$, from the definition of S_η^i there is no jump of v between $z+se_i$ and $z+(s+\eta)e_i$ (for a.e. s) when $(1-\chi_{S_\eta^i})(z+se_i)\neq 0$, so that in this case

$$|v(z+(s+\eta)e_i)-v(z+se_i)|^p \;=\; \left|\int_0^\eta \frac{\partial v}{\partial x_i}(z+(s+t)e_i)\,dt\right|^p \;\le\; \eta^{p-1}\int_0^\eta\left|\frac{\partial v}{\partial x_i}(z+(s+t)e_i)\right|^p\,dt\,.$$

We deduce

$$\text{(18)}\qquad \int_{e_i^\perp} d\mathcal{H}^{N-1}(z)\int_{\{s:z+se_i\in\mathcal{O}_\eta^i\}}(1-\chi_{S_\eta^i})(z+se_i)\frac{|v(z+(s+\eta)e_i)-v(z+se_i)|^p}{\eta^p} \;\le\; \int_{\Omega_f^\delta}\left|\frac{\partial v}{\partial x_i}(x)\right|^p\,dx\,.$$

On the other hand, for $\mathcal{H}^{N-1}$-a.e. z, we have (from the definition of S_η^i)

$$\left|\{s:z+se_i\in\mathcal{O}_\eta^i\cap S_\eta^i\}\right| \;\le\; \eta\mathcal{H}^0\left(\{s:z+se_i\in\mathcal{O}_\eta^i\cap(\partial_*\Omega_h\cup\Sigma)\}\right)$$

so that

$$
\int_{e_i^\perp} d\mathcal{H}^{N-1}(z) \int_{\{s: z+se_i \in \mathcal{O}_\eta^i\}} \frac{\chi_{S_\eta^i}(z+se_i)}{\eta}\, ds \le \int_{(\partial_*\Omega_h \cup \Sigma)\cap\Omega_f} |e_i\cdot\nu(x)|\, d\mathcal{H}^{N-1}(x)\,, \tag{19}
$$

where ν is the normal to $\partial_*\Omega_h \cup \Sigma$, defined $\mathcal{H}^{N-1}$-a.e. (up to a change of sign which is not relevant here). Collecting (18) and (19), we get

$$
\int_{(0,1)^N} D_\eta^{i,y}\, dy \;\le\; \int_{\Omega_f^\delta} \left|\frac{\partial v}{\partial x_i}(x)\right|^p dx \;+\; \alpha \int_{(\partial_*\Omega_h\cup\Sigma)\cap\Omega_f} |e_i\cdot\nu(x)|\, d\mathcal{H}^{N-1}(x)\,.
$$

By construction, we have $\mathcal{H}^{N-1}((\partial_*\Omega_h \cup \Sigma)\cap\Omega_f) \le \varepsilon$ (see (14)); hence,

$$
D_\eta^y \;=\; \sum_{i=1}^N \int_{(0,1)^N} D_\eta^{i,y} \;\le\; \int_{\Omega_f^\delta} W_p(\nabla v(x))\, dx \;+\; \alpha\sqrt{N}\varepsilon\,. \tag{20}
$$

Now, for any y and $\eta > 0$ (small), we introduce the interpolate (known as "Q1" in finite elements theory) of $(v_k^{y,\eta})_k$:

$$
v^{y,\eta}(x) \;=\; \sum_{k\in(\mathbb{Z}/n\mathbb{Z})^{N-1}\times\mathbb{Z}} v_k^{y,\eta}\Delta\left(\frac{x}{\eta} - (k+y)\right)\,, \quad x \in \omega\times\mathbb{R}\,,
$$

where (as before $a^+ = a \vee 0 = \max\{a, 0\}$)

$$
\Delta(x) \;=\; \prod_{i=1}^N (1-|x_i|)^+\,. \tag{21}
$$

It is classical [2, 14] that there exists a sequence $(\eta_l)_{l\ge1}$ such that $v^{y,\eta} \to v$ in $L^1(\omega\times(-1,+\infty))$ as $l\to\infty$ for a.e. $y\in(0,1)^N$. Then, possibly extracting a subsequence, we deduce from (20) that there exists $y\in(0,1)^N$ such that both

$$
\lim_{l\to\infty} D_{\eta_l}^y \;\le\; \int_{\omega\times(-1-\delta,+\infty)} W_p(\nabla v)\, dx \;+\; \alpha\sqrt{N}\varepsilon \tag{22}
$$

and $\|v^{y,\eta_l} - v\|_{L^1} \to 0$ as $l\to\infty$. In what follows, we fix y to this value and drop the corresponding superscript.

Consider now a cube $C_k = (y+k)\eta_l + (0,\eta_l)^N$ such that $C_k \subset \Omega_f^\delta$. We say that C_k is a "regular cube" if $\partial_*\Omega_h \cup \Sigma$ does not cross any edge of C_k, that is, when $l_{\hat{k}}^{i,\eta_l} = 0$ for any i and $\hat{k} \in k + \{0,1\}^N$ with $\hat{k}_i = k_i$. On the other hand, if $\partial_*\Omega_h\cup\Sigma$ crosses at least one of the edges of C_k, we say the cube is a "jump cube." Notice that in the latter case, since $\partial_*\Omega_h\cup\Sigma$ is a generalized subgraph, all cubes "above" C_k are also jump cubes as long as they intersect Ω_h: Precisely, every other cube $C' = C_{k',k_N+m}$, $m\ge1$, with $C'\subset\Omega_f^\delta$ has at least one edge that crosses $\partial_*\Omega_h\cup\Sigma$ unless $C'\subset\Omega_f\setminus\Omega_h^1$ (in which case $v=0$ and $v^{\eta_l}=0$ a.e. in C'). We denote by $\mathcal{J}$ the union of all jump cubes and of all cubes in Ω_f^δ that lie above a jump cube (i.e., either jump cubes or regular cubes where v vanishes) and by $\mathcal{R}$ the union of all the other regular cubes (the regular cubes that lie below $\partial_*\Omega_h\cup\Sigma$) so that $\mathcal{C}_f = \mathcal{R}\cup\mathcal{J}$ is the union of all cubes $C_k\subset\Omega_f^\delta$ (see Figure 2).

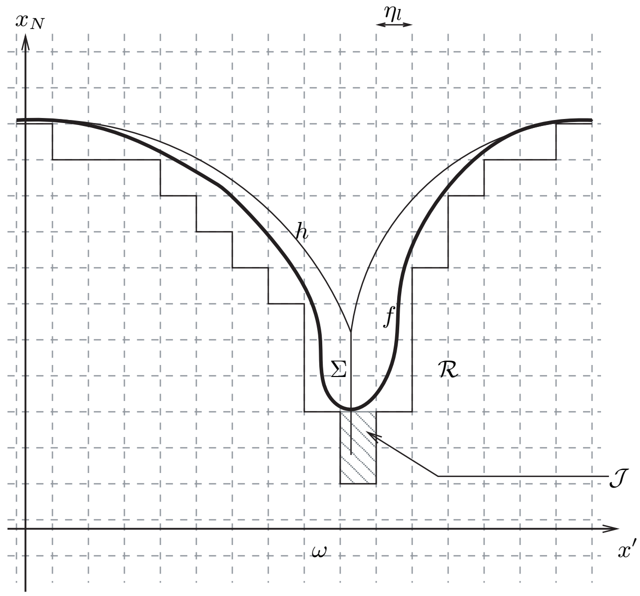


FIG. 2. *The cubes below f are grouped into two regions: a region $\mathcal{R}$ where the total bulk energy of v^{η_l} is estimated by the "bulk" part of D_{η_l} and a region $\mathcal{J}$ whose common boundary with $\mathcal{R}$ is estimated by the "surface term," or order ε, of D_{η_l}.*

Since each edge $[(y+k)\eta_l,(y+k+e_i)\eta_l]$ (for any i) is shared by at most 2^{N-1} cubes, one can decompose the energy D_{η_l} as a sum of contributions of the cubes in the following way:

$$D_{\eta_l} \;\geq\; \sum_{C_k \text{ "regular"}} (\eta_l)^N \frac{1}{2^{N-1}} \sum_{i=1}^{N} \sum_{\substack{\hat{k}\in k+\{0,1\}^N \\ \hat{k}_i=k_i}} \frac{\left| v^{\eta_l}_{\hat{k}+e_i} - v^{\eta_l}_{\hat{k}} \right|^p}{(\eta_l)^p}$$
$$+\; \alpha\left(\frac{\eta_l}{2}\right)^{N-1} \times (\text{number of "jump cubes"})$$

(this is a very rough estimate since only one edge of each jump cube is taken into account even if many edges cross $\partial_*\Omega_h \cup \Sigma$). In particular, from (20) we find that the number of the jump cubes is bounded by a constant times η_l^{1-N}, so that their total Lebesgue measure is $O(\eta_l)$.

By inequality (36) in Lemma A.1, the term in the sum over regular cubes is larger or equal to $\int_{C_k} W_p(\nabla v^{\eta_l}(x))\,dx$; on the other hand, the term involving the jump cubes bounds the measure of the boundary of these cubes since clearly $\alpha(\eta_l/2)^{N-1} = \alpha\mathcal{H}^{N-1}(\partial C_k)/(N2^N)$. In particular, we have

$$D_{\eta_l} \;\geq\; \int_{\mathcal{R}} W_p(\nabla v^{\eta_l}(x))\,dx \;+\; \frac{\alpha}{N2^N}\mathcal{H}^{N-1}(\partial\mathcal{J}\cap\partial\mathcal{R})\,. \tag{23}$$

Now, we will "move" $\mathcal{C}_f = \mathcal{R}\cup\mathcal{J}$ upwards (in the direction x_N) in order to cover Ω_f (we will then translate v^{η_l} accordingly): Let $\kappa = 1+\sqrt{N}\max_{\xi\in\omega}|\nabla f(\xi)|$; this constant is such that

$$\mathcal{C}_f + \kappa\eta_l e_N \;\supset\; \Omega_f$$

as soon as l is large enough (so that $x_N > -1$ yields $x_N - \kappa\eta_l > -1 - \delta + \eta_l$ which clearly holds as soon as $\eta_l \le \delta/(1+\kappa)$).

We then define, for any l (large enough), the function $f_l \in BV(\omega)$ by $f_l(x') = \sup\{x_N < f(x') : (x', x_N - \kappa\eta_l) \in \mathcal{R}\}$, and for any $x \in \omega \times (-1,+\infty)$, we also define $v_l(x)$ by

$$v_l(x) \;=\; \begin{cases} v^{\eta_l}(x', x_N - \kappa\eta_l) & \text{if } -1 < x_N < f_l(x'), \\ 0 & \text{otherwise.} \end{cases}$$

By construction, the boundary of Ω_{f_l} (in $\omega \times (-1,+\infty)$) is a piecewise smooth compact set made of two parts: One part is contained in the (smooth) graph of f, $\partial\Omega_f$, and the rest, $\partial\Omega_{f_l} \cap \Omega_f$, is a subset of $(\partial\mathcal{J} \cap \partial\mathcal{R}) + \kappa\eta_l e_N$, which is a finite union of facets of hypercubes. On the other hand, $v_l \in W^{1,p}(\Omega_{f_l})$, with as a consequence of (23),

$$\int_{\Omega_{f_l}} W_p(\nabla v_l(x))\,dx \;+\; \frac{\alpha}{N2^N}\mathcal{H}^{N-1}(\partial\Omega_{f_l} \cap \Omega_f) \;\le\; D_{\eta_l}\,. \tag{24}$$

We fix $\alpha = N2^N$ and make the observation that $v_l = v^{\eta_l}(\cdot - \kappa\eta_l e_N)$ except on a set of measure $O(\eta_l)$ (the union of the cubes of $\mathcal{J}$ such that $\partial_*\Omega_h \cup \Sigma$ crosses an edge of the cube). Therefore, $v_l \to v$ as $l \to \infty$ in $L^1(\omega \times (-1,+\infty))$ (and, as well, $f_l \to f$). We can now fix l large enough so that $\|f_l - f\|_{L^1(\omega)} + \|v_l - v\|_{L^1(\omega\times(-1,+\infty))} < \varepsilon$ and

$$\begin{aligned}\int_{\Omega_{f_l}} W_p(\nabla v_l(x))\,dx \;+\; \mathcal{H}^{N-1}(\partial\Omega_{f_l}) \;&\le\; D_{\eta_l} \;+\; \mathcal{H}^{N-1}(\partial\Omega_f)\\ &\le\; \int_{\omega\times(-1,+\infty)} W_p(\nabla u(x))\,dx \;+\; \mathcal{H}^{N-1}(\partial_*\Omega_h) \;+\; 2\mathcal{H}^{N-1}(\Sigma) \;+\; (3 + 2^N N\sqrt{N})\varepsilon\,,\end{aligned}$$

where we have used (15), (16), (22), and (24). Observe eventually that if l is large enough, we also have (since $\liminf_{l\to\infty} \mathcal{H}^{N-1}(\partial\Omega_{f_l}) \ge \mathcal{H}^{N-1}(\partial\Omega_f)$ and using (15))

$$\mathcal{H}^{N-1}(\partial\Omega_{f_l}) \;\ge\; \mathcal{H}^{N-1}(\partial_*\Omega_h) \;+\; 2\mathcal{H}^{N-1}(\Sigma)) \;-\; 2\varepsilon\,.$$

Using now Lemma 4.3, we can find a smooth $f' \in C^\infty(\omega;\mathbb{R}^N)$ with $f' \le f_l$, close enough to f_l, in such a way that if $v' = v_l$ in Ω'_f and 0 in $(\omega \times (-1,+\infty)) \setminus \Omega'_f$, one has $\|f' - f\|_{L^1(\omega)} + \|v' - v\|_{L^1(\omega\times(-1,+\infty))} < 2\varepsilon$; hence, both $\|f' - h\|_{L^1(\omega)} < 3\varepsilon$ and $\|v' - u\|_{L^1(\omega\times(-1,+\infty))} < 3\varepsilon$, and

$$\begin{aligned}&\int_{\Omega_{f'}} W_p(\nabla v'))\,dx \;+\; \mathcal{H}^{N-1}(\partial\Omega_{f'})\\ &\qquad\le\; \int_{\omega\times(-1,+\infty)} W_p(\nabla u(x))\,dx \;+\; \mathcal{H}^{N-1}(\partial_*\Omega_h) \;+\; 2\mathcal{H}^{N-1}(\Sigma) \;+\; \beta\varepsilon,\end{aligned}$$

where $\beta = 4 + 2^N N\sqrt{N}$ is a constant and, as well,

$$\mathcal{H}^{N-1}(\partial\Omega_{f'}) \;\ge\; \mathcal{H}^{N-1}(\partial_*\Omega_h) \;+\; 2\mathcal{H}^{N-1}(\Sigma)) \;-\; 3\varepsilon\,.$$

Performing this construction for $\varepsilon = 1/n$, $n \ge 1$, yields the existence of two sequences $(f_n)_{n\ge1}$, $(u_n)_{n\ge1}$, with $f_n \in C^\infty(\omega)$, $u_n \in W^{1,p}(\Omega_{f_n})$, $f_n \to h$ in $L^1(\omega)$,

$u_n \to u$ in $L^1(\omega \times (-1,+\infty))$,

$$\limsup_{n\to\infty} \int_{\Omega_{f_n}} W_p(\nabla u_n(x))\,dx + \int_\omega \sqrt{1+|\nabla f_n(x)|^2}\,dx \\ \le \int_{\omega\times(-1,+\infty)} W_p(\nabla u(x))\,dx + \mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(\Sigma)\,, \tag{25}$$

and

$$\mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(\Sigma) \le \liminf_{n\to\infty} \int_\omega \sqrt{1+|\nabla f_n(x)|^2}\,dx\,. \tag{26}$$

The function u_n, extended with 0 out of Ω_{f_n}, is in $GSBV(\omega \times (-1,+\infty))$, and its gradient is ∇u_n in Ω_{f_n} and 0 outside. Invoking now Ambrosio's compactness theorem for $GSBV$ functions, we find that $\nabla u_n \rightharpoonup \nabla u$ in $L^p(\omega\times(-1,+\infty);\mathbb{R}^N)$, so that

$$\int_{\omega\times(-1,+\infty)} W_p(\nabla u(x))\,dx \le \liminf_{n\to\infty} \int_{\omega\times(-1,+\infty)} W_p(\nabla u_n(x))\,dx\,,$$

which, combined with (25) and (26), yields that

$$\lim_{n\to\infty} \int_{\omega\times(-1,+\infty)} W_p(\nabla u_n(x))\,dx = \int_{\omega\times(-1,+\infty)} W_p(\nabla u(x))\,dx\,, \tag{27}$$

$$\lim_{n\to\infty} \int_\omega \sqrt{1+|\nabla f_n(x)|^2}\,dx = \mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(\Sigma))\,. \tag{28}$$

In particular, we deduce from (27) (since $1<p<+\infty$) that ∇u_n goes strongly to ∇u in $L^p(\omega\times(-1,+\infty);\mathbb{R}^N)$. We also find that $u_n \to u^0$ strongly in $W^{1,p}(\omega\times(-1,0))$. Modifying u_n in order to ensure that $u_n \equiv u^0$ in $\omega\times(-1,0)$ is now not difficult. A simple way is as follows: We choose a continuous extension operator from $W^{1,p}(\omega\times(-1,0))$ to $W^{1,p}(\omega\times(-1,+\infty))$ and define, for all n, a function w_n as the extension of $(u_n|_{\omega\times(-1,0)} - u^0)$. Clearly, $w_n \to 0$ strongly in $W^{1,p}(\omega\times(-1,+\infty))$. The sequence u_n is then modified in the following way: We replace u_n with $u_n - w_n$ in Ω_{f_n}, letting it keep the value 0 outside. This new u_n satisfies the same properties as before, but additionally, $u_n = u^0$ a.e. in $\omega\times(-1,0)$. This shows the thesis. ◻

5. An approximation result. We introduce in this section, as in [9], a phase-field approximation of the functional $\overline{F}$. The idea is to represent the subgraph $\Omega_h\setminus\Sigma$ by a field v that will be an approximation of the characteristic function of this set, at a scale of order ε. Then, numerically, the minimization of our new functional will provide an approximation of (u,h) minimizing $\overline{F}$. Our approximated functional is the following:

$$F_\varepsilon(u,v) = \int_{\omega\times(0,+\infty)} (\eta_\varepsilon + v^2(x)) W(\nabla u(x))\,dx \\ + c_V\left(\frac{\varepsilon}{2}\int_{\omega\times(0,+\infty)} |\nabla v(x)|^2\,dx + \frac{1}{\varepsilon}\int_{\omega\times(0,+\infty)} V(v(x))\,dx\right) \tag{29}$$

if $u \in W^{1,p}(\omega\times(0,+\infty))$, with $u=u^0$ on $\omega\times\{0\}$, and $v\in H^1(\omega\times(0,+\infty))$, with $v=1$ on $\omega\times\{0\}$ and $\partial_N v \le 0$ a.e. in $\omega\times(0,+\infty)$. Otherwise, for all other $u,v \in L^1(\omega\times(0,+\infty))$, we let $F_\varepsilon(u,v)=+\infty$. Here the potential V is a two-wells potential with $V(t) > 0$ except if $t \in \{0,1\}$, $V(0) = V(1) = 0$, and $c_V^{-1} = \int_0^1 \sqrt{2V(t)}\,dt$. The parameter η_ε is any function of ε with $\eta_\varepsilon/(\varepsilon^{p-1}) \to 0$ as $\varepsilon \to 0$. The function u^0 is assumed to be the trace of a function in $W^{1,p}(\omega\times(-1,0))$, still denoted by u^0, and for technical reasons we also have to assume that it is bounded: $u^0 \in L^\infty(\omega\times(-1,0))$. The following results generalize in arbitrary dimension Theorem 3.1 of [9]. However, its proof also owes a lot to [13, sect. 5.2], where a similar approximation is studied.

THEOREM 5.1. *Let $(\varepsilon_j)_{j\geq 1}$ be a decreasing sequence of positive numbers, going to* 0. *Then the following hold.*

(i) *For any (u_j, v_j), if $\limsup_{j\to\infty} F_{\varepsilon_j}(u_j, v_j) < +\infty$, then up to a subsequence there exist u, v such that $v_j \to v$ in $L^1(\omega\times(0,+\infty))$ and $u_j(x) \to u(x)$ a.e. in $\{v = 1\}$, and there exists $h \in BV(\omega;\mathbb{R}_+)$ such that $\{v = 1\} = \Omega_h$ and*

$$\overline{F}(u,h) \ \leq\ \liminf_{j\to\infty} F_{\varepsilon_j}(u_j, v_j)\,. \tag{30}$$

(ii) *For any $h \in BV(\omega;\mathbb{R}_+)$ and $u \in GSBV_p(\omega\times(-1,+\infty))$ with $u = u^0$ in $\omega\times(-1,0)$ and $u(x) = 0$ a.e. in $\{x_N > h(x')\}$, there exists (u_j, v_j) such that $u_j \to u$ and $v_j \to \chi_{\Omega_h}$ in $L^1(\omega\times(0,+\infty))$ and*

$$\limsup_{j\to\infty} F_{\varepsilon_j}(u_j, v_j) \ \leq\ \overline{F}(u,h)\,. \tag{31}$$

This is almost a Γ-convergence result. We deduce, in particular, that if for all j, (u_j, v_j) is a minimizer of F_{ε_j}, then up to a subsequence, $v_j \to \chi_{\Omega_h}$ and $u_j \to u$ a.e. in Ω_h, where (u,h) minimize the relaxed functional $\overline{F}$.

Remark 5.2. The thesis of the theorem is still valid if (as in [9, Thm. 3.1]) the set Ω_h must satisfy a volume constraint $|\Omega_h| = V > 0$ (which is imposed in the approximation by a constraint on v_j: $\int_{\omega\times(0,+\infty)} v_j(x)\,dx = V$). The adaption of the proofs is easy; see Remark 4.2 above.

Proof of Theorem 5.1. We first show the first point. Let (u_j, v_j) be as in (i). Since $F_{\varepsilon_j}(u_j, v_j)$ is finite, v_j must be nondecreasing in x_N. Now, if we replace v_j by $\tilde v_j(x) = 0 \vee ((v_j(x) - \delta_j x_N)\wedge 1)$ and if δ_j is small enough, one can ensure that $F_{\varepsilon_j}(u_j, v_j) = F_{\varepsilon_j}(u_j, \tilde v_j) + O(1/j)$, and $\tilde v_j$ is strictly decreasing.

Assume first that v_j is smooth, so that $\tilde v_j$ is smooth in $\{0 < \tilde v_j < 1\}$. For any $s \in (0,1)$, let $h_j^s : \omega \to \mathbb{R}_+$ be the function such that $\tilde v_j(x', h_j^s(x')) = s$ for any $x' \in \omega$; then clearly, h_j^s is in $C^1(\omega)$, with

$$|\nabla' h_j^s(x')| \ =\ \frac{|\nabla'\tilde v_j(x', h_j^s(x'))|}{|\partial_N \tilde v_j(x', h_j^s(x'))|} \ \leq\ \frac{1}{\delta_j}|\nabla'\tilde v_j(x', h_j^s(x'))|$$

for any $x' \in \omega$. Now, we deduce that

$$\begin{aligned}
\int_\omega |\nabla' h_j^s(x')|^2\,dx' \ &\leq\ \frac{1}{\delta_j}\int_\omega \frac{|\nabla'\tilde v_j(x', h_j^s(x'))|^2}{|\partial_N \tilde v_j(x', h_j^s(x'))|}\,dx' \\
&\leq\ \frac{1}{\delta_j}\int_\omega |\nabla'\tilde v_j(x', h_j^s(x'))|^2 \left(\frac{\sqrt{1 + |\nabla' h_j^s(x')|^2}}{|\nabla \tilde v_j(x', h_j^s(x'))|}\right) dx' \\
&=\ \frac{1}{\delta_j}\int_{\partial\{\tilde v_j > s\}} \frac{|\nabla'\tilde v_j(x)|^2}{|\nabla\tilde v_j(x)|}\,d\mathcal{H}^{N-1}(x)\,.
\end{aligned}$$

Using the coarea formula, we find that

$$\int_0^1 \left(\int_\omega |\nabla' h_j^s(x')|^2 \, dx' \right) ds \;\le\; \frac{1}{\delta_j} \int_{\{1>\tilde v_j>0\}} |\nabla' \tilde v_j(x)|^2 \, dx \;<\; +\infty \,.$$

By approximation, we easily deduce that this remains true when v_j is just in $H^1(\omega \times (0,+\infty))$: We get that for a.e. level $s \in (0,1)$, the set $\{\tilde v_j > s\}$ can be represented as the subgraph of a function $h_j^s \in H^1(\omega)$. We may also assume that this is true for all $j \ge 1$.

Now, we notice that (using $a^2 + b^2 \ge 2ab$ and the coarea formula)

$$\begin{aligned}(32)\quad \frac{\varepsilon_j}{2} \int_{\omega\times(0,+\infty)} |\nabla \tilde v_j(x)|^2 \, dx \;+\; \frac{1}{\varepsilon_j} \int_{\omega\times(0,+\infty)} V(\tilde v_j(x)) \, dx \\ \ge \int_{\omega\times(0,+\infty)} \sqrt{2V(\tilde v_j(x))} |\nabla \tilde v_j(x)| \, dx \\ \ge \int_0^1 \sqrt{2V(s)} \left(\int_\omega \sqrt{1 + |\nabla' h_j^s(x')|^2} \, dx' \right)\end{aligned}$$

and, in particular, using Fatou's lemma, we see that

$$\begin{aligned}\int_0^1 \sqrt{2V(s)} \left(\liminf_{j\to\infty} \int_\omega \sqrt{1 + |\nabla' h_j^s(x')|^2} \, dx' \right) \\ \le \liminf_{j\to\infty} \left(\frac{\varepsilon_j}{2} \int_{\omega\times(0,+\infty)} |\nabla \tilde v_j(x)|^2 \, dx \;+\; \frac{1}{\varepsilon_j} \int_{\omega\times(0,+\infty)} V(\tilde v_j(x)) \, dx \right).\end{aligned}$$

In particular, for a.e. $s \in (0,1)$, $h_j^s \in H^1(\omega)$ for all $j \ge 1$, and in addition, $\liminf_{j\to\infty} \sqrt{1 + |\nabla' h_j^s|^2}$ is finite.

By a diagonal argument, we can find a subsequence (still denoted by (ε_j)) and a decreasing sequence $(s_n)_{n\ge1}$ of real numbers in $(0,1)$ with $\lim_{n\to\infty} s_n = 0$ and such that, for each n,

$$\lim_{j\to\infty} \int_\omega \sqrt{1 + |\nabla' h_j^{s_n}(x')|^2} \, dx' \;=\; \liminf_{j\to\infty} \int_\omega \sqrt{1 + |\nabla' h_j^{s_n}(x')|^2} \, dx' \;<\; +\infty \,.$$

We can also assume that, for each n, $h_j^{s_n}$ converges in $L^1(\omega)$ to some function h^{s_n}, and since it is then clear (since $V(\tilde v_j(x)) \to 0$ a.e. in $\omega \times (0,+\infty)$) that $\tilde v_j(x) \to 0$ for a.e. x with $x_N > h^{s_n}(x')$ and $\tilde v_j(x) \to 1$ for a.e. x with $x_N < h^{s_n}(x')$, this function is independent on n and will be denoted simply by h.

For any $n \ge 1$, let us denote by u_j^n the function given by $u_j(x)$ if $x_N < h_j^{s_n}(x')$ and by 0 otherwise; let us show that $(u_j^n)_{j\ge1}$ is compact in $GSBV(\omega \times (-1,+\infty))$. One has $u_j^n \in W^{1,p}(\{x \,:\, -1 < x_N < h_j^{s_n}(x')\})$; hence, $u_j^n \in GSBV(\omega \times (-1,+\infty))$ with $S_{u_j^n} \subseteq \{(x', h_j^{s_n}(x')) \,:\, x' \in \omega\}$. In particular,

$$\mathcal{H}^{N-1}(S_{u_j^n}) \;\le\; \int_\omega \sqrt{1 + |\nabla' h_j^{s_n}(x')|^2} \, dx'$$

is uniformly bounded (in j). On the other hand,

$$F_{\varepsilon_j}(u_j, \tilde v_j) \;\ge\; (\eta_{\varepsilon_j} + s_n^2) \int_{\omega\times(0,+\infty)} W(\nabla u_j^n(x)) \, dx$$

showing that ∇u_j^n is uniformly bounded in $L^p(\omega \times (-1,+\infty); \mathbb{R}^N)$.

Now, for any $x' \in \omega$, if we denote by $\hat{u}_j^n$ the function $u_j^n - u^0$ (where u^0 is appropriately extended to a function in $W^{1,p}(\omega \times (-1,+\infty))$ that vanishes for $x_N \geq 1$), one sees that, for any x with $x_N < h_j^{s_n}(x')$,

$$|\hat{u}_j^n(x)| \;\leq\; \int_0^{x_N} |\partial_N \hat{u}_j^n(x',s)|\,ds \;\leq\; x_N^{1-1/p} \left(\int_0^{x_N} |\partial_N \hat{u}_j^n(x',s)|^p\,ds\right)^{1/p}$$

so that, for any $M > 0$ and a.e. $x' \in \omega$,

$$\int_0^{M\wedge h_j^{s_n}(x')} |\hat{u}_j^n(x',s)|\,ds \;\leq\; \frac{M^{2-1/p}}{2^{1-1/p}} \left(\int_0^{h_j^{s_n}(x')} |\partial_N \hat{u}_j^n(x',s)|^p\,ds\right)^{1/p}.$$

We get

$$\|\hat{u}_j^n\|_{L^1(\omega\times(-1,M))} \;\leq\; C(M)\|\partial_N \hat{u}_j^n\|_{L^p(\omega\times(-1,+\infty))}.$$

Therefore, $u_j^n = \hat{u}_j^n + u^0$ is uniformly bounded in $L^1_{\text{loc}}(\omega \times (-1,+\infty))$. By Ambrosio's compactness theorem we deduce that there exists $u^n \in GSBV_p(\omega\times(-1,+\infty))$ such that $u_j^n(x) \to u^n(x)$ a.e. in $\omega \times (-1,+\infty)$, up to a subsequence.

By a diagonal argument, we can extract a subsequence (still denoted by $(\varepsilon_j)_{j\geq 1}$) such that for each $n \geq 0$, $u_j^n(x) \to u^n(x)$ a.e. as $\varepsilon_j \to 0$. Now, by construction we have that if $n' \geq n$, then $u_j^{n'}(x) = u_j^n(x)$ a.e. in $\{x_N < h_j^n(x')\}$. From this we deduce that $u^{n'}(x) = u^n(x)$ a.e. in $\{x_N < h(x')\}$, and since moreover one checks easily that both functions vanish a.e. in $\{x_N > h(x')\}$, one deduces that u^n, which is simply denoted by u in the following, is independent on n.

We have shown the first assertion of point (i) of Theorem 5.1: Indeed, if we let $v = \chi_{\Omega_h}$, one sees that $\tilde{v}_j(x) \to v(x)$ a.e., and by construction also $v_j(x) \to v(x)$ a.e. in $\omega \times (0,+\infty)$. Moreover, $u_j(x) \to u(x)$ a.e. in $\{x \in \omega\times(-1,+\infty) \,:\, x_N < h(x)\}$, with $u = u^0$ in $\omega \times (-1,0)$. The function u is in $GSBV_p(\omega\times(-1,+\infty))$ and vanishes above the graph of h.

Let us now show (30). We follow a similar proof as that in [13]. We have

$$\begin{aligned}\int_{\omega\times(0,+\infty)} (\eta_{\varepsilon_j} + \tilde{v}_j^2(x)) W(\nabla u_j(x))\,dx \;&\geq\; \int_{\omega\times(0,+\infty)} \left(2\int_0^{\tilde{v}_j(x)} s\,ds\right) W(\nabla u_j(x))\,dx \\ &\geq\; \int_0^1 2s \left(\int_{\{\tilde{v}_j(x)>s\}} W(\nabla u_j(x))\,dx\right) ds\,.\end{aligned}$$

This inequality, together with (32), yields

$$\begin{aligned}&F_{\varepsilon_j}(u_j,\tilde{v}_j) \;\geq \\ &\qquad \int_0^1 \left(2s\int_{\{\tilde{v}_j(x)>s\}} W(\nabla u_j(x))\,dx \,+\, c_V\sqrt{2V(s)}\int_\omega \sqrt{1+|\nabla' h_j^s(x')|^2}\,dx'\right) ds\,.\end{aligned}$$

By Fatou's lemma, we deduce that

$$\begin{aligned}&\int_0^1 \liminf_{j\to\infty} \left(2s\int_{\{\tilde{v}_j(x)>s\}} W(\nabla u_j(x))\,dx \,+\, c_V\sqrt{2V(s)}\int_\omega \sqrt{1+|\nabla' h_j^s(x')|^2}\,dx'\right) ds \\ &\qquad\qquad \leq\; \liminf_{j\to\infty} F_{\varepsilon_j}(u_j,\tilde{v}_j) \;<\; +\infty\,.\end{aligned} \tag{33}$$

Therefore, for a.e. $s \in (0,1)$,

$$\liminf_{j\to\infty}\ 2s\int_{\{\tilde v_j(x)>s\}} W(\nabla u_j(x))\,dx \,+\, c_V\sqrt{2V(s)}\int_\omega \sqrt{1+|\nabla' h_j^s(x')|^2}\,dx' \ <\ +\infty\,.$$

Let us choose such an s, with additionally $h_j^s \in H^1(\omega)$ for all $j \ge 1$, and let us consider a subsequence $(j_k)_{k\ge1}$ such that

$$\begin{aligned}\lim_{k\to\infty} 2s\int_{\{\tilde v_{j_k}(x)>s\}} & W(\nabla u_{j_k}(x))\,dx \,+\, c_V\sqrt{2V(s)}\int_\omega \sqrt{1+|\nabla' h_{j_k}^s(x')|^2}\,dx' \\ &= \liminf_{j\to\infty}\ 2s\int_{\{\tilde v_j(x)>s\}} W(\nabla u_j(x))\,dx \,+\, c_V\sqrt{2V(s)}\int_\omega \sqrt{1+|\nabla' h_j^s(x')|^2}\,dx'\,.\end{aligned}$$

As above, let us introduce the sequence of functions $u_{j_k}^s \in GSBV_p(\omega\times(-1,+\infty))$ such that $u_{j_k}^s(x) = u_{j_k}(x)$ if $x_N < h_{j_k}^s(x')$ and 0 otherwise. By compactness, we easily check that $u_{j_k}^s(x) \to u(x)$ a.e. in $\omega\times(-1,+\infty)$, while $h_{j_k}^s \to h$ in $L^1(\omega)$. By the l.s.c. property (P1), we deduce

$$\begin{aligned} 2s\int_{\Omega_h^+} & W(\nabla u) \,+\, c_V\sqrt{2V(s)}(\mathcal{H}^{N-1}(\partial_*\Omega_h) + 2\mathcal{H}^{N-1}(S'_u\cap\Omega_h^1)) \\ &\le \lim_{k\to\infty} 2s\int_{\{\tilde v_{j_k}(x)>s\}} W(\nabla u_{j_k}(x))\,dx \,+\, c_V\sqrt{2V(s)}\int_\omega \sqrt{1+|\nabla' h_{j_k}^s(x')|^2}\,dx'\,.\end{aligned}$$

Integrating (33) on $(0,1)$ and recalling that by construction $F_{\varepsilon_j}(u_j,\tilde v_j) = F_{\varepsilon_j}(u_j,v_j) + o(1)$, we deduce (30).

Let us now show point (ii) of Theorem 5.1. The proof follows the same lines as in [9], where the same inequality is shown in the 2D case, and we will only sketch it.

Let $h \in BV(\omega;\mathbb{R}_+)$, and let $u \in GSBV_p(\omega\times(-1,+\infty))$, with $u = u^0$ in $\omega\times(-1,0)$ and $u(x) = 0$ a.e. in $\{x_N > h(x')\}$, with $\overline{F}(u,h) < +\infty$. By Theorem 2.2, there exist h_n in $C^1(\omega;\mathbb{R}_+)$ and $u^n \in W^{1,p}(\Omega_h;\mathbb{R})$, with $u^n = u^0$ in $\omega\times(-1,0)$, $h_n \to h$ in $L^1(\omega)$, and $u^n \to u$ a.e. in $\omega\times(0,+\infty)$, with

$$\limsup_{n\to\infty} F(u^n,h_n) = \overline{F}(u,h)\,.$$

By construction (since we have assumed $u^0 \in L^\infty(\omega\times(-1,0))$), one also has that $u^n \in L^\infty(\omega\times(0,+\infty))$. Now, we construct sequences $(u_j^n)_j$ and $(v_j^n)_j$ with $u_j^n \to u^n$ in $L^1(\omega\times(0,+\infty))$ and $v_j^n \to \chi_{\Omega_{h_n}}$ in $L^1(\omega\times(0,+\infty))$ such that

$$\limsup_{j\to\infty} F_{\varepsilon_j}(u_j^n,v_j^n) \le F(u^n,h_n). \tag{34}$$

Let us condider the sequence of functionals

$$H_\varepsilon(v) = \frac{\varepsilon}{2}\int_{\omega\times(0,+\infty)} |\nabla v(x)|^2\,dx + \frac{1}{\varepsilon}\int_{\omega\times(0,+\infty)} V(v(x))\,dx;$$

the celebrated Γ-convergence result of Modica and Mortola for such functionals (see [1]) allows us to find, for each n, a sequence $(v_j^n)_j$ converging to the characteristic function $\chi_{\Omega_{h_n}}$ such that

$$\begin{aligned}\limsup_{j\to\infty} H_{\varepsilon_j}(v_j^n) &= \int_0^1 \sqrt{2V(s)}\,ds\ \mathcal{H}^{N-1}(S_{\chi_{\Omega_{h_n}}}\cap\omega\times(0,+\infty)) \\ &= c_V^{-1}\mathcal{H}^{N-1}(\partial\Omega_{h_n}).\end{aligned}$$

We recall that the explicit construction of the recovery sequence $(v_j^n)_j$ can be obtained in the following way: One considers γ_j solution of the Euler's equation of the functional with appropriate boundary conditions, namely,

$$\begin{cases} -\gamma_j'' + V'(\gamma_j) = 0, \\ \gamma_j(0) = 1, \quad \gamma_j\Big(\dfrac{1}{\sqrt{\varepsilon_j}}\Big) = 0\,. \end{cases}$$

This function is extended by 0 beyond $1/\sqrt{\varepsilon_j}$. One then lets

$$v_j^n(x) = \gamma_j\Big(\frac{\operatorname{dist}(x,\Omega_{h_n}^+)}{\varepsilon_j}\Big).$$

Then, the sequence $(u_j^n)_j$ is constructed by translating u_n and multiplying by an appropriate cut-off function, as in [9]. We first choose $c_n \geq \max\{1, \|\nabla h_n\|_{L^\infty(\omega)}\}$, and let $w_j^n(x) := v_j^n(x', x_N - c_n\sqrt{2\varepsilon_j})$. This function is 1 on the support of v_j^n and vanishes shortly beyond. Then, we let $u_j^n(x) = u^n(x', x_N - 2c_n\sqrt{2\varepsilon_j})w_j^n(x)$. (As in the end of the proof of Proposition 4.1, we have to modify slightly u_j^n in order to ensure $u_j^n = u^0$ in $\omega\times(-1,0)$; however, this is easily done, and one checks that this modified u_j^n satisfies a uniform (in j) L^∞ bound.) In order to show that (34) holds, we just need to check

$$\limsup_{j\to\infty} \int_{\omega\times(0,+\infty)} (\eta_{\varepsilon_j} + (v_j^n(x))^2) W(\nabla u_j^n(x))\,dx \;\leq\; \int_{\Omega_{h_n}^+} W(\nabla u^n(x))\,dx\,. \tag{35}$$

Since $\nabla u_j^n(x) = w_j^n(x)\nabla u^n(x', x_N - 2c_n\sqrt{2\varepsilon_j}) + u^n(x', x_N - 2c_n\sqrt{2\varepsilon_j})\nabla w_j^n(x)$, this inequality is clear as soon as we have established that

$$\limsup_{j\to\infty}\ \eta_{\varepsilon_j} \int_{\omega\times(0,+\infty)} |u^n(x', x_N - 2c_n\sqrt{2\varepsilon_j})\nabla w_j^n(x)|^p\,dx = 0,$$

and since u^n is bounded in L^∞, we need to show

$$\limsup_{j\to\infty}\ \eta_{\varepsilon_j} \int_{\omega\times(0,+\infty)} |\nabla w_j^n(x)|^p\,dx = 0\,.$$

This integral is bounded by

$$\begin{aligned}
&\int_{\{0<\operatorname{dist}(x,\Omega_{h_n}^+)<\sqrt{\varepsilon_j}\}} \frac{|\gamma_j'|^p\,\big(\operatorname{dist}(x,\Omega_{h_n}^+)/\varepsilon_j\big)}{\varepsilon_j^p}\,dx \\
&\qquad = \int_0^{\sqrt{\varepsilon_j}} \frac{|\gamma_j'|^p(s/\varepsilon_j)}{\varepsilon_j^p}\mathcal{H}^{N-1}(\{\operatorname{dist}(\cdot,\Omega_{h_n}^+) = s\})\,ds \\
&\qquad\qquad = \frac{1}{\varepsilon_j^{p-1}} \int_0^{1/\sqrt{\varepsilon_j}} |\gamma_j'|^p(s)\mathcal{H}^{N-1}(\{\operatorname{dist}(\cdot,\Omega_{h_n}^+) = \varepsilon_j s\})\,ds\,.
\end{aligned}$$

Now, one can show that

$$\int_0^{1/\sqrt{\varepsilon_j}} |\gamma_j'|^p(s)\mathcal{H}^{N-1}(\{\operatorname{dist}(\cdot,\Omega_{h_n}^+) = \varepsilon_j s\})\,ds \to \mathcal{H}^{N-1}(\partial\Omega_{h_n}^+)\int_0^1 \sqrt{2V(t)}^{\,p-1}\,dt$$

as $j \to \infty$; hence, since we have assumed $\eta_\varepsilon/\varepsilon^{p-1} \to 0$ as $\varepsilon \to 0$, we deduce (35) and (34).

Since (34) holds, a standard diagonal extraction argument allows us to find subsequences $(u^{n_k}_{j_k})_k$, $(v^{n_k}_{j_k})_k$ satisfying point (ii) of Theorem 5.1, and this completes the proof of the theorem.

Appendix A. A simple inequality.

LEMMA A.1. *Let $w \in C^1([0,1]^N)$ satisfy, for any $x \in [0,1]^N$,*

$$w(x) \;=\; \sum_{k\in\{0,1\}^N} w(k)\Delta(x-k),$$

where Δ is defined in (21)*. Then, for any $p \geq 1$,*

$$\int_{(0,1)^N} W_p(\nabla w(x))\,dx \;\leq\; \frac{1}{2^{N-1}}\sum_{i=1}^{N}\sum_{\substack{k\in\{0,1\}^N\\ k_i=0}} |w(k+e_i)-w(k)|^p. \tag{36}$$

Proof. We show that, for each i,

$$\int_{(0,1)^N}\left|\frac{\partial w}{\partial x_i}(x)\right|^p dx \;\leq\; \frac{1}{2^{N-1}}\sum_{\substack{k\in\{0,1\}^N\\ k_i=0}} |w(k+e_i)-w(k)|^p\,.$$

We will show this inequality for $i = N$. Let us denote, for $x' = (x_1,\dots,x_{N-1})$,

$$\Delta_{N-1}(x') \;=\; \prod_{i=1}^{N-1}(1-|x_i|)^+\,.$$

Then, for any $x \in (0,1)^N$,

$$\begin{aligned} w(x) \;&=\; \sum_{k\in\{0,1\}^N} w(k)\Delta(x-k)\\ &=\; \sum_{k'\in\{0,1\}^{N-1}} \Delta_{N-1}(x'-k')(w_{k',0}(1-x_N)+w_{k',1}x_N) \end{aligned}$$

so that

$$\frac{\partial w}{\partial x_N}(x) \;=\; \sum_{k'\in\{0,1\}^{N-1}} \Delta_{N-1}(x'-k')(w_{k',1}-w_{k',0})\,.$$

Now, at any x, we have $\sum_{k'\in\{0,1\}^{N-1}} \Delta_{N-1}(x'-k') = 1$ so that this is a convex combination of $(w_{k',1}-w_{k',0})_{k'\in\{0,1\}^{N-1}}$. Hence, by convexity of the function $|\cdot|^p$,

$$\int_{(0,1)^N}\left|\frac{\partial w}{\partial x_N}(x)\right|^p dx \;\leq\; \int_{(0,1)^N}\sum_{k'\in\{0,1\}^{N-1}} \Delta_{N-1}(x'-k')|w_{k',1}-w_{k',0}|^p\,dx\,.$$

We deduce (36) by simply observing that, for any $k' \in \{0,1\}^{N-1}$,

$$\int_{(0,1)^N}\Delta_{N-1}(x'-k')\,dx \;=\; \int_0^1 dx_N \times \prod_{i=1}^{N-1}\int_0^1 (1-|x_i-k_i|)^+\,dx_i \;=\; \frac{1}{2^{N-1}}\,. \qquad \square$$

Acknowledgments. The authors would like to thank A. Braides, E. Bonnetier, and F. Jouve for stimulating discussions and the two latter for providing the result of the numerical experiment shown in Figure 1. The authors are also grateful to the anonymous referees, whose comments helped improve a previous version of this work.

REFERENCES

[1] G. Alberti, *Variational models for phase transitions, an approach via* Γ*-convergence*, in Calculus of Variations and Partial Differential Equations, Springer, Berlin, 2000, pp. 95–114.

[2] R. Alicandro, M. Focardi, and M. S. Gelli, *Finite-difference approximation of energies in fracture mechanics*, Ann. Sc. Norm. Super. Pisa Cl. Sci. (4), 29 (2000), pp. 671–709.

[3] L. Ambrosio, *A compactness theorem for a new class of functions with bounded variation*, Boll. Un. Mat. Ital. Ser. B, (7) (1989), pp. 857–881.

[4] L. Ambrosio, *On the lower semicontinuity of quasiconvex integrals in* $SBV(\Omega, \mathbb{R}^k)$, Nonlinear Anal., 23 (1994), pp. 405–425.

[5] L. Ambrosio, *A new proof of the SBV compactness theorem*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 127–137.

[6] L. Ambrosio, A. Coscia, and G. Dal Maso, *Fine properties of functions with bounded deformation*, Arch. Ration. Mech. Anal., 139 (1997), pp. 201–238.

[7] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Clarendon Press, New York, 2000.

[8] G. Bellettini, A. Coscia, and G. Dal Maso, *Compactness and lower semicontinuity properties in* $\mathrm{SBD}(\Omega)$, Math. Z. 228 (1998), pp. 337–351.

[9] E. Bonnetier and A. Chambolle, *Computing the equilibrium configuration of epitaxially strained crystalline films*, SIAM J. Appl. Math., 62 (2002), pp. 1093–1121.

[10] E. Bonnetier, A. Chambolle, and F. Jouve, *A variational model for equilibrium shapes of hetero-epitaxial films*, in PICOF'02: 2$^{\text{é}}$ Colloque sur les problèmes inverses, le contrôle et l'optimisation de formes, M. Jaoua, J. Jaffré, and T. Ha Duong, eds., Carthage, Tunisia, 2002, pp. 235–240.

[11] E. Bonnetier, R. S. Falk, and M. A. Grinfeld, *Analysis of a one-dimensional variational model of the equilibrium shape of a deformable crystal*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 573–591.

[12] B. Bourdin and A. Chambolle, *Implementation of an adaptive finite-element approximation of the Mumford-Shah functional*, Numer. Math., 85 (2000), pp. 609–646.

[13] A. Braides, A. Chambolle, and M. Solci, *A relaxation result for energies defined on pairs set-function and applications*, ESAIM Control Optim. Calc. Var., to appear.

[14] A. Chambolle, *Finite-differences discretizations of the Mumford-Shah functional*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 261–288.

[15] A. Chambolle, *An approximation result for special functions with bounded deformation*, J. Math. Pures Appl. (9), 83 (2004), pp. 929–954.

[16] A. Chambolle, *Addendum to: "An approximation result for special functions with bounded deformation" [J. Math. Pures Appl. (9), 83 (2004), pp. 929–954; mr2074682]*, J. Math. Pures Appl. (9), 84 (2005), pp. 137–145.

[17] G. Dal Maso, G. A. Francfort, and R. Toader, *Quasi-static evolution in brittle fracture: The case of bounded solutions*, in Calculus of Variations: Topics from the Mathematical Heritage of E. De Giorgi, Quad. Mat. 14, Dept. Math., Seconda Univ. Napoli, Caserta, 2004, pp. 245–266.

[18] G. Dal Maso, G. A. Francfort, and R. Toader, *Quasistatic crack growth in nonlinear elasticity*, Arch. Ration. Mech. Anal., 176 (2005), pp. 165–225.

[19] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[20] G. A. Francfort and C. J. Larsen, *Existence and convergence for quasi-static evolution in brittle fracture*, Comm. Pure Appl. Math., 56 (2003), pp. 1465–1500.

[21] A. Giacomini and M. Ponsiglione, *A* Γ*-convergence approach to stability of unilateral minimality properties*, in *Fracture mechanics and applications*, Arch. Ration. Mech. Anal., 180 (2006), pp. 399–447.

[22] M. Gobbino, *Finite difference approximation of the Mumford–Shah functional*, Comm. Pure Appl. Math., 51 (1998), pp. 197–228.

[23] J. Kristensen, *Lower semicontinuity in spaces of weakly differentiable functions*, Math. Ann., 313 (1999), pp. 653–710.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 103–125

TRAVELING WAVEFRONTS IN A DELAYED FOOD-LIMITED POPULATION MODEL*

CHUNHUA OU† AND JIANHONG WU‡

Abstract. In this paper we develop a new method to establish the existence of traveling wavefronts for a food-limited population model with nonmonotone delayed nonlocal effects. Our approach is based on a combination of perturbation methods, the Fredholm theory, and the Banach fixed point theorem. We also develop and theoretically justify Canosa's asymptotic method for the wavefronts with large wave speeds. Numerical simulations are provided to show that there exists a prominent hump when the delay is large.

Key words. traveling wave fronts, nonmonotone, nonlocal, food-limited

AMS subject classifications. 35K55, 35K57, 35R10, 92D25

DOI. 10.1137/050638011

1. Introduction.

There has been some success in establishing the existence of traveling wavefronts for the reaction-diffusion equation with nonlocal delayed nonlinearity. When the nonlinearity is monotone, the existence of traveling wavefronts can be obtained by extension of the methods of the super/subsolution pair [1], [7], [29], homotopy [3], and Leray-Schauder degree [28]. Unfortunately, when the delayed nonlinearity is no longer monotone, very little has been achieved (except for the work in [9]). While one suspects that the method developed by Wu and Zou [29] and based on a nonstandard ordering could be applicable, the construction of a supersolution and subsolution pair is nontrivial, and it is almost as difficult as solving the original given equations. In this paper we develop a new approach to establish the existence of traveling wavefronts in the case when the delayed nonlinearity is nonmonotone. We shall demonstrate this approach by considering the following food-limited reaction-diffusion equation

$$\frac{\partial u}{\partial t}(t,x) = \frac{\partial^2 u}{\partial x^2}(t,x) + u(t,x)\frac{1-(f*u)(t,x)}{1+\gamma(f*u)(t,x)}, \tag{1.1}$$

where the parameter $\gamma > 0$, and the spatiotemporal convolution $f*u$ is defined by

$$f*u = \int_{-\infty}^{t}\int_{-\infty}^{\infty} f(t,s,x,y)u(s,y)dyds, \tag{1.2}$$

with the kernel $f(t,s,x,y)$ satisfying the normalization condition

$$\int_{-\infty}^{t}\int_{-\infty}^{\infty} f(t,s,x,y)dyds = 1.$$

*Received by the editors August 11, 2005; accepted for publication (in revised form) December 27, 2006; published electronically April 13, 2007. This work was partially supported by the Canada Research Chairs Program, by the NCE Center of Mathematics for Information Technology and Complex Systems, and by the Natural Sciences and Engineering Research Council of Canada.

http://www.siam.org/journals/sima/39-1/63801.html

†Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics, York University, Toronto, Ontario, M3J 1P3, Canada. Current address: Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, Canada (ou@math.mun.ca).

‡Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics, York University, Toronto, Ontario, M3J 1P3, Canada (wujh@mathstat.yorku.ca).

The simplest version of (1.1) without diffusion is the following ODE:

$$\frac{du}{dt} = ru(t)\frac{K - u(t)}{K + \gamma u(t)}, \tag{1.3}$$

where r, K, and γ are positive constants. This equation was first proposed by Smith [25] as a mathematical model for population of *Daphnia* (water flea), and a derivation of this equation is given in [23]. The equation can also be used to study the effects of environmental toxicants on aquatic populations [16].

The delayed food-limited model

$$\frac{du}{dt} = ru(t)\frac{K - u(t-\tau)}{K + \gamma u(t-\tau)}, \quad \tau > 0, \tag{1.4}$$

has been studied recently by several authors; see [13], [17], [27], [18], and [8]. It seems that the best result for the local stability of the positive equilibrium $u = K$ is given in [27]. For the first time, the global stability of the positive equilibrium was established in [18]; see also [8] for further generalizations.

Equation (1.3) incorporating spatial dispersal was investigated by Feng and Lu [11]. They considered both the reaction-diffusion equation without time delay

$$\frac{\partial u}{\partial t} - Au(t,x) = r(x)u(t,x)\frac{K(x) - u(t,x)}{K(x) + \gamma(x)u(t,x)} \tag{1.5}$$

and the corresponding time-delay model

$$\frac{\partial u}{\partial t} - Au(t,x) = r(x)u(t,x)\frac{K(x) - au(t,x) - bu(t-\tau,x)}{K(x) + a\gamma(x)u(t,x) + b\gamma(x)u(t-\tau,x)}, \tag{1.6}$$

where $x = (x_1, x_2, \ldots, x_n) \in \Omega \subseteq R^n$, with Ω bounded and the operator A, given by

$$A = \sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{j=1}^{n} \beta_j(x)\frac{\partial}{\partial x_j},$$

is uniformly strongly elliptic and has coefficient functions that are uniformly Hölder continuous in $\bar{\Omega}$. Feng and Lu studied the above problems subject to general boundary conditions that include both the zero-Dirichlet and zero-Neumann cases, and they established a global convergence result for the nonzero steady state.

We are here concerned about the general case (1.1), and we first note that this includes various types of special cases by choosing the kernel function f.

(i) If the kernel f is taken to be

$$f(t,s,x,y) = \delta(t-s)\delta(x-y),$$

(1.1) becomes the reaction-diffusion equation without delay

$$\frac{\partial u}{\partial t}(t,x) = \frac{\partial^2 u}{\partial x^2}(t,x) + u(t,x)\frac{1 - u(t,x)}{1 + \gamma u(t,x)}, \tag{1.7}$$

which is a special case of (1.5).

(ii) If the kernel function f has a discrete time lag τ and spatial averaging, that is,

$$f(t,s,x,y) = \frac{1}{\sqrt{4\pi(t-s)}}e^{-(x-y)^2/4(t-s)}\delta(t-s-\tau),$$

then (1.1) becomes

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u\frac{1 - \int_{-\infty}^{\infty} \frac{e^{-(x-y)^2/4\tau}}{\sqrt{4\pi\tau}} u(t-\tau, y)dy}{1 + \gamma \int_{-\infty}^{\infty} \frac{e^{-(x-y)^2/4\tau}}{\sqrt{4\pi\tau}} u(t-\tau, y)dy}. \tag{1.8}$$

A derivation of this type of model, using probabilistic arguments, was given in [4]. In this model, the movement of individuals to their present positions from where they have been at previous times is accounted for by a spatial convolution with a kernel that spreads normally with a dependence on the delay.

(iii) If $f(t, s, x, y) = \delta(x - y)G(t - s)$, where

$$G(t) = \frac{1}{\tau}e^{-t/\tau} \quad \text{or} \quad G(t) = \frac{t}{\tau^2}e^{-t/\tau}, \tag{1.9}$$

(1.1) becomes a model of reaction diffusion equation with distributed delay:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u\frac{1 - \int_{-\infty}^{t} G(t-\eta)u(\eta, x)d\eta}{1 + \gamma \int_{-\infty}^{t} G(t-\eta)u(\eta, x)d\eta}. \tag{1.10}$$

The parameter τ measures time delay and is comparable to the discrete delay τ in (1.8). The two kernel functions G in (1.9) are used frequently in the literature on delay differential equations. The first of the two functions G is sometimes called the "weak" generic kernel because it implies that the importance of events in the past decreases exponentially. The second kernel (the "strong" generic case) is different because it implies that a particular time in the past, namely, τ time units ago, is more important than any other since this kernel achieves its unique maximum when $t = \tau$. This kernel can be viewed as a smoothed out version of the case $G(t) = \delta(t-\tau)$, which gives rise to the discrete delay model.

(iv) If the kernel f is taken to be

$$f(t, s, x, y) = \frac{1}{\sqrt{4\pi(t-s)}}e^{-(x-y)^2/4(t-s)}G(t-s),$$

then (1.1) is a reaction diffusion equation with both distributed delay and spatial averaging. In the distributed delay case with $G(t) = \frac{1}{\tau}e^{-t/\tau}$, a formal asymptotic expansion of traveling wavefront to (1.1) when τ is small was found recently by Gourley and Chaplain by using the so-called linear chain techniques; see [14], [15]. But the convergence of this series or the proof of validity of this expansion has been absent. The central idea of this trick is to recast the traveling wave equation into a higher dimensional system of ODEs without delay. When τ is small, Fenichel's geometrical singular perturbation theory (see [12] or part two of [2]) is applicable. As mentioned in [14], if $G(t) = \frac{t}{\tau^2}e^{-t/\tau}$, linear chain techniques are still applicable, but the system of traveling wave profile equation is six-dimensional. While the trick remains to be effective theoretically, it will be much more difficult in practice. Apparently, it is well known that the drawback of this method is that it is applicable only for models with the special distributed delays. One cannot extend this technique to the discrete case. Another disadvantage of this approach is that if the unperturbed system ($\tau = 0$) is a higher dimensional system, the construction of a traveling wavefront is extraordinarily difficult.

As mentioned in section 3 of [14], traveling wave solutions to (1.1) in the discrete case are much more difficult to study than in the distributed case with specific kernels, because we are no longer able to recast the wave profile equation of (1.8) into a nondelay equation, and thus Fenichel's geometrical singular perturbation theory cannot be directly used to find a heteroclinic connection in a finite dimensional manifold. Generally, it is well known that in this case the search for traveling wavefronts becomes a much more difficult task.

The purpose of this paper is to develop a completely new approach suitable for all of the aforementioned cases for the existence of traveling wavefronts. The principal result can be stated as follows.

THEOREM 1.1. *For any fixed constant* $c \geq 2$, *there exists a real number* $\delta = \delta(c) > 0$ *so that for* $\tau \in [0, \delta]$, (1.1) *possesses a traveling wavefront* $u(t, x) = U(x - ct)$ *satisfying* $U(-\infty) = 1$ *and* $U(\infty) = 0$.

We should remark that our approach here can be developed to study more general equations including

$$\frac{\partial u}{\partial t}(t,x) = \frac{\partial^2 u}{\partial t^2}(t,x) + u(t,x)H((f * u)(t,x)),$$

with $H(\cdot)$ a decreasing function; we refer to [20] for this development.

Although the main result should be of interest, we wish to emphasize the novelty of the approach that we develop. This approach is based on a combination of the perturbation analysis, the Fredholm operator theory, and the fixed point theorems and is expected to find applications in other models as well. A detailed proof is given for case (ii) and is sketched to emphasize the key differences for cases (iii) and (iv). This approach does not work when the delay is not small. In the case where the delay is arbitrary, we develop in section 5 Faria, Huang, and Wu's perturbation method [9] for traveling wavefronts with large wave speeds. We shall provide both theoretical justifications and numerical simulations for this method.

2. The discrete-delay and spatial-averaging case. In the discrete-delay and spatial-averaging case, i.e., the case when the delayed term involves an evaluation of the dependence exactly time τ ago and a convolution in space to account for the movement of individuals to their present positions from their past positions at previous times, (1.1) becomes

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u(t,x)\left(\frac{1 - \int_{-\infty}^{\infty} 1/\sqrt{4\pi\tau}e^{-(x-y)^2/(4\tau)}u(t-\tau,y)dy}{1 + \gamma\int_{-\infty}^{\infty} 1/\sqrt{4\pi\tau}e^{-(x-y)^2/(4\tau)}u(t-\tau,y)dy}\right). \tag{2.1}$$

Our intention here is to establish the existence of traveling waves to (2.1) connecting the two uniform steady states $u = 0$ and $u = 1$. For this purpose, we first show the existence of such wavefronts when the delay τ is zero.

Letting $\tau \to 0^+$, we arrive at the following nondelay version of the food-limited model:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u\frac{1-u}{1+\gamma u}, \tag{2.2}$$

which is actually a modified version of the well-known Fisher's equation. Obviously, (2.2) has two uniform steady-state solutions $u = 0$ and $u = 1$. Considering the traveling wavefront form by setting $u(t, x) = U_0(z) = U_0(x - ct)$ in (2.2), we obtain the following second-order ODE for $U_0(z)$:

$$U_0'' + cU_0' + U_0\frac{1-U_0}{1+\gamma U_0} = 0 \tag{2.3}$$

or equivalently the following first-order coupled system

$$\begin{cases} U_0' = V_0, \\ V_0' = -cV_0 - U_0 \frac{1-U_0}{1+\gamma U_0}. \end{cases} \tag{2.4}$$

The existence of traveling wavefronts to (2.4) can be established by using the standard phase-plane techniques. Here we present only the result below.

THEOREM 2.1. *If $c \geq 2$, then in the (U_0, V_0) phase plane, a heteroclinic connection exists between the critical points $(U_0, V_0) = (1, 0)$ and $(0, 0)$. Furthermore, the traveling front $U_0(z)$ is strictly monotonically decreasing.*

It is easy to see that the equilibrium $(1, 0)$ is a saddle and the origin $(0, 0)$ is a stable node.

To obtain an extension when $\tau > 0$, we need the following estimate on the derivative of the wave profile U_0.

THEOREM 2.2. *Let U_0 be a traveling wavefront solution to* (2.3). *Then*

$$-\frac{1}{2\sqrt{\gamma}} < U_0'(z) \leq 0 \ \textit{ for all } z \in (-\infty, \infty). \tag{2.5}$$

Proof. Since U_0 is strictly monotonically decreasing, it is obvious that $U_0'(z) \leq 0$. It remains to prove that $U_0'(z) > -1/(2\sqrt{\gamma})$. Note that (2.3) can be rewritten as

$$U_0'' + cU_0' \ - \frac{1}{\gamma} U_0 + \left(1 + \frac{1}{\gamma}\right) \frac{U_0}{1 + \gamma U_0} = 0. \tag{2.6}$$

Let

$$\lambda_1 = \frac{-c - \sqrt{c^2 + 4/\gamma}}{2} < 0, \quad \lambda_2 = \frac{-c + \sqrt{c^2 + 4/\gamma}}{2} > 0.$$

Then it follows from (2.6) that

$$U_0(z) = \frac{(1 + \frac{1}{\gamma})}{\lambda_2 - \lambda_1} \left[\int_{-\infty}^{z} e^{\lambda_1 (z-s)} \frac{U_0}{1 + \gamma U_0} ds + \int_{z}^{\infty} e^{\lambda_2 (z-s)} \ \frac{U_0}{1 + \gamma U_0} ds \right],$$

and hence using the fact that $0 < U_0 < 1$, we have

$$\begin{aligned} U_0'(z) &= \frac{(1 + \frac{1}{\gamma})}{\lambda_2 - \lambda_1} \left[\lambda_1 \int_{-\infty}^{z} e^{\lambda_1 (z-s)} \frac{U_0}{1 + \gamma U_0} ds + \lambda_2 \int_{z}^{\infty} e^{\lambda_2 (z-s)} \frac{U_0}{1 + \gamma U_0} ds \right] \\ &> \frac{(1 + \frac{1}{\gamma})\lambda_1}{\lambda_2 - \lambda_1} \int_{-\infty}^{z} e^{\lambda_1 (z-s)} \frac{U_0}{1 + \gamma U_0} ds \\ &\geq \frac{(1 + \frac{1}{\gamma})\lambda_1}{\lambda_2 - \lambda_1} \max \left(\frac{U_0}{1 + \gamma U_0} \right) \int_{-\infty}^{z} e^{\lambda_1 (z-s)} ds \\ &= -\frac{(1 + \frac{1}{\gamma})}{(1 + \gamma)\sqrt{c^2 + 4/\gamma}} \geq -\frac{1}{2\sqrt{\gamma}}. \end{aligned}$$

The proof is complete. $\square$

Now we are in a position to establish the existence of traveling wavefronts to (2.1). We will show that the traveling fronts to (2.1) can be approximated by the corresponding wavefronts $U_0(z)$ of (2.3) when τ is small. First, we introduce some notations. Let $C(R,R)$ be the Banach space of continuous and bounded functions from R to R equipped with the standard norm $||\phi||_C = \sup\{|\phi(t)|, t \in R\}$. Let $C^1 = C^1(R,R) = \{\phi \in C : \phi' \in C\}$, $C^2 = \{\phi \in C : \phi'' \in C\}$, $C_0 = \{\phi \in C : \lim_{t\to\pm\infty} \phi = 0\}$, and $C_0^1 = \{\phi \in C_0 : \phi' \in C_0\}$, where the corresponding norms are defined by

$$||\phi||_{C_0} = ||\phi||_C,\ ||\phi||_{C^1} = ||\phi||_{C_0^1} = ||\phi||_C + ||\phi'||_C,$$

and

$$||\phi||_{C^2} = ||\phi||_C + ||\phi'||_C + ||\phi''||_C.$$

Set $u(t,x) = U(z) = U(x-ct)$ in (2.1). Then $U(z)$ satisfies the profile equation

$$-cU' = U'' + \ U\frac{1-H(U)(z)}{1+\gamma H(U)(z)}, \tag{2.7}$$

where

$$H(U)(z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi\tau}} e^{-y^2/(4\tau)} U(z-y+c\tau)dy.$$

We suppose that U can be approximated by U_0 and hence assume that $U = U_0 + W$. Then an equation for W is given by

$$-cW' = W'' \ + (U_0+W)\frac{1-H(U_0+W)(z)}{1+\gamma H(U_0+W)(z)} - \ U_0\frac{1-U_0(z)}{1+\gamma U_0(z)}\ . \tag{2.8}$$

Applying Taylor's expansions to

$$(U_0+W)\frac{1-H(U_0+W)(z)}{1+\gamma H(U_0+W)(z)}$$

yields

$$\begin{aligned}(U_0+W)\frac{1-H(U_0+W)(z)}{1+\gamma H(U_0+W)(z)} = \ & U_0\frac{1-H(U_0)}{1+\gamma H(U_0)}\\ &+W\frac{1-H(U_0)}{1+\gamma H(U_0)} - \frac{(1+\gamma)U_0}{(1+\gamma H(U_0))^2}H(W)\\ &+R_1(z,\tau,W),\end{aligned} \tag{2.9}$$

where $R_1(z,\tau,W)$ is the remainder (higher order terms) of this expansion, and for the time being we write it as

$$\begin{aligned}R_1(z,\tau,W) = & (U_0+W)\frac{1-H(U_0+W)(z)}{1+\gamma H(U_0+W)(z)}\\ &-U_0\frac{1-H(U_0)}{1+\gamma H(U_0)}\\ &-W\frac{1-H(U_0)}{1+\gamma H(U_0)} + \frac{(1+\gamma)U_0}{(1+\gamma H(U_0))^2}H(W).\end{aligned} \tag{2.10}$$

Let $g(x) = x\frac{1-x}{1+\gamma x}$. Then we have

$$g'(x) = \frac{1-x}{1+\gamma x} - \frac{(1+\gamma)x}{(1+\gamma x)^2}.$$

Therefore, by (2.9), (2.8) becomes

$$\begin{aligned}(2.11)\qquad -cW' &= W'' + g'(U_0(z))W(z)\\ &\quad + R_1(z,\tau,W) + R_2(z,\tau) + R_3(z,\tau,W),\end{aligned}$$

where

$$R_2(z,\tau) = U_0\frac{1-H(U_0)}{1+\gamma H(U_0)} - g(U_0),$$

and

$$R_3(z,\tau,W) = W\frac{1-H(U_0)}{1+\gamma H(U_0)} - \frac{(1+\gamma)U_0}{(1+\gamma H(U_0))^2}H(W) - g'(U_0(z))W(z).$$

Next we transform (2.11) into an integral equation as follows. Recall that the equation

$$W'' + cW' - W = 0$$

has the characteristic equation

$$\lambda^2 + c\lambda - 1 = 0$$

that has two real roots

$$\lambda_1 = \frac{-c-\sqrt{c^2+4}}{2} < 0, \quad \lambda_2 = \frac{-c+\sqrt{c^2+4}}{2} > 0.$$

Thus (2.11) is equivalent to the following integral equation:

$$(2.12)\quad W = \frac{1}{\lambda_2-\lambda_1}\left(\begin{array}{c}\int_{-\infty}^{z} e^{\lambda_1(z-s)}\left[(1+g'(U_0(s)))W(s) + R_1 + R_2 + R_3\right]ds\\ +\int_{z}^{\infty} e^{\lambda_2(z-s)}\left[(1+g'(U_0(s)))W(s) + R_1 + R_2 + R_3\right]ds\end{array}\right).$$

We will study the existence of a solution $W \in C_0$ to (2.12). For this purpose, we define a linear operator $L : C_0 \to C_0$ by

$$\begin{aligned}L(W)(z) = W &- \frac{\int_{-\infty}^{z} e^{\lambda_1(z-s)}(1+g'(U_0(s)))W(s)ds}{\lambda_2-\lambda_1}\\ &- \frac{\int_{z}^{\infty} e^{\lambda_2(z-s)}(1+g'(U_0(s)))W(s)ds}{\lambda_2-\lambda_1}.\end{aligned}$$

It is obvious that $L(W) \in C_0$ if $W \in C_0$. In order to verify the existence of a solution $W \in C_0$ to (2.12), we need to establish some estimates for the terms in the right-hand side of (2.12) when $W \in C_0$. We have the following.

LEMMA 2.3. *For each $\delta > 0$, there is a $\sigma > 0$ such that*

$$(2.13)\qquad ||R_1(z,\tau,\phi) - R_1(z,\tau,\varphi)||_{C_0} \le \delta||\phi-\varphi||_{C_0}$$

and

$$\int_{-\infty}^{z} e^{\lambda_1(z-s)}|R_1(s,\tau,\phi) - R_1(s,\tau,\varphi)|ds + \int_{z}^{\infty} e^{\lambda_2(z-s)}|R_1(s,\tau,\phi) - R_1(s,\tau,\varphi)|ds$$
$$\le \delta||\phi-\varphi||_{C_0}$$

(2.14)

for all finite τ and all $\phi, \varphi \in B(\sigma)$, where $B(\sigma)$ is the ball in C_0 with radius σ and the center at the origin.

Proof. Since R_1 is the remainder of Taylor's expansion and $||H\phi|| \le ||\phi||$ for any $\phi \in C$, we have

$$||R_1(\cdot,\tau,\phi)|| = O(||\phi||^2_{C_0}) \text{ as } ||\phi||_{C_0} \to 0, \tag{2.15}$$

uniformly for all finite τ. Obviously $R_1(\cdot,\tau,\phi)$, $(R_1)_\phi(\cdot,\tau,\phi)$ (the derivative of R_1 with respect to ϕ), and $(R_1)_{\phi\phi}(\cdot,\tau,\phi)$ (the second derivative of R_1 with respect to ϕ) are continuous for ϕ in a neighborhood of the origin in C_0, and $\tau \in [0,\tau_0]$, where τ_0 is a positive number. Therefore, (2.13) and (2.14) follow from (2.15). □

LEMMA 2.4. *As $\tau \to 0$, we have*

$$\left|\int_{-\infty}^{z} e^{\lambda_1(z-s)}R_2(s,\tau)ds + \int_z^\infty e^{\lambda_2(z-s)}R_2(s,\tau)ds\right| = O(\sqrt{\tau})$$

uniformly for all $z \in (-\infty,\infty)$.

Proof. Since

$$\begin{aligned} R_2(z,\tau) &= U_0\frac{1-H(U_0)}{1+\gamma H(U_0)} - g(U_0)\\ &= U_0\frac{1-H(U_0)}{1+\gamma H(U_0)} - U_0\frac{1-U_0}{1+\gamma U_0}, \end{aligned}$$

we need to show only that when τ is small, the following:

$$\int_{-\infty}^{z} e^{\lambda_1(z-s)}\,|H(U_0)(s)-U_0(s)|\,ds = O(\sqrt{\tau}) \tag{2.16}$$

and

$$\int_z^\infty e^{\lambda_1(z-s)}\,|H(U_0)(s)-U_0(s)|\,ds = O(\sqrt{\tau}) \tag{2.17}$$

hold. In fact, we have

$$\begin{aligned} &\int_{-\infty}^{z} e^{\lambda_1(z-s)}\,|H(U_0)(s)-U_0(s)|\,ds\\ =&\int_{-\infty}^{z} e^{\lambda_1(z-s)}\left|\int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi\tau}}e^{-y^2/4\tau}U_0(s-y+c\tau)dy - U_0(s)\right|ds\\ =&\int_{-\infty}^{z} e^{\lambda_1(z-s)}\left|\int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi\tau}}e^{-y^2/4\tau}\left[U_0(s-y+c\tau)-U_0(s)\right]dy\right|ds\\ \le&\int_{-\infty}^{z} e^{\lambda_1(z-s)}\int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi\tau}}e^{-y^2/4\tau}(|y|+c\tau)dyds||U_0'||\\ =&\ O(\sqrt{\tau}). \end{aligned}$$

Similarly we can show (2.17). □

LEMMA 2.5. *There exists an $M_0 > 0$ such that for all $W \in C_0$, the following inequality:*

$$\left|\int_{-\infty}^{z} e^{\lambda_1(z-s)}R_3(s,\tau,W)ds + \int_z^\infty e^{\lambda_2(z-s)}R_3(s,\tau,W)ds\right| \le \sqrt{\tau}M_0||W||_{C_0} \tag{2.18}$$

holds. Furthermore for any two elements ϕ and φ in C_0, we have

$$\left|\int_{-\infty}^{z} e^{\lambda_1(z-s)}(R_3(s,\tau,\phi)-R_3(s,\tau,\varphi))ds+\int_{z}^{\infty} e^{\lambda_2(z-s)}(R_3(s,\tau,\phi)-R_3(s,\tau,\varphi))ds\right|$$
$$=O(\sqrt{\tau})||\phi-\varphi||_{C_0}. \tag{2.19}$$

Proof. We rewrite R_3 as

$$\begin{aligned} R_3(z,\tau,W)= {} & W\left(\frac{1-H(U_0)}{1+\gamma H(U_0)}-\frac{1-U_0}{1+\gamma U_0}\right)\\ & -H(W)\left(\frac{(1+\gamma)U_0}{(1+\gamma H(U_0))^2}-\frac{(1+\gamma)U_0}{(1+\gamma U_0)^2}\right)\\ & -\frac{(1+\gamma)U_0}{(1+\gamma U_0)^2}\left(H(W)-W\right). \end{aligned} \tag{2.20}$$

Therefore, for the integrations of the first and the second lines on the right-hand side of (2.20), we have from (2.16) the following estimates:

$$\left|\int_{-\infty}^{z} e^{\lambda_1(z-s)}W\left(\frac{1-H(U_0)}{1+\gamma H(U_0)}-\frac{1-U_0}{1+\gamma U_0}\right)ds\right|=O(\sqrt{\tau}||W||), \tag{2.21}$$

and

$$\left|\int_{-\infty}^{z} e^{\lambda_1(z-s)}H(W)\left(\frac{(1+\gamma)U_0}{(1+\gamma H(U_0))^2}-\frac{(1+\gamma)U_0}{(1+\gamma U_0)^2}\right)ds\right|=O(\sqrt{\tau}||W||), \tag{2.22}$$

due to the fact $||H(W)||\le||W||$. For the integration of the function in the last line of (2.20), if $W\in C_0^1$, by exchanging the order of integration and integration by parts, we have

$$\begin{aligned} & \int_{-\infty}^{z} e^{\lambda_1(z-s)}\frac{(1+\gamma)U_0}{(1+\gamma U_0)^2}\left(H(W)-W\right)ds\\ ={} & \int_{-\infty}^{z} e^{\lambda_1(z-s)}\;\frac{(1+\gamma)U_0}{(1+\gamma U_0)^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi\tau}}e^{-y^2/(4\tau)}\left(W(s-y+c\tau)-W(s)\right)dyds\\ ={} & \int_{-\infty}^{z} e^{\lambda_1(z-s)}\frac{(1+\gamma)U_0}{(1+\gamma U_0)^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi\tau}}e^{-y^2/(4\tau)}\int_0^{-y+c\tau}W'(s+\eta)d\eta dyds\\ ={} & \int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi\tau}}e^{-y^2/(4\tau)}\int_0^{-y+c\tau}\int_{-\infty}^{z} e^{\lambda_1(z-s)}\frac{(1+\gamma)U_0}{(1+\gamma U_0)^2}W'(s+\eta)dsd\eta dy\\ ={} & \frac{(1+\gamma)U_0(z)}{(1+\gamma U_0(z))^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi\tau}}e^{-y^2/(4\tau)}\int_0^{-y+c\tau}W(z+\eta)d\eta dy\\ & -\int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi\tau}}e^{-y^2/(4\tau)}\int_0^{-y+c\tau}\int_{-\infty}^{z}W(s+\eta)\left[e^{\lambda_1(z-s)}\frac{(1+\gamma)U_0}{(1+\gamma U_0)^2}\right]'dsd\eta dy\\ ={} & O(\sqrt{\tau}||W||). \end{aligned} \tag{2.23}$$

To obtain the above result, we have used the fact that

$$\int_{-\infty}^{z}\left|\frac{d}{ds}\left[e^{\lambda_1(z-s)}\frac{(1+\gamma)U_0(s)}{(1+\gamma U_0(s))^2}\right]\right|ds$$

is uniformly bounded for all $z \in (-\infty, \infty)$. Therefore, from (2.21), (2.22), and (2.23) it follows that there exists a constant M_1 such that

$$\left| \int_{-\infty}^{z} e^{\lambda_1 (z-s)} R_3(s, \tau, W) ds \right| \leq \sqrt{\tau} M_1 ||W||_{C_0}. \tag{2.24}$$

Similarly, we can prove that there exists a constant M_2 so that

$$\left| \int_{z}^{\infty} e^{\lambda_2 (z-s)} R_3(s, \tau, W) ds \right| \leq \sqrt{\tau} M_2 ||W||_{C_0}. \tag{2.25}$$

Therefore, it follows from (2.24) and (2.25) that for any $W \in C_0^1$, we have

$$\left| \int_{-\infty}^{z} e^{\lambda_1 (z-s)} R_3(s, \tau, W) ds + \int_{z}^{\infty} e^{\lambda_2 (z-s)} R_3(s, \tau, W) ds \right| \leq \sqrt{\tau} M_0 ||W||_{C_0},$$

with

$$M_0 = \sum_{j=1}^{2} M_j.$$

Since C_0^1 is dense in C_0, the inequality (2.18) holds for all $W \in C_0$. Thus (2.19) is satisfied due to the fact that $R_3(s, \tau, W)$ is a linear functional of W.

We should mention that for each $\tau > 0$ and $W \in C_0$, we have R_1, R_2, and $R_3 \in C_0$ and hence

$$\int_{-\infty}^{z} e^{\lambda_1 (z-s)} (R_1 + R_2 + R_3) ds + \int_{z}^{\infty} e^{\lambda_2 (z-s)} \left[R_1 + R_2 + R_3 \right] ds \in C_0.$$

Now we are ready to prove our main result. □

THEOREM 2.6. *For any $c \geq 2$, there exists a constant $\delta = \delta(c) > 0$ so that for any $\tau \in [0, \delta]$,* (2.1) *possesses a traveling wavefront $u(t, x) = U(x - ct)$ satisfying $U(-\infty) = 1$ and $U(\infty) = 0$.*

Proof. Define an operator $T : \Psi \in C^2 \to C$ from the homogeneous part of (2.11) as follows:

$$T\Psi(z) = c\Psi'(z) + \Psi''(z) + g'(U_0(z))\Psi(z). \tag{2.26}$$

The formal adjoint equation of $T\Psi = 0$ is given by

$$-c\Phi'(z) + \Phi''(z) + g'(U_0(z))\Phi(z) = 0, \quad z \in R. \tag{2.27}$$

We now divide our proof into five steps.

Step 1. We claim that if $\Phi \in C$ is a solution of (2.27) and Φ is C^2-smooth, then $\Phi = 0$. Moreover, we have $\Re(T) = C$, where $\Re(T)$ is the range space of T.

Indeed, when $z \to \infty$, $U_0(z) \to 0$ and $g'(U_0(z)) \to 1$. Then (2.27) tends asymptotically to an equation with constant coefficients

$$-c\Phi'(z) + \Phi''(z) + \Phi(z) = 0. \tag{2.28}$$

The corresponding characteristic equation of (2.28) is

$$\lambda^2 - c\lambda + 1 = 0. \tag{2.29}$$

Both roots of (2.29) have a positive real part as $c \geq 2$, and thus we can conclude that any bounded solution to (2.28) must be the zero solution. So as $z \to \infty$, any solution to (2.27) other than the zero solution must grow exponentially for large z. Then the only solution satisfying $\Phi(\pm\infty) = 0$ is the zero solution. By the Fredholm theory (see Lemma 4.2 in [22]) we have that $\Re(T) = C$.

Step 2. Let $\Theta \in C_0$ be given. We conclude that if Ψ is a bounded solution of $T\Psi = \Theta$, then we have $\lim_{z\to\pm\infty} \Psi(z) = 0$.

In fact when $z \to \infty$, the equation

$$c\Psi'(z) + \Psi''(z) + g'(U_0(z))\Psi(z) = \Theta \tag{2.30}$$

tends asymptotically to

$$c\Psi'(z) + \Psi''(z) + \Psi(z) = 0. \tag{2.31}$$

Note for (2.31), the ω-limit set of every bounded solution is just the critical point $\Psi = 0$. Using Theorem 1.8 from [19], we find that every bounded solution of (2.30) also satisfies

$$\lim_{z\to\infty} \Psi(z) = 0.$$

When $z \to -\infty$, (2.30) tends asymptotically to

$$c\Psi'(z) + \Psi''(z) + g'(1)\Psi(z) = 0. \tag{2.32}$$

Since $g'(1) = -1$, the characteristic equation of (2.32) has two eigenvalues: $\bar{\lambda}_1 < 0$ and $\bar{\lambda}_2 > 0$. Thus every bounded solution of (2.32) must satisfy

$$\lim_{z\to-\infty} \Psi(z) = 0.$$

Inverting the time from z to $-z$ and using the result in [19] again, we know that any bounded solution to (2.30) satisfies $\lim_{z\to-\infty} \Psi(z) = 0$. Hence the claim of Step 2 holds.

Step 3. For a linear operator $L : C_0 \to C_0$ defined by

$$L(W)(z) = W - \frac{1}{\lambda_2 - \lambda_1} \begin{pmatrix} \int_{-\infty}^{z} e^{\lambda_1(z-s)}(1 + g'(U_0(s)))W(s)ds \\ + \int_{z}^{\infty} e^{\lambda_2(z-s)}(1 + g'(U_0(s)))W(s)ds \end{pmatrix},$$

we want to prove that $\Re(L) = C_0$; that is, for each $Z \in C_0$, we have a $W \in C_0$ so that

$$W - \frac{1}{\lambda_2 - \lambda_1} \begin{pmatrix} \int_{-\infty}^{z} e^{\lambda_1(z-s)}(1 + g'(U_0(s)))W(s)ds \\ + \int_{z}^{\infty} e^{\lambda_2(z-s)}(1 + g'(U_0(s))W(s))ds \end{pmatrix} = Z(z).$$

To see this, we assume that $\xi(z) = W(z) - Z(z)$ and obtain an equation for ξ as follows:

$$\begin{aligned} \xi(z) = {} & \frac{1}{\lambda_2 - \lambda_1} \begin{pmatrix} \int_{-\infty}^{z} e^{\lambda_1(z-s)}(1 + g'(U_0(s)))\xi(s)ds \\ + \int_{z}^{\infty} e^{\lambda_2(z-s)}(1 + g'(U_0(s)))\xi(s)ds \end{pmatrix} \\ & + \frac{1}{\lambda_2 - \lambda_1} \begin{pmatrix} \int_{-\infty}^{z} e^{\lambda_1(z-s)}(1 + g'(U_0(s)))Z(s)ds \\ + \int_{z}^{\infty} e^{\lambda_2(z-s)}(1 + g'(U_0(s)))Z(s)ds \end{pmatrix}. \end{aligned}$$

Differentiating both sides twice yields

$$-c\xi' - \xi''(z) - g'(U_0(z))\xi(z) = (1 + g'(U_0(z)))Z(z). \tag{2.33}$$

Using the results that $\Re(T) = C$ in Step 1 and $Z \in C_0$, we obtain by Step 2 that there exists a solution $\xi(z)$ satisfying (2.33) and $\xi(\pm\infty) = 0$. Returning to the variable W, we have $W = \xi + Z \in C_0$.

Step 4. Let $N(L)$ be the null space of operator L. By Lemma 5.1 in [9], there is a subspace $N^{\perp}(L)$ in C_0 so that

$$C_0 = N^{\perp}(L) \oplus N(L);$$

see also [10]. It is clear that $N^{\perp}(L)$ is a Banach space. If we let $S = L|_{N^{\perp}(L)}$ be the restriction of L to $N^{\perp}(L)$, then $S : N^{\perp}(L) \to C_0$ is one-to-one and onto. By the well-known Banach inverse operator theorem, we have that $S^{-1} : C_0 \to N^{\perp}(L)$ is a linear bounded operator.

Step 5. When L is restricted to $N^{\perp}(L)$, (2.12) can be written as

$$S(W)(z) = \frac{1}{\lambda_2 - \lambda_1} \begin{pmatrix} \int_{-\infty}^{z} e^{\lambda_1(z-s)} \left[R_1 + R_2 + R_3\right] ds \\ + \int_{z}^{\infty} e^{\lambda_2(z-s)} \left[R_1 + R_2 + R_3\right] ds \end{pmatrix}.$$

Since the norm $||S^{-1}||$ is independent of τ, it follows from Lemmas 2.3, 2.4, and 2.5 that there exist $\sigma > 0, \delta > 0$, and $0 < \rho < 1$ such that for all $\tau \in (0, \delta]$ and $W, \varphi, \psi \in B(\sigma) \cap N^{\perp}(L)$,

$$||F(z, W)\ || \le \frac{1}{3}\left(||W|| + \sigma\right)$$

and

$$||F(z, \varphi) - F(z, \psi)||\ \le \rho||\varphi - \psi||,$$

where

$$F(z, W) = \frac{1}{\lambda_2 - \lambda_1} S^{-1} \begin{pmatrix} \int_{-\infty}^{z} e^{\lambda_1(z-s)} \left[R_1 + R_2 + R_3(\tau, s, W)\right] ds \\ + \int_{z}^{\infty} e^{\lambda_2(z-s)} \left[R_1 + R_2 + R_3(\tau, s, W)\right] ds \end{pmatrix}.$$

It is easy to know that for any $W \in B(\sigma) \cap N^{\perp}(L)$, we have

$$||F(z, W)\ || \le \frac{1}{3}\left(||W|| + \sigma\right) \le \sigma.$$

Hence $F(z, \varphi)$ is a uniform contractive mapping for $W \in N^{\perp}(L) \cap B(\sigma)$. By using the Banach contraction principle, it follows that for $\tau \in [0, \delta]$, equation (2.12) has a unique solution $W \in N^{\perp}(L)$. Returning to the original variable, $W + U_0$ is a heteroclinic connection between the two equilibria 1 and 0. This completes the proof. $\square$

3. The distributed delay case. In this section we consider (1.1) with the kernel function

$$f(t, s, x, y) = G(t - s)\delta(x - y),$$

where

$$G(t) = \frac{1}{\tau} e^{-t/\tau} \quad \text{or} \quad G(t) = \frac{t}{\tau^2} e^{-t/\tau}. \tag{3.1}$$

We shall focus on the following equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u \frac{1 - \int_{-\infty}^{t} \frac{t-\eta}{\tau^2} e^{-(t-\eta)/\tau} u(\eta, x) d\eta}{1 + \gamma \int_{-\infty}^{t} \frac{t-\eta}{\tau^2} e^{-(t-\eta)/\tau} u(\eta, x) d\eta} \tag{3.2}$$

since the corresponding analysis for the weak kernel $G(t) = \frac{1}{\tau} e^{-t/\tau}$ is much easier. Instead of using the linear chain trick which is valid only for the kernels in (3.1), we shall use the approach in section 2 to prove rigorously that traveling fronts exist when τ is small.

As in section 2, assume that $u(t, x) = U(z)$, $z = x - ct$, where $c \geq 2$. Substituting $u = U(z)$ into (3.2), we have a wave equation for U

$$-cU' = U'' + U \frac{1 - \int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} U(z + c\eta) d\eta}{1 + \gamma \int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} U(z + c\eta) d\eta}.$$

Let $U = U_0 + W$, where U_0 is the traveling fronts for (2.4). Then we have an equation for W of the form

$$-cW' = W'' - W + (U_0 + W) h[U_0 + W] - U_0 \frac{1 - U_0}{1 + \gamma U_0}, \tag{3.3}$$

where the functional h is defined by

$$h[U](z) = \frac{1 - \int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} U(z + c\eta) d\eta}{1 + \gamma \int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} U(z + c\eta) d\eta}.$$

Applying Taylor's expansion to $h[U_0 + W]$, we have

$$\begin{aligned} h[U_0 + W](z) &= \frac{1 - \int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} [U_0(z + c\eta) + W(z + c\eta] dy d\eta}{1 + \gamma \int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} [U_0(z + c\eta) + W(z + c\eta] dy d\eta} \\ &= h[U_0](z) - \frac{(1 + \gamma)}{\left(1 + \gamma \int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} U_0(z + c\eta) d\eta\right)^2} \int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} W(z + c\eta) d\eta + \cdots . \end{aligned}$$

Thus (3.3) becomes

$$\begin{aligned} -cW' &= W'' + g'(U_0(z)) W(z) \\ &\quad + R_1(z, \tau, W) + R_2(z, \tau) + R_3(z, \tau, W), \end{aligned} \tag{3.4}$$

where

$$\begin{aligned} R_1(z, \tau, W) = {} & (U_0 + W) h[U_0 + W](z) - U_0 h[U_0](z) \\ & + U_0 \frac{(1 + \gamma)}{\left(1 + \gamma \int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} U_0(z + c\eta) d\eta\right)^2} \int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} W(z + c\eta) d\eta \\ & - h[U_0] W, \end{aligned} \tag{3.5}$$

$$R_2(z, \tau) = U_0 h[U_0](z) - U_0 \frac{1 - U_0}{1 + \gamma U_0}, \tag{3.6}$$

and

$$R_3(z,\tau,W) = \frac{(1+\gamma)\ U_0 \int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} W(z+c\eta)d\eta}{\left(1+\gamma\int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} U_0(z+c\eta)d\eta\right)^2} \tag{3.7}$$
$$+h[U_0]W(z) - g'(U_0(z))W(z).$$

To obtain the existence of the traveling fronts when τ is small, we need to prove that Lemmas 2.3, 2.4, and 2.5 hold. The proofs are quite similar to those in the discrete case, so we shall prove only Lemma 2.5 as an illustration and leave the proofs of Lemmas 2.3 and 2.4 to interested readers.

Proof of Lemma 2.5 *in the case of distributed delay.* Note that

$$g'(U_0(z)) = \frac{1-U_0}{1+\gamma U_0} - \frac{(1+\gamma)U_0}{(1+\gamma U_0)^2}.$$

So

$$\begin{aligned} R_3(z,\tau,W) = {} & \bar{F}(U_0)(z)\left(\int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} W(z+c\eta)d\eta - W(z)\right) \\ & - \left(\frac{(1+\gamma)\ U_0}{\left(1+\gamma\int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} U_0(z+c\eta)d\eta\right)^2} - \frac{(1+\gamma)U_0}{(1+\gamma U_0)^2}\right) W \\ & + \left(h[U_0] - \frac{1-U_0}{1+\gamma U_0}\right) W, \end{aligned}$$

where

$$\bar{F}(U_0)(z) = -\frac{(1+\gamma)U_0}{\left(1+\gamma\int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} U_0(z+c\eta)d\eta\right)^2}.$$

We note that when τ is small,

$$\int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} U_0(z+c\eta)d\eta - U_0(z) = O(\tau)$$

holds uniformly for any $z \in (-\infty, \infty)$. Therefore, we have

$$\begin{aligned} & \int_{-\infty}^z e^{\lambda_1(z-s)} \left(\frac{(1+\gamma)\ U_0}{\left(1+\gamma\int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} U_0(s+c\eta)d\eta\right)^2} - \frac{(1+\gamma)U_0}{(1+\gamma U_0)^2}\right) W(s)ds \\ & = O(\tau\|W\|) \end{aligned}$$

and

$$\int_{-\infty}^z e^{\lambda_1(z-s)} \left(h[U_0] - \frac{1-U_0}{1+\gamma U_0}\right) W(s)ds = O(\tau\|W\|).$$

We now prove that

$$\int_{-\infty}^z e^{\lambda_1(z-s)} \bar{F}(U_0)(s) \left(\int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} \left(W(s+c\eta) - W(s)\right) d\eta\right) ds = O(\tau\|W\|).$$

Using the fact that

$$\bar{F}(U_0)(s) = O(1)$$

for any $s \in (-\infty, \infty)$, we need only prove that

$$\int_{-\infty}^{z} e^{\lambda_1(z-s)} \left(\int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} \left(W(s+c\eta) - W(s) \right) d\eta \right) ds = O(\tau ||W||).$$

Indeed, when $W \in C_0^1$, we have

$$W(s+c\eta) - W(s) = \int_0^{c\eta} W'(s+v) dv.$$

Exchanging the order of integration and using the integration by parts, we obtain

$$\begin{aligned} &\int_{-\infty}^{z} e^{\lambda_1(z-s)} \left(\int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} \left(W(s+c\eta) - W(s) \right) d\eta \right) ds \\ &= \int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} \left(\int_0^{c\eta} \int_{-\infty}^{z} e^{\lambda_1(z-s)} W'(s+v) ds dv \right) d\eta \\ &= O(\tau ||W||_{C_0}). \end{aligned}$$

Continuing the process as in section 2, we can show that Lemma 2.5 remains true, and so does the result in Theorem 2.6 for (3.2).

Similarly, we can prove that Theorem 2.6 is true if the kernel function is replaced by

$$f(t,s,x,y) = \frac{1}{\tau} e^{-(t-s)} \delta(x-y). \qquad \square$$

4. The distributed delay and spatial-averaging case. In this section, we consider (1.1) with the distributed delay and spatial averaging. Namely, we study the following equation:

(4.1)
$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u(t,x) \left(\frac{1 - \int_{-\infty}^0 \frac{t-\eta}{\tau^2} e^{-(t-\eta)/\tau} \int_{-\infty}^\infty \frac{\exp(-(x-y)^2/(4(t-\eta)))}{\sqrt{4\pi(t-\eta)}} u(\eta,y) dy d\eta}{1 + \int_{-\infty}^0 \frac{t-\eta}{\tau^2} e^{-(t-\eta)/\tau} \int_{-\infty}^\infty \frac{\exp(-(x-y)^2/(4(t-\eta)))}{\sqrt{4\pi(t-\eta)}} u(\eta,y) dy d\eta} \right).$$

As before, by a traveling wavefront, we mean a solution $u(t,x) = U(z) = U(-ct+x)$, where $c > 0$ is the wave speed. Thus this specific kind of solution satisfies the following second-order ODE:

$$-cU' = U'' + U \frac{1 - H_1(U)(z)}{1 + \gamma H_1(U)(z)}, \tag{4.2}$$

where

$$H_1(U)(z) = \int_0^\infty \frac{\eta}{\tau^2} e^{-\eta/\tau} \int_{-\infty}^\infty \frac{1}{\sqrt{4\pi\eta}} e^{-y^2/4\eta} U(z-y+c\eta) dy d\eta.$$

We suppose that U can be approximated by U_0 and hence assume that $U = U_0 + W$. Then we obtain the equation for W as follows:

$$-cW' = W'' + (U_0 + W) \frac{1 - H_1(U_0+W)(z)}{1 + \gamma H_1(U_0+W)(z)} - U_0 \frac{1 - U_0(z)}{1 + \gamma U_0(z)}. \tag{4.3}$$

Applying Taylor's expansions to

$$(U_0+W)\frac{1-H_1(U_0+W)(z)}{1+\gamma H_1(U_0+W)(z)},$$

we have

$$\begin{aligned}(U_0+W)\frac{1-H_1(U_0+W)(z)}{1+\gamma H_1(U_0+W)(z)} = {} & U_0\frac{1-H_1(U_0)}{1+\gamma H_1(U_0)}\\ & +W\frac{1-H_1(U_0)}{1+\gamma H_1(U_0)}-\frac{(1+\gamma)\ U_0}{(1+\gamma H_1(U_0))^2}H_1(W)\\ & +R_1(z,\tau,W),\end{aligned} \tag{4.4}$$

where $R_1(z,\tau,W)$ is the remainder (higher order terms) of this expansion, and this can be rewritten as

$$\begin{aligned}R_1(z,\tau,W) = {} & (U_0+W)\frac{1-H_1(U_0+W)(z)}{1+\gamma H_1(U_0+W)(z)}\\ & -U_0\frac{1-H_1(U_0)}{1+\gamma H_1(U_0)}\\ & -W\frac{1-H_1(U_0)}{1+\gamma H_1(U_0)}+\frac{(1+\gamma)U_0}{(1+\gamma H_1(U_0))^2}H_1(W).\end{aligned} \tag{4.5}$$

Recall that

$$g(x)=x\frac{1-x}{1+\gamma x}\ \text{ and } g'(x)=\frac{1-x}{1+\gamma x}-\frac{(1+\gamma)x}{(1+\gamma x)^2}.$$

Therefore, in view of (4.4), (4.3) becomes

$$\begin{aligned}-cW' = {} & W''+g'(U_0(z))W(z)\\ & +\ R_1(z,\tau,W)+R_2(z,\tau)+R_3(z,\tau,W),\end{aligned} \tag{4.6}$$

where

$$R_2(z,\tau)=U_0\frac{1-H_1(U_0)}{1+\gamma H_1(U_0)}-\ g(U_0)$$

and

$$R_3(z,\tau,W)=\ W\frac{1-H_1(U_0)}{1+\gamma H_1(U_0)}-\frac{(1+\gamma)U_0}{(1+\gamma H_1(U_0))^2}H_1(W)-g'(U_0(z))W(z).$$

As before, we transform (4.6) into the following integral equation

$$W=\frac{1}{\lambda_2-\lambda_1}\begin{pmatrix}\int_{-\infty}^{z}e^{\lambda_1(z-s)}\left[(1+g'(U_0(s)))W(s)+R_1+R_2+R_3\right]ds\\ +\int_{z}^{\infty}e^{\lambda_2(z-s)}\left[(1+g'(U_0(s)))W(s)+R_1+R_2+R_3\right]ds\end{pmatrix}. \tag{4.7}$$

Now the above argument can be repeated to show that Theorem 2.6 holds for (4.1).

Similarly, we can prove Theorem 2.6 for (1.1) with the kernel function

$$f(t,s,x,y)=\frac{1}{\tau}e^{-(t-s)}\frac{1}{\sqrt{4\pi(t-s)}}e^{-(x-y)^2/(4(t-s))}. \tag{4.8}$$

5. Traveling wavefronts with large wave speed. In section 2, we obtained a traveling wavefront for our model by assuming that the maturation time τ is small. In this section we utilize the idea of Canosa [6] to investigate the existence of traveling wavefront for (2.1) without the smallness requirement of τ. Although the method is originally a formal asymptotic analysis as the front speed approaches infinity, it is known that for Fisher's equation the method generates a solution that is accurate within a few percent of the true solution, even at the minimum speed. The method has also been applied to other reaction-diffusion equations, including coupled systems, with a very good accuracy; see [21] and [24]. The main purpose here is to give a theoretical justification of the method for our food-limited model by showing the fact that when the wave speed tends to infinity, our traveling wavefront approaches a heteroclinic solution (the leading term of Canosa's expansions) of the original model without diffusion. The main idea of this section is from [9], except that we use some known results of global stability of the positive equilibrium instead of applying Smith and Thieme's order preserving semiflows theory [26].

Linearizing (2.7) for U far ahead of the front, where $U \to 0$, gives

$$-cU'(z) = U''(z) - U(z).$$

To ensure that we are studying ecologically realistic fronts that are positive for all values of z, we assume, as in Fisher's equation, that the wave speed $c \geq 2$. Following Canosa's approach, we introduce the small parameter

$$\varepsilon = 1/c^2 \leq \frac{1}{4}$$

and seek a solution of the form

$$U(z) = G(\zeta), \quad \zeta = \sqrt{\varepsilon} z.$$

Equation (2.7) becomes

$$\varepsilon G'' + G' + G \frac{1 - \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi\tau}} e^{-y^2/4\tau} G(\zeta - \sqrt{\varepsilon} y + \tau) dy}{1 + \gamma \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi\tau}} e^{-y^2/4\tau} G(\zeta - \sqrt{\varepsilon} y + \tau) dy} = 0. \tag{5.1}$$

When $\varepsilon = 0$, (5.1) reduces to

$$G' + G \frac{1 - G(\zeta + \tau)}{1 + \gamma G(\zeta + \tau)} = 0. \tag{5.2}$$

For (5.2), we have the following result concerning the heteroclinic orbit connecting the two equilibria $G = 0$ and $G = 1$.

THEOREM 5.1. *Assume $\tau/(1+\gamma) < \frac{3}{2}$. Then* (5.2) *has a heteroclinic orbit $g_0(\zeta)$ connecting the two equilibria $G = 1$ and $G = 0$.*

Proof. When $\varepsilon = 0$, we set $g_0(\zeta) = G(-\zeta)$ to invert (5.2) into a delay differential equation

$$g_0' = g_0 \frac{1 - g_0(\zeta - \tau)}{1 + \gamma g_0(\zeta - \tau)}. \tag{5.3}$$

By the result in [18] or [8], we know that the equilibrium $g_0 = 1$ is a global attractor as long as the initial value $g_0(s) = \phi(s)$, $s \in [-\tau, 0]$, satisfies

$$\phi(0) > 0 \text{ and } \phi(s) \geq 0 \text{ for s} \in [-\tau, 0].$$

Linearizing (5.3) around $g_0 = 0$, we have

$$g_0' = g_0. \tag{5.4}$$

Therefore, the unstable space E_u of the trivial solution in the usual phase space $C_\tau = C([-\tau, 0]; R)$ of continuous functions equipped with the sup-norm $||\cdot||$ is spanned by $\chi(s) = e^s, s \in [-\tau, 0]$. Let E_s be the subspace in C_τ so that $C_\tau = E_s \oplus E_u$; then there exists $\epsilon_0 > 0$ and a C^1-map $w : E_u \to E_s$, with $w(0) = 0$ and $Dw(0) = 0$ so that a local unstable manifold of $g_0 = 0$ is given by $\epsilon\chi + w(\epsilon\chi)$ for $\epsilon \in (-\epsilon_0, \epsilon_0)$. Choose $\epsilon_0 > 0$ sufficiently small so that the operator norm $||Dw(\epsilon\chi)|| < e^{-\tau}$ for $\epsilon \in (0, \epsilon_0)$. Then pick up $\epsilon \in (0, \epsilon_0)$ and consider $\phi = \epsilon\chi + w(\epsilon\chi)$. We have

$$\phi(s) = \epsilon e^s + w(\epsilon\chi)(s) > \epsilon e^s - e^{-\tau}\epsilon||\chi|| \geq \epsilon(e^{-\tau} - e^{-\tau}) \geq 0.$$

So the solution from the point ϕ on the local unstable manifold of $g_0 = 0$ is positive and tends to 1 due to the global attractivity of the positive equilibrium $g_0 = 1$. Returning to the original variable, we have an orbit G connecting the two equilibria $G = 1$ and $G = 0$. This completes the proof. ∎

For (5.3), the positive equilibrium 1 is a node, and all of the conditions in Theorem 1.1 in [9] are satisfied. Thus direct application of this result gives the following.

THEOREM 5.2. *Assume $\tau/(1+\gamma) < \frac{3}{2}$. There is a constant $c^* > 0$ such that for any $c > c^*$,* (5.1) *has a traveling wave solution $G(x - ct)$ connecting the two equilibria* 0 *and* 1. *When the wave speed $c \to \infty$, the wave profile $G(\xi)$ converges to a solution to* (5.2).

Although the result is a consequence of Theorem 1.1 in [9], for the completeness of this paper and the convenience of readers, we outline the proof of this theorem as follows.

For (5.1), set $\bar{G}(\zeta) = G(-\zeta)$. Then $\bar{G}$ satisfies the equation

$$\varepsilon\bar{G}'' - \bar{G}' + \bar{G}\frac{1 - \int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi\tau}}e^{-y^2/4\tau}\bar{G}(\zeta + \sqrt{\varepsilon}y - \tau)dy}{1 + \gamma\int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi\tau}}e^{-y^2/4\tau}\bar{G}(\zeta + \sqrt{\varepsilon}y - \tau)dy} = 0.$$

Now when ε is small, we use g_0 to approximate the wavefront $\bar{G}(\zeta)$ in (5.1). Let $\bar{G} = g_0 + W$. Then we have an equation for W

$$W' = \varepsilon W'' + \varepsilon g_0'' + (g_0 + W)\frac{1 - h_1(g_0 + W)}{1 + \gamma h_1(g_0 + W)} - g_0\frac{1 - g_0(\zeta - \tau)}{1 + \gamma g_0(\zeta - \tau)}, \tag{5.5}$$

where the functional h_1 is given by

$$h_1[U](\zeta) = \int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi\tau}}e^{-\eta^2/4\tau}U(\zeta + \sqrt{\varepsilon}\eta - \tau)d\eta. \tag{5.6}$$

By means of Taylor's expansion, we have

$$\begin{aligned}(g_0 + W)\frac{1 - h_1(g_0 + W)(\zeta)}{1 + \gamma h_1(g_0 + W)(\zeta)} = \;& g_0\frac{1 - h_1(g_0)}{1 + \gamma h_1(g_0)} \\ & + W\frac{1 - h_1(g_0)}{1 + \gamma h_1(g_0)} - \frac{(1+\gamma)g_0}{(1 + \gamma h_1(g_0))^2}h_1(W) \\ & + R_1(\zeta, \tau, W),\end{aligned} \tag{5.7}$$

where $R_1(\zeta,\tau,W)$ is the remainder (higher order terms) of this expansion. Therefore by (5.7), (5.5) becomes

$$W' = \varepsilon W'' + P^0 W(z) + R_1(\zeta,\tau,W) + R_2(\zeta,\tau) + R_3(\zeta,\tau,W), \tag{5.8}$$

where the linear operator $P^0 : C \to C$ is defined by

$$\begin{aligned}
P^0 W(\zeta) &= \frac{1-g_0(\zeta-\tau)}{1+\gamma g_0(\zeta-\tau)} W(\zeta) - g_0 \frac{(1+\gamma)}{(1+\gamma g_0(\zeta-\tau))^2} W(\zeta-\tau),\\
R_2(\zeta,\tau) &= g_0 \frac{1-h_1(g_0)}{1+\gamma h_1(g_0)} - g_0 \frac{1-g_0(\zeta-\tau)}{1+\gamma g_0(\zeta-\tau)} + \varepsilon g_0'',
\end{aligned}$$

and

$$\begin{aligned}
R_3(\zeta,\tau,W) = {} & W \frac{1-h_1(g_0)}{1+\gamma h_1(g_0)} - g_0 \frac{(1+\gamma)}{(1+\gamma h_1(g_0))^2} h_1(W)\\
& - W \frac{1-g_0(\zeta-\tau)}{1+\gamma g_0(\zeta-\tau)} + g_0 \frac{(1+\gamma)}{(1+\gamma g_0(\zeta-\tau))^2} h_1(W).
\end{aligned}$$

Now we prove that there exists a $W \in C_0$ satisfying (5.8) when ε is small. Equation (5.8) can be transformed into an integral equation as follows. We first write (5.8) as

$$\varepsilon W'' - W' - W = -W - P^0 W - R_1 - R_2 - R_3. \tag{5.9}$$

Since the equation

$$\varepsilon \lambda^2 - \lambda - 1 = 0$$

has two real zeros λ_1 and λ_2, with

$$\lambda_1 = \frac{1-\sqrt{1+4\varepsilon}}{2\varepsilon} < 0, \quad \lambda_2 = \frac{1+\sqrt{1+4\varepsilon}}{2\varepsilon} > 0, \tag{5.10}$$

(5.9) is equivalent to the integral equation

$$\begin{aligned}
W(\zeta) = {} & \frac{1}{\sqrt{1+4\varepsilon}} \int_{-\infty}^{\zeta} e^{\lambda_1(\zeta-t)} [W(t) + P^0 W(t)] dt\\
& + \frac{1}{\sqrt{1+4\varepsilon}} \int_{\zeta}^{\infty} e^{\lambda_2(\zeta-t)} [W(t) + P^0 W(t)] dt\\
& + \frac{1}{\sqrt{1+4\varepsilon}} \int_{-\infty}^{\zeta} e^{\lambda_1(\zeta-t)} [R_1 + R_2 + R_3] dt\\
& + \frac{1}{\sqrt{1+4\varepsilon}} \int_{\zeta}^{\infty} e^{\lambda_2(\zeta-t)} [R_1 + R_2 + R_3] dt.
\end{aligned} \tag{5.11}$$

It is easy to show that

$$\lim_{\varepsilon \to 0^+} \lambda_1 = -1, \quad \lim_{\varepsilon \to 0^+} \lambda_2 = +\infty.$$

Thus we have, from (5.11), that

$$
\begin{aligned}
&W(\zeta) - \int_{-\infty}^{\zeta} e^{-(\zeta-t)}[W(t) + P^0W(t)]dt \\
&= \int_{-\infty}^{\zeta} \left[\frac{e^{\lambda_1(\zeta-t)}}{\sqrt{1+4\varepsilon}} - e^{-(\zeta-t)}\right][W(t) + P^0W(t)]dt \\
&\quad + \frac{1}{\sqrt{1+4\varepsilon}} \int_{\zeta}^{\infty} e^{\lambda_2(\zeta-t)}[W(t) + P^0W(t)]dt \\
&\quad + \frac{1}{\sqrt{1+4\varepsilon}} \int_{-\infty}^{\zeta} e^{\lambda_1(\zeta-t)}[R_1 + R_2 + R_3\ ]dt \\
&\quad + \frac{1}{\sqrt{1+4\varepsilon}} \int_{\zeta}^{\infty} e^{\lambda_2(\zeta-t)}[R_1 + R_2 + R_3]dt. \qquad (5.12)
\end{aligned}
$$

For the right-hand side of (5.12), in a similar manner as in section 2, we can prove that

$$\int_{-\infty}^{\zeta} \left[\frac{e^{\lambda_1(\zeta-t)}}{\sqrt{1+4\varepsilon}} - e^{-(\zeta-t)}\right][W(t) + P^0W(t)]dt = O(\sqrt{\varepsilon}||W||_{C_0}),$$

$$\frac{1}{\sqrt{1+4\varepsilon}} \int_{\zeta}^{\infty} e^{\lambda_2(\zeta-t)}[W(t) + P^0W(t)]dt = O(\sqrt{\varepsilon}||W||_{C_0}),$$

$$\frac{1}{\sqrt{1+4\varepsilon}} \left(\int_{-\infty}^{\zeta} e^{\lambda_1(\zeta-t)}R_1\ dt + \int_{\zeta}^{\infty} e^{\lambda_2(\zeta-t)}R_1 dt\right) = O(||W||^2),$$

$$\frac{1}{\sqrt{1+4\varepsilon}} \left(\int_{-\infty}^{\zeta} e^{\lambda_1(\zeta-t)}R_2\ dt + \int_{\zeta}^{\infty} e^{\lambda_2(\zeta-t)}R_2 dt\right) = O(\sqrt{\varepsilon}),$$

and

$$\frac{1}{\sqrt{1+4\varepsilon}} \left(\int_{-\infty}^{\zeta} e^{\lambda_1(\zeta-t)}R_3\ dt + \int_{\zeta}^{\infty} e^{\lambda_2(\zeta-t)}R_3 dt\right) = O(\sqrt{\varepsilon}||W||_{C_0}).$$

Let L be the linear operator defined by the left-hand side of (5.12), namely,

$$[LW](\zeta) = W(\zeta) - \int_{-\infty}^{\zeta} e^{-(\zeta-t)}[W(t) + P^0W(t)]dt.$$

It is obvious that if $W \in C_0$, then $LW \in C_0$. In order to use the argument in section 2 to prove our result, we need to prove that $\Re(L) = C_0$, where $\Re(L)$ is the range space of L; that is, for each $u \in C_0$, we need to show that equation $LW = u$ or, equivalently,

$$W(\zeta) - \int_{-\infty}^{\zeta} e^{-(\zeta-t)}[W(t) + P^0W(t)]dt = u(\zeta), \quad \zeta \in (-\infty, \infty)$$

has a solution in C_0. For this purpose, we set $w = W - u$. Upon substitution, we have an equation for w :

$$w' = P^0w(\zeta) + u(\zeta) + P^0u(\zeta). \tag{5.13}$$

Define an operator $T : C_0^1 \to C_0$ by

$$[Tw](\zeta) = w'(\zeta) - P^0w(\zeta)$$

and the formal adjoint equation of $Tw = 0$ by

$$\phi'(t) = -\frac{1 - g_0(t-\tau)}{1+\gamma g_0(t-\tau)}\phi(t) + \frac{g_0(1+\gamma)}{(1+\gamma g_0(t-\tau))^2}\phi\,(t+\tau)\,,\ t \in (-\infty, \infty). \tag{5.14}$$

When $t \to \infty$, (5.14) tends asymptotically to

$$\phi'(t) = \frac{1}{1+\gamma}\phi(t+\tau).$$

When $\tau/(1+\gamma) < \frac{\pi}{2}$, it is easy to see that if ϕ is a bounded solution to (5.14), then $\phi = 0$. From p. 7 of Chow, Lin, and Mallet-Paret [5], we see that T is Fredholm and $\Re(T) = C_0$. Therefore, (5.13) has a solution $w \in C_0$. From now on we can use the same argument as in sections 4 and 5 in [9] to verify that (5.12) has a solution $W \in C_0$.

6. Summary and simulations. In this paper we have studied the existence of traveling wavefronts for the food-limited population model that involves nonmonotone delayed nonlocal response. The classical phase-plane approach or super/subsolution technique does not work for this type of model due to the lack of monotonicity. Hence, we develop a perturbation argument based on some analytical tools such as the contraction mapping principle and the Fredholm theory to establish the existence of traveling wavefronts. We consider three cases with spatiotemporal averaging when the delay is small. In the general case where the smallness condition on the delay is no longer required, we also developed Canosa's method to establish traveling wavefronts with large wave speeds.

Our work shows how our perturbation analyses based on some analytical tools are particularly useful for models with small delay or for wavefronts with large wave speeds. We believe the smallness condition on τ and the largeness condition on the wave speed can be removed by a certain homotopy argument, and this remains to be a subject for future study.

We should emphasize the difficulty caused by the nonmonotonicity of the delayed nonlocal response. In particular, we note that the traveling wavefronts obtained have prominent humps as the following two numerical simulations show.

The first numerical simulation, reported in Figure 1, is for (2.1), carried out by using Matlab on a spatial domain $-L_0 \le x \le L_0$ (for some $L_0 > 0$) with homogeneous Neumann boundary conditions at both ends. For initial data, we set a nonzero steady state value 1 at the left side and zero elsewhere for all $t \in [-\tau, 0]$. The solution stabilizes to a wavefront when time t goes on. For τ sufficiently small, the resulting traveling fronts appear to be strictly monotone. Increasing the value τ, we find that the monotonicity is lost and a prominent hump is exhibited. When $\gamma = 1$ and $\tau = 2$, the solutions at two different times are shown in Figure 1.

The second numerical simulation is about (4.1) with the kernel given by (4.8). Set

$$v(t,x) = \int_{-\infty}^{t} \frac{1}{\tau}e^{-(t-s)} \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi(t-s)}} e^{-(x-y)^2/(4(t-s))} u(s,y)dyds.$$

Then it is easy to recast (1.1) into the form

$$\begin{cases} u_t = u_{xx} + u\frac{1-v}{1+\gamma v} \\ v_t = v_{xx} + \frac{1}{\tau}(u-v). \end{cases}$$

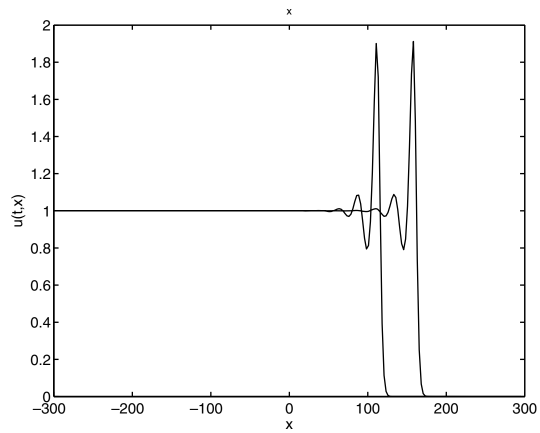


FIG. 1. $\gamma = 1, \tau = 2$. *There exists a prominent hump in the front.*

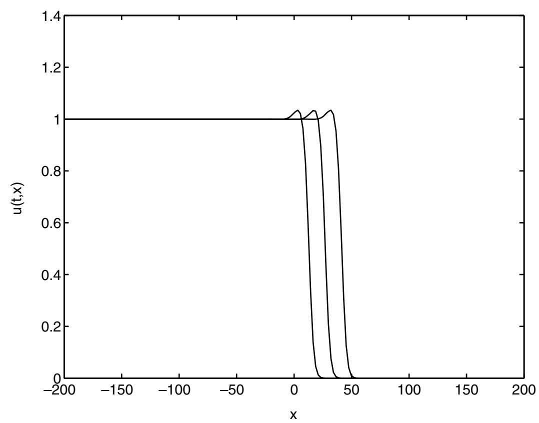


FIG. 2. $\gamma = 1, \tau = 1$. *There exists a prominent hump in the front.*

Using the method of lines, we find again the solution to the above equations with step initial functions and the Neumann boundary conditions stabilizes to a wavefront with a hump. The solution pattern at three different times with $\gamma = 1, \tau = 1$ are shown in Figure 2.

Acknowledgment. We thank the reviewers whose constructive comments lead to a significant improvement of the manuscript.

REFERENCES

[1] J. F. M. AL-OMARI AND S. A. GOURLEY, *Stability and traveling fronts in Lotka-Volterra competition models with stage structure*, SIAM J. Appl. Math., 63 (2003), pp. 2063–2086.
[2] L. ARNOLD, C. JONES, K. MISCHAIKOW, AND G. RAUGEL, *Dynamical systems*, Lectures given at the Second C.I.M.E. Session held in Montecatini Terme, 1994, R. Johnson, ed., Lect. Notes Math. 1609, Springer-Verlag, Berlin, 1995.
[3] P. W. BATES, P. C. FIFE, X. REN, AND X. WANG, *Traveling waves in a convolution model for phase transitions*, Arch. Ration. Mech. Anal., 138 (1997), pp. 105–136.
[4] N. F. BRITTON, *Spatial structures and periodic travelling waves in an integro-differential reaction-diffusion population model*, SIAM J. Appl. Math., 50 (1990), pp. 1663–1688.

[5] S.-N. Chow, X. -B. Lin, and J. Mallet-Paret, *Transition layers for singularly perturbed delay differential equations with monotone nonlinearities*, J. Dynam. Differential Equations, 1 (1989), pp. 3–43.
[6] J. Canosa, *On a nonlinear diffusion equation describing population growth*, IBM J. Res. Develop., 17 (1973), pp. 307–313.
[7] X. Chen, *Existence, uniqueness, and asymptotic stability of traveling waves in nonlocal evolution equations*, Adv. Differential Equations, 2 (1997), pp. 125–160.
[8] T. Faria, E. Liz, J. Oliveira, and S. Trofimchuk, *On a generalized Yorke condition for scalar delayed population models*, Discrete Contin. Dyn. Syst., 12 (2005), pp. 481–500.
[9] T. Faria, W. Huang, and J. Wu, *Traveling waves for delayed reaction-diffusion equations with global response*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 462 (2006), pp. 229–261.
[10] T. Faria and S.Trofimchuk, *Non-monotone traveling waves in a single species reaction-diffusion equation with delay*, J. Differential Equations, 228 (2006), pp. 357–376.
[11] W. Feng and X. Lu, *On diffusive population models with toxicants and time delays*, J. Math. Anal. Appl., 233 (1999), pp. 373–386.
[12] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.
[13] K. Gopalsamy, M. S. C. Kulenovic, and G. Ladas, *Time lags in a "food-limited" population model*, Appl. Anal., 31 (1988), pp. 225–237.
[14] S. A. Gourley and M. A. J. Chaplain, *Travelling fronts in a food-limited population model with time delay*, Proc. Roy. Soc. Edinburgh Sect. A, 132 (2002), pp. 75–89.
[15] S. A. Gourley and S. Ruan, *Convergence and travelling fronts in functional differential equations with nonlocal terms: a competition model*, SIAM J. Math. Anal., 35 (2003), pp. 806–822.
[16] T. G. Hallam, R. R. Lassiter, and S. A. L. M. Kooijman, *Effects of toxicants on aquatic populations*, in Applied Mathematical Ecology (Trieste, 1986), Biomathematics 18, Springer-Verlag, Berlin, 1989, pp. 352–382, 92A17.
[17] E. Liz, V. Tkachenko, and S. Trofimchuk, *York and Wright* 3/2*-stability theorems from a unified point of view*, Discrete Contin. Dyn. Syst., 2003, suppl., pp. 580–589.
[18] M. Pinto, G. Robledo, E. Liz, V. Tkachenko, and S. Trofimchuk, *Wright type delay differential equations with negative Schwarzian*, Discrete Contin. Dyn. Syst., 9 (2003), pp. 309–321.
[19] K. Mischaikow, H. Smith, and H. R. Thieme, *Asymptotically autonomous semiflows: chain recurrence and Lyapunov functions*, Trans. Amer. Math. Soc., 347 (1995), pp. 1669–1685.
[20] C. H. Ou and J. Wu, *Persistence of wavefronts in delayed non-local reaction diffusion equations*, J. Differential Equations, 235 (2007), pp. 219–261.
[21] J. D. Murray, *Mathematical Biology*, in Biomathematics 19, Springer-Verlag, Berlin, 1989.
[22] K. J. Palmer, *Exponential dichotomies and transversal homoclinic points*, J. Differential Equations, 55 (1984), pp. 225–256.
[23] E. C. Pielou, *Introduction to Mathematical Ecology*, Wiley, New York, 1969.
[24] J. A. Sherratt, *Wavefront propagation in a competition equation with a new motility term modelling contact inhibition between cell populations*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 456 (2000), pp. 2365–2386.
[25] F. E. Smith, *Population dynamics in Daphnia magna*, Ecology, 44 (1963), pp. 651–663.
[26] H. L. Smith and H. R. Thieme, *Strongly order preserving semiflows generated by functional-differential equations*, J. Differential Equations, 93 (1991), pp. 332–363.
[27] J. W.-H.So and J. S. Yu, *On the uniform stability for a "food-limited" population model with time delay*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 991–1002.
[28] A. I. Volpert, V. A. Volpert, and V. A. Volpert, *Traveling Wave Solutions of Parabolic Systems*, Trans. Math. Monogr. 140, American Mathematical Society, Providence, RI, 1994. Translated from the Russian manuscript by James F. Heyda.
[29] J. Wu and X. Zou, *Traveling wave fronts of reaction-diffusion systems with delay*, J. Dynam. Differential Equations, 13 (2001), pp. 651–687.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 126–137

EXISTENCE, UNIQUENESS, AND REGULARITY OF OPTIMAL TRANSPORT MAPS*

ALESSIO FIGALLI†

Abstract. Adapting some techniques and ideas of McCann [*Duke Math. J.*, 80 (1995), pp. 309–323], we extend a recent result with Fathi [*Optimal Transportation on Manifolds*, preprint] to yield existence and uniqueness of a unique transport map in very general situations, without any integrability assumption on the cost function. In particular this result applies for the optimal transportation problem on an n-dimensional noncompact manifold M with a cost function induced by a C^2-Lagrangian, provided that the source measure vanishes on sets with σ-finite $(n-1)$-dimensional Hausdorff measure. Moreover we prove that in the case $c(x, y) = d^2(x, y)$, the transport map is approximatively differentiable a.e. with respect to the volume measure, and we extend some results of [D. Cordero-Erasquin, R. J. McCann, and M. Schmuckenschlager, *Invent. Math.*, 146 (2001), pp. 219–257] about concavity estimates and displacement convexity.

Key words. optimal transportation, existence, uniqueness, approximate differentiability, concavity estimate, displacement convexity

AMS subject classifications. 28A12, 28A15, 37J50, 49J99, 49N60, 53B21

DOI. 10.1137/060665555

1. Introduction and main result.

Let M be an n-dimensional manifold (Hausdorff and with a countable basis), N a Polish space, $c : M \times N \to \mathbb{R}$ a cost function, and μ and ν two probability measures on M and N, respectively.

In a recent work with Fathi [6], we proved, under general assumption on the cost function, existence and uniqueness of optimal transport maps for the Monge–Kantorovich problem. More precisely, the result is as follows.

THEOREM 1.1. *Assume that $c : M \times N \to \mathbb{R}$ is lower semicontinuous, bounded from below, and such that*

$$\int_{M\times N} c(x, y)\, d\mu(x)\, d\nu(y) < +\infty.$$

If

(i) *$x \mapsto c(x, y) = c_y(x)$ is locally semiconcave in x locally uniformly in y;*
(ii) *$\frac{\partial c}{\partial x}(x, \cdot)$ is injective on its domain of definition;*
(iii) *and the measure μ gives zero mass to sets with σ-finite $(n-1)$-dimensional Hausdorff measure,*

then there exists a measurable map $T : M \to N$ such that any plan γ optimal for the cost c is concentrated on the graph of T.

More precisely, there exists a sequence of Borel subsets $B_n \subset M$, with $B_n \subset B_{n+1}$, $\mu(B_n) \nearrow 1$, and a sequence of locally semiconcave functions $\varphi_n : M \to \mathbb{R}$, where φ_n is differentiable on B_n, such that, thanks to assumption (ii)*, the map $T : M \to N$ is uniquely defined on B_n by*

$$\frac{\partial c}{\partial x}(x, T(x)) = d_x\varphi_n. \tag{1}$$

---

*Received by the editors July 20, 2006; accepted for publication (in revised form) October 9, 2006; published electronically April 20, 2007.
http://www.siam.org/journals/sima/39-1/66555.html

†Scuola Normale Superiore, Piazza dei Cavalieri 7, 56100 Pisa, Italy (a.figalli@sns.it).

This implies both existence of an optimal transport map and uniqueness for the Monge problem.

Now we want to generalize this existence and uniqueness result for optimal transport maps without any integrability assumption on the cost function, adapting the ideas of [8]. We observe that, without the hypothesis

$$\int_{M\times N} c(x,y)\,d\mu(x)\,d\nu(y) < +\infty,$$

denoting with $\Pi(\mu,\nu)$ the set of probability measures on $M\times N$ whose marginals are μ and ν, in general the minimization problem

$$C(\mu,\nu) := \inf_{\gamma\in\Pi(\mu,\nu)} \left\{ \int_{M\times N} c(x,y)\,d\gamma(x,y) \right\} \tag{2}$$

is ill-posed, as it may happen that $C(\mu,\nu) = +\infty$. However, it is known that the optimality of a transport plan γ is equivalent to the c-cyclical monotonicity of the measure-theoretic support of γ whenever $C(\mu,\nu) < +\infty$ (see [2], [11], [13]), and so one may ask whether the fact that the measure-theoretic support of γ is c-cyclically monotone implies that γ is supported on a graph. Moreover one can also ask whether this graph is unique, that is, it does not depend on γ, which is the case when the cost is $\mu\otimes\nu$ integrable, as Theorem 1.1 tells us. In that case, uniqueness follows by the fact that the functions φ_n are constructed using a pair of functions (φ,ψ) which is optimal for the dual problem, and so they are independent of γ (see [6] for more details). The result we now want to prove is the following.

THEOREM 1.2. *Assume that $c : M\times N \to \mathbb{R}$ is lower semicontinuous and bounded from below, and let γ be a plan concentrated on a c-cyclically monotone set. If*

(i) *the family of maps $x \mapsto c(x,y) = c_y(x)$ is locally semiconcave in x locally uniformly in y;*
(ii) *$\frac{\partial c}{\partial x}(x,\cdot)$ is injective on its domain of definition;*
(iii) *and the measure μ gives zero mass to sets with σ-finite $(n-1)$-dimensional Hausdorff measure,*

then γ is concentrated on the graph of a measurable map $T : M \to N$ (existence). Moreover, if $\tilde\gamma$ is another plan concentrated on a c-cyclically monotone set, then $\tilde\gamma$ is concentrated on the same graph (uniqueness).

Once the above result is proven, the uniqueness of the Wasserstein geodesic between absolutely continuous measures will follow as a simple corollary (see section 3). Finally, in subsection 3.1, we will prove that in the particular case $c(x,y) = \frac{1}{2}d^2(x,y)$, the optimal transport map is approximatively differentiable a.e. with respect to the volume measure, and we will obtain a concavity estimate on the Jacobian of the optimal transport map, which will allows us to generalize to noncompact manifolds a displacement convexity result proven in [4].

2. Proof of Theorem 1.2.

Existence. We want to prove that γ is concentrated on a graph. First we recall that since γ is concentrated on a c-cyclically monotone set, there exists a pair of functions (φ,ψ), with φ μ-measurable and ψ ν-measurable, such that

$$\varphi(x) = \inf_{y\in N} \psi(y) + c(x,y) \quad \forall x\in M,$$

which implies

$$\varphi(x) - \psi(y) \le c(x,y) \quad \forall (x,y)\in M\times N.$$

Moreover we have

$$\varphi(x) - \psi(y) = c(x,y) \quad \gamma\text{-a.e.} \tag{3}$$

and there exists a point $x_0 \in M$ such that $\varphi(x_0) = 0$ (see [13, Theorem 5.9]). In particular, this implies

$$\psi(y) \geq -c(x_0,y) > -\infty \quad \forall y \in N.$$

So, we can argue as in [6]. More precisely, given a suitable increasing sequence of compact sets $(K_n) \subset N$ such that $\nu(K_n) \nearrow 1$ and $\psi \geq -n$ on K_n (it suffices to take an increasing sequence of compact sets $K_n \subset \{\psi \geq -n\}$ such that $\nu(\{\psi \geq -n\} \setminus K_n) \leq \frac{1}{n}$), we consider the locally semiconcave function

$$\varphi_n(x) := \inf_{y \in K_n} \psi(y) + c(x,y). \tag{4}$$

Then, thanks to (3), it is possible to find an increasing sequence of Borel sets $D_n \subset \operatorname{supp}(\mu)$, with $\mu(D_n) \nearrow 1$, such that φ_n is differentiable on D_n, $\varphi_n \equiv \varphi$ on D_n, the set $\{\varphi_n = \varphi\}$ has μ-density 1 at all the points of D_n, and γ is concentrated on the graph of the map T uniquely determined on D_n by

$$\frac{\partial c}{\partial x}(x, T(x)) = d_x\varphi_n \quad \text{for } x \in D_n.$$

Moreover one has

$$\varphi(x) = \psi(T(x)) + c(x, T(x)) \quad \forall x \in \bigcup_n D_n \tag{5}$$

(see [6] for more details).

Uniqueness. As we observed before, the difference here with the case of Theorem 1.1 is that the function φ_n depends on the pair (φ, ψ), which in this case depends on γ. Let $(\tilde{\varphi}, \tilde{\psi})$ be a pair associated to $\tilde{\gamma}$ as above, and let $\tilde{\varphi}_n$ and $\tilde{D}_n$ be such that $\tilde{\gamma}$ is concentrated on the graph of the map $\tilde{T}$ determined on $\tilde{D}_n$ by

$$\frac{\partial c}{\partial x}(x, \tilde{T}(x)) = d_x\tilde{\varphi}_n \quad \text{for } x \in \tilde{D}_n.$$

We need to prove that $T = \tilde{T}$ μ-a.e.

Let us define $C_n := D_n \cap \tilde{D}_n$. Then $\mu(C_n) \nearrow 1$. We want to prove that if x is a μ-density point of C_n for a certain n, then $T(x) = \tilde{T}(x)$ (we recall that since $\mu(\cup_n C_n) = 1$, the union of the μ-density points of C_n is also of full μ-measure; see, for example, [5, Chapter 1.7]).

Let us assume by contradiction that $T(x) \neq \tilde{T}(x)$, that is,

$$d_x\varphi_n \neq d_x\tilde{\varphi}_n.$$

Since $x \in \operatorname{supp}(\mu)$, each ball around x must have positive measure under μ. Moreover, the fact that the sets $\{\varphi_n = \varphi\}$ and $\{\tilde{\varphi}_n = \tilde{\varphi}\}$ have μ-density 1 in x implies that the set

$$\{\varphi = \tilde{\varphi}\}$$

has μ-density 0 in x. In fact, as φ_n and $\tilde{\varphi}_n$ are locally semiconcave, up to adding a C^1 function they are concave in a neighborhood of x and their gradients differ at x. So we can apply the nonsmooth version of the implicit function theorem proven in [8], which tells us that $\{\varphi_n = \tilde{\varphi}_n\}$ is a set with finite $(n-1)$-dimensional Hausdorff measure in a neighborhood of x (see [8, Theorem 17 and Corollary 19]). So we have

$$\limsup_{r\to 0} \frac{\mu(\{\varphi = \tilde{\varphi}\} \cap B_r(x))}{\mu(B_r(x))} \le \limsup_{r\to 0} \Bigg[\frac{\mu(\{\varphi \ne \varphi_n\} \cap B_r(x))}{\mu(B_r(x))} + \frac{\mu(\{\varphi_n = \tilde{\varphi}_n\} \cap B_r(x))}{\mu(B_r(x))} + \frac{\mu(\{\tilde{\varphi}_n \ne \tilde{\varphi}\} \cap B_r(x))}{\mu(B_r(x))} \Bigg] = 0.$$

Therefore, exchanging φ with $\tilde{\varphi}$ if necessary, we may assume that

$$\mu(\{\varphi < \tilde{\varphi}\} \cap B_r(x)) \ge \frac{1}{4}\mu(B_r(x)) \quad \text{for } r > 0 \text{ sufficiently small.} \tag{6}$$

Let us define $A := \{\varphi < \tilde{\varphi}\}$, $A_n := \{\varphi_n < \tilde{\varphi}_n\}$, $E_n := A \cap A_n \cap C_n$. Since the sets $\{\varphi_n = \varphi\}$ and $\{\tilde{\varphi}_n = \tilde{\varphi}\}$ have μ-density 1 in x, and x is a μ-density point of C_n, we have

$$\lim_{r\to 0} \frac{\mu((A \setminus E_n) \cap B_r(x))}{\mu(B_r(x))} = 0,$$

and so, by (6), we get

$$\mu(E_n \cap B_r(x)) \ge \frac{1}{5}\mu(B_r(x)) \quad \text{for } r > 0 \text{ sufficiently small.} \tag{7}$$

Now, arguing as in the proof of Aleksandrov's lemma (see [8, Lemma 13]), we can prove that

$$X := \tilde{T}^{-1}(T(A)) \subset A$$

and $X \cap E_n$ lies a positive distance from x. In fact let us assume, without loss of generality, that

$$\varphi(x) = \varphi_n(x) = \tilde{\varphi}(x) = \tilde{\varphi}_n(x) = 0, \quad d_x\varphi_n \ne d_x\tilde{\varphi}_n = 0.$$

To obtain the inclusion $X \subset A$, let $z \in X$ and $y := \tilde{T}(z)$. Then $y = T(m)$ for a certain $m \in A$. For any $w \in M$, recalling (5), we have

$$\varphi(w) \le c(w, y) - c(m, y) + \varphi(m),$$

$$\tilde{\varphi}(m) \le c(m, y) - c(z, y) + \tilde{\varphi}(z).$$

Since $\varphi(m) < \tilde{\varphi}(m)$ we get

$$\varphi(w) < c(w, \tilde{T}(z)) - c(z, \tilde{T}(z)) + \tilde{\varphi}(z) \quad \forall w \in M.$$

In particular, taking $w = z$, we obtain $z \in A$, which proves the inclusion $X \subset A$.

Let us suppose now, by contradiction, that there exists a sequence $(z_k) \subset X \cap E_n$ such that $z_k \to x$. Again there exists m_k such that $\tilde{T}(z_k) = T(m_k)$. As $d_x\tilde{\varphi}_n = 0$, the closure of the superdifferential of a semiconcave function implies that $d_{z_k}\tilde{\varphi}_n \to 0$. We now observe that, arguing exactly as above with φ_n and $\tilde{\varphi}_n$ instead of φ and $\tilde{\varphi}$, by using (4), (5), and the fact that $\varphi = \varphi_n$ and $\tilde{\varphi} = \tilde{\varphi}_n$ on C_n, one obtains

$$\varphi_n(w) < c(w, \tilde{T}(z_k)) - c(z_k, \tilde{T}(z_k)) + \tilde{\varphi}_n(z_k) \quad \forall w \in M.$$

Taking w sufficiently near to x, we can assume that we are in $\mathbb{R}^n \times N$. We now remark that since $z_k \in E_n \subset \tilde{D}_n$, $\tilde{T}(z_k)$ vary in a compact subset of N (this follows by the construction of $\tilde{T}$). So, by hypothesis (i) on c, we can find a common modulus of continuity ω in a neighborhood of x for the family of uniformly semiconcave functions $z \mapsto c(z, \tilde{T}(z_k))$. Then we get

$$\begin{aligned}\varphi_n(w) &< \frac{\partial c}{\partial x}(z_k, \tilde{T}(z_k))(w - z_k) + \omega(|w - z_k|)|w - z_k| + \tilde{\varphi}_n(z_k) \\ &= d_{z_k}\tilde{\varphi}_n(w - z_k) + \omega(|w - z_k|)|w - z_k| + \tilde{\varphi}_n(z_k).\end{aligned}$$

Letting $k \to \infty$ and recalling that $d_{z_k}\tilde{\varphi}_n \to 0$ and $\tilde{\varphi}_n(x) = \varphi_n(x) = 0$, we obtain

$$\varphi_n(w) - \varphi_n(x) \le \omega(|w - x|)|w - x| \Rightarrow d_x\varphi_n = 0,$$

which is absurd.

Thus there exists $r > 0$ such that $B_r(x) \cap E_n$ and $X \cap E_n$ are disjoint, and (7) holds. Defining now $Y := T(A)$, by (7) we obtain

$$\begin{aligned}\nu(Y) = \mu(T^{-1}(Y)) \ge \mu(A) = \mu(E_n) + \mu(A \setminus E_n) \ge \mu(B_r(x) \cap E_n) \\ + \mu(X \cap E_n) + \mu(X \setminus E_n) = \mu(B_r(x) \cap E_n) + \mu(X) \ge \frac{1}{5}\mu(B_r(x)) + \nu(Y),\end{aligned}$$

which is absurd. □

Let us now consider the special case $N = M$, with M a complete manifold. As shown in [6], this theorem applies in the following cases:

1. $c : M \times M \to \mathbb{R}$ is defined by

$$c(x, y) := \inf_{\gamma(0)=x,\ \gamma(1)=y} \int_0^1 L(\gamma(t), \dot{\gamma}(t))\, dt,$$

where the infimum is taken over all the continuous piecewise C^1 curves, and the Lagrangian $L(x, v) \in C^2(TM, \mathbb{R})$ is C^2-strictly convex and uniform superlinear in v, and verifies a uniform boundedness in the fibers.

2. $c(x, y) = d^p(x, y)$ for any $p \in (1, +\infty)$, where $d(x, y)$ denotes a complete Riemannian distance on M.

Moreover, in the cases above, the following important fact holds.

Remark 2.1. For μ-a.e. x, there exists a unique curve from x to $T(x)$ that minimizes the action. In fact, since $\frac{\partial c}{\partial x}(x, y)$ exists at $y = T(x)$ for μ-a.e. x, the fact that $\frac{\partial c}{\partial x}(x, \cdot)$ is injective on its domain of definition tells us that its velocity at time 0 is μ-a.e. uniquely determined (see [6]).

Let us recall the following definition; see [1, Definition 5.5.1, p. 129].

DEFINITION 2.2 (approximate differential). *We say that $f : M \to \mathbb{R}^m$ has an approximate differential at $x \in M$ if there exists a function $h : M \to \mathbb{R}^m$ differentiable at x such that the set $\{f = h\}$ has density 1 at x with respect to the Lebesgue measure (this just means that the density is 1 in the charts). In this case, the approximate* *value of f at x is defined as $\tilde{f}(x) = h(x)$, and the approximate differential of f at x is defined as $\tilde{d}_x f = d_x h$. It is not difficult to show that this definition makes sense. In fact, neither $h(x)$ nor $d_x h$ depend on the choice of h, provided x is a density point of the set $\{f = h\}$ for the Lebesgue measure.*

We recall that many standard properties of the differential still hold for the approximate differential, such as linearity and additivity. In particular, it is simple to check that the property of being approximatively differentiable is stable by right composition with smooth maps (say C^1), and in this case the standard chain rule formula for the differentials holds. Moreover we remark that it makes sense to speak of approximate differential for maps between manifolds.

In [6], the following formula is proven: In the particular case $c(x, y) = d^2(x, y)$, if μ is absolutely continuous with respect to the Lebesgue measure, then the optimal transport map is given by

$$T(x) = \exp_x[-\tilde{\nabla}_x \varphi],$$

where $\tilde{\nabla}_x \varphi$ denotes the approximate gradient of φ at x, which simply corresponds to the element of $T_x M$ obtained from $\tilde{d}_x \varphi$ using the isomorphism with $T_x^* M$ induced by the Riemannian metric (the above formula generalizes the one found by McCann on compact manifolds; see [10]).

3. The Wasserstein space W_2. Let (M, g) be a smooth complete Riemannian manifold, equipped with its geodesic distance d and its volume measure vol. We denote with $P(M)$ the set of probability measures on M. The space $P(M)$ can be endowed with the so-called Wasserstein distance W_2:

$$W_2(\mu_0, \mu_1)^2 := \min_{\gamma \in \Pi(\mu_0, \mu_1)} \left\{ \int_{M \times M} d^2(x, y)\, d\gamma(x, y) \right\}.$$

The quantity $W_2(\mu_0, \mu_1)$ will be called the *Wasserstein distance of order* 2 between μ_0 and μ_1. It is well known that it defines a metric on $P(M)$ (not necessarily finite), and so one can speak about geodesic in the metric space $(P(M), W_2)$. This space turns out, indeed, to be a length space (see, for example, [12], [13]). Now, whenever $W_2(\mu_0, \mu_1) < +\infty$, we know that any optimal transport plan is supported on a c-cyclical monotone set (see, for example, [2], [11], [13]). We denote with $P^{ac}(M)$ the subset of $P(M)$ that consists of the Borel probability measures on M that are absolutely continuous with respect to vol. Thus, if $\mu_0, \mu_1 \in P^{ac}(M)$ and $W_2(\mu_0, \mu_1) < +\infty$, we know that there exists a unique transport map between μ_0 and μ_1.

PROPOSITION 3.1. *$P^{ac}(M)$ is a geodesically convex subset of $P(M)$. Moreover, if $\mu_0, \mu_1 \in P^{ac}(M)$ and $W_2(\mu_0, \mu_1) < +\infty$, then there is a unique Wasserstein geodesic $\{\mu_t\}_{t \in [0,1]}$ joining μ_0 to μ_1, which is given by*

$$\mu_t = (T_t)_\sharp \mu_0 := (\exp[-t\tilde{\nabla}\varphi])_\sharp \mu_0,$$

where $T(x) = \exp_x[-\tilde{\nabla}_x \varphi]$ is the unique transport map from μ_0 to μ_1, which is optimal for the cost $\frac{1}{2} d^2(x, y)$ (and so also optimal for the cost $d^2(x, y)$). Moreover,

1. *T_t is the unique optimal transport map from μ_0 to μ_t for all $t \in [0, 1]$;*
2. *T_t^{-1} is the unique optimal transport map from μ_t to μ_0 for all $t \in [0, 1]$ (and, if $t \in [0, 1)$, it is countably Lipschitz);*
3. *$T \circ T_t^{-1}$ is the unique optimal transport map from μ_t to μ_1 for all $t \in [0, 1]$ (and, if $t \in (0, 1]$, it is countably Lipschitz).*

Proof. Regarding the fact that $\mu_t \in P^{ac}(M)$ (which corresponds to saying that $P^{ac}(M)$ is geodesically convex) and the countably Lipschitz regularity of the transport maps (i.e., there exists a countable partition of M such that the map is Lipschitz on each set), they follow from the results in [6].

Thanks to the results proved in the last section, the proof of the rest of the proposition is quite standard. In fact, a basic representation theorem (see [13, Corollary 7.20]) states that any Wasserstein geodesic curve necessarily takes the form $\mu_t = (e_t)_{\#}\Pi$, where Π is a probability measure on the set Γ of minimizing geodesics $[0,1] \to M$, and $e_t : \Gamma \to M$ is the evaluation at time t: $e_t(\gamma) := \gamma(t)$. Thus the thesis follows from Remark 2.1. □

The above result tells us that also $(P^{ac}(M), W_2)$ is a length space.

3.1. Regularity, concavity estimate, and a displacement convexity result. We now consider the cost function $c(x,y) = \frac{1}{2}d^2(x,y)$. Let $\mu, \nu \in P^{ac}(M)$ with $W_2(\mu,\nu) < +\infty$, and let us denote with f and g their respective densities with respect to vol. Let

$$T(x) = \exp_x[-\tilde{\nabla}_x \varphi]$$

be the unique optimal transport map from μ to ν.

We recall that locally semiconcave functions with linear modulus admit vol-a.e. a second order Taylor expansion (see [3], [4]). Let us recall the definition of approximate hessian.

Definition 3.2 (approximate hessian). *We say that $f : M \to \mathbb{R}^m$ has a approximate hessian at $x \in M$ if there exists a function $h : M \to \mathbb{R}$ such that the set $\{f = h\}$ has density 1 at x with respect to the Lebesgue measure and h admits a second order Taylor expansion at x, that is, there exists a self-adjoint operator $H : T_xM \to T_xM$ such that*

$$h(\exp_x w) = h(x) + \langle \nabla_x h, w\rangle + \frac{1}{2}\langle Hw, w\rangle + o(\|w\|_x^2).$$

In this case the approximate hessian is defined as $\tilde{\nabla}_x^2 f := H$.

As in the case of the approximate differential, it is not difficult to show that this definition makes sense.

Observing that $d^2(x,y)$ is locally semiconcave with linear modulus (see [6, Appendix]), we get that φ_n is locally semiconcave with linear modulus for each n. Thus we can define μ-a.e. an approximate hessian for φ (see Definition 3.2):

$$\tilde{\nabla}_x^2 \varphi := \nabla_x^2 \varphi_n \quad \text{for } x \in D_n \cap E_n,$$

where D_n was defined in the proof of Theorem 1.2, E_n denotes the full μ-measure set of points where φ_n admits a second order Taylor expansion, and $\nabla_x^2 \varphi_n$ denotes the self-adjoint operator on T_xM that appears in the Taylor expansion on φ_n at x. Let us now consider, for each set $F_n := D_n \cap E_n$, an increasing sequence of compact sets $K_m^n \subset F_n$ such that $\mu(F_n \setminus \cup_m K_m^n) = 0$. We now define the measures $\mu_m^n := \mu \llcorner K_m^n$ and $\nu_m^n := T_\sharp \mu_m^n = (\exp[-\nabla\varphi_n])_\sharp \mu_m^n$, and we renormalize them in order to obtain two probability measures:

$$\hat{\mu}_m^n := \frac{\mu_m^n}{\mu_m^n(M)} \in P_2^{ac}(M), \quad \hat{\nu}_m^n := \frac{\nu_m^n}{\nu_m^n(M)} = \frac{\nu_m^n}{\mu_m^n(M)} \in P_2^{ac}(M).$$

We now observe that T is still optimal. In fact, if this were not the case, we would have

$$\int_{M\times M} c(x,S(x))\,d\hat{\mu}_m^n(x) < \int_{M\times M} c(x,T(x))\,d\hat{\mu}_m^n(x)$$

for a certain S transport map from $\hat{\mu}_m^n$ to $\hat{\nu}_m^n$. This would imply that

$$\int_{M\times M} c(x,S(x))\,d\mu_m^n(x) < \int_{M\times M} c(x,T(x))\,d\mu_m^n(x),$$

and so the transport map

$$\tilde{S}(x) := \begin{cases} S(x) & \text{if } x\in K_m^n, \\ T(x) & \text{if } x\in M\setminus K_m^n \end{cases}$$

would have a cost strictly less than the cost of T, which would contradict the optimality of T.

We will now apply the results of [4] to the compactly supported measures $\hat{\mu}_m^n$ and $\hat{\nu}_m^n$ in order to get information on the transport problem from μ to ν. In what follows we will denote by $\nabla_x d_y^2$ and by $\nabla_x^2 d_y^2$, respectively, the gradient and the hessian with respect to x of $d^2(x,y)$, and by $d_x\exp$ and $d(\exp_x)_v$ the two components of the differential of the map $TM\ni(x,v)\mapsto\exp_x[v]\in M$ (whenever they exist). By [4, Theorem 4.2], we get the following.

THEOREM 3.3 (Jacobian identity a.e.). *There exists a subset $E\subset M$ such that $\mu(E)=1$ and, for each $x\in E$, $Y(x):=d(\exp_x)_{-\tilde{\nabla}_x\varphi}$ and $H(x):=\frac{1}{2}\nabla_x^2 d_{T(x)}^2$ both exist and we have*

$$f(x)=g(T(x))\det[Y(x)(H(x)-\tilde{\nabla}_x^2\varphi)]\neq 0.$$

Proof. It suffices to observe that [4, Theorem 4.2] applied to $\hat{\mu}_m^n$ and $\hat{\nu}_m^n$ gives that, for μ-a.e. $x\in K_m^n$,

$$\frac{f(x)}{\mu_m^n(M)}=\frac{g(T(x))}{\mu_m^n(M)}\det[Y(x)(H(x)-\nabla_x^2\varphi_n)]\neq 0,$$

which implies

$$f(x)=g(T(x))\det[Y(x)(H(x)-\tilde{\nabla}_x^2\varphi)]\neq 0 \quad \text{for } \mu\text{-a.e. } x\in K_m^n.$$

Passing to the limit as $m,n\to+\infty$ we get the result. □

We can thus define μ-a.e. the (weak) differential of the transport map at x as

$$d_xT := Y(x)\big(H(x)-\tilde{\nabla}_x^2\varphi\big).$$

Let us prove now that, indeed, $T(x)$ is approximately differentiable μ-a.e., and that the above differential coincides with the approximate differential of T. In order to prove this fact, let us first make a formal computation. Observe that since the map $x\mapsto\exp_x[-\frac{1}{2}\nabla_x d_y^2]=y$ is constant, we have

$$0=d_x(\exp_x[-\tfrac{1}{2}\nabla_x d_y^2])=d_x\exp[-\tfrac{1}{2}\nabla_x d_y^2]-d(\exp_x)_{-\frac{1}{2}\nabla_x d_y^2}\big(\tfrac{1}{2}\nabla_x^2 d_y^2\big)\quad\forall y\in M,$$

By differentiating (in the approximate sense) the equality $T(x) = \exp[-\tilde{\nabla}_x\varphi]$ and recalling the equality $\tilde{\nabla}_x\varphi = \frac{1}{2}\nabla_x d^2_{T(x)}$, we obtain

$$\begin{aligned}\tilde{d}_x T &= d(\exp_x)_{-\tilde{\nabla}_x\varphi}\big(-\tilde{\nabla}^2_x\varphi\big) + d_x \exp[-\tilde{\nabla}_x\varphi]\\ &= d(\exp_x)_{-\tilde{\nabla}_x\varphi}\big(-\tilde{\nabla}^2_x\varphi\big) + d(\exp_x)_{-\frac{1}{2}\nabla_x d^2_{T(x)}}\big(\tfrac{1}{2}\nabla^2_x d^2_{T(x)}\big)\\ &= d(\exp_x)_{-\tilde{\nabla}_x\varphi}\big(H(x) - \tilde{\nabla}^2_x\varphi\big),\end{aligned}$$

as wanted. In order to make the above proof rigorous, it suffices to observe that for μ-a.e. x, $T(x) \notin cut(x)$, where $cut(x)$ is defined as the set of points $z \in M$ which cannot be linked to x by an extendable minimizing geodesic. Indeed we recall that the square of the distance fails to be semiconvex at the cut locus, that is, if $x \in cut(y)$, then

$$\inf_{0<\|v\|_x<1} \frac{d^2_y(\exp_x[v]) - 2d^2_y(x) + d^2_y(\exp_x[-v])}{|v|^2} = -\infty$$

(see [4, Proposition 2.5]). Now fix $x \in F_n$. Since we know that $\frac{1}{2}d^2(z, T(x)) \geq \varphi_n(z) - \psi(T(x))$ with equality for $z = x$, we obtain a bound from below of the hessian of $d^2_{T(x)}$ at x in terms of the hessian of φ_n at x (see the proof of [4, Proposition 4.1(a)]). Thus, since each φ_n admits vol-a.e. a second order Taylor expansion, we obtain that, for μ-a.e. x,

$$x \notin cut(T(x)), \quad \text{or equivalently} \quad T(x) \notin cut(x).$$

This implies that all the computations we made above in order to prove the formula for $\tilde{d}_x T$ are correct. Indeed the exponential map $(x, v) \mapsto \exp_x[v]$ is smooth if $\exp_x[v] \notin cut(x)$, the function d^2_y is smooth around any $x \notin cut(y)$ (see [4, Paragraph 2]), and $\tilde{\nabla}_x\varphi$ is approximatively differentiable μ-a.e. Thus, recalling that, once we consider the right composition of an approximatively differentiable map with a smooth map, the standard chain rule holds (see the remarks after Definition 2.2), we have proved the following regularity result for the transport map.

Proposition 3.4 (approximate differentiability of the transport map). *The transport map is approximatively differentiable for μ-a.e. x, and its approximate differential is given by the formula*

$$\tilde{d}_x T = Y(x)\big(H(x) - \tilde{\nabla}^2_x\varphi\big),$$

where Y and H are defined in Theorem 3.3.

To prove our displacement convexity result, the following change of variables formula will be useful.

Proposition 3.5 (change of variables for optimal maps). *If $A : [0 + \infty) \to \mathbb{R}$ is a Borel function such that $A(0) = 0$, then*

$$\int_M A(g(y))\, d\,\mathrm{vol}(y) = \int_E A\left(\frac{f(x)}{J(x)}\right) J(x)\, d\,\mathrm{vol}(x),$$

where $J(x) := \det[Y(x)(H(x) - \tilde{\nabla}^2_x\varphi)] = \det[\tilde{d}_x T]$ (*either both integrals are undefined or both take the same value in* $\overline{\mathbb{R}}$).

The proof follows by the Jacobian identity proved in Theorem 3.3, exactly as in [4, Corollary 4.7].

Let us now define for $t \in [0,1]$ the measure $\mu_t := (T_t)_\sharp \mu$, where

$$T_t(x) = \exp_x[-t\tilde{\nabla}_x \varphi].$$

By the results in [6] and Proposition 3.1, we know that T_t coincides with the unique optimal map pushing μ forward to μ_t, and that μ_t is absolutely continuous with respect to vol for any $t \in [0,1]$.

Given $x, y \in M$, following [4], we define for $t \in [0,1]$

$$Z_t(x,y) := \{z \in M \mid d(x,z) = td(x,y) \text{ and } d(z,y) = (1-t)d(x,y)\}.$$

If N is now a subset of M, we set

$$Z_t(x,N) := \cup_{y \in N} Z_t(x,y).$$

Letting $B_r(y) \subset M$ denote the open ball of radius $r > 0$ centered at $y \in M$, for $t \in (0,1]$ we define

$$v_t(x,y) := \lim_{r \to 0} \frac{\mathrm{vol}(Z_t(x, B_r(y)))}{\mathrm{vol}(B_{tr}(y))} > 0$$

(the above limit always exists, though it will be infinite when x and y are conjugate points; see [4]). Arguing as in the proof of Theorem 3.3, by [4, Lemma 6.1] we get the following.

THEOREM 3.6 (Jacobian inequality). *Let E be the set of full μ-measure given by Theorem* 3.3. *Then for each $x \in E$, $Y_t(x) := d(\exp_x)_{-t\tilde{\nabla}_x \varphi}$ and $H_t(x) := \frac{1}{2}\nabla_x^2 d^2_{T_t(x)}$ both exist for all $t \in [0,1]$ and the Jacobian determinant*

$$J_t(x) := \det[Y_t(x)(H_t(x) - t\tilde{\nabla}_x^2 \varphi)] \tag{8}$$

satisfies

$$J_t^{\frac{1}{n}}(x) \geq (1-t)\,[v_{1-t}(T(x),x)]^{\frac{1}{n}} + t\,[v_t(x,T(x))]^{\frac{1}{n}}\, J_1^{\frac{1}{n}}(x).$$

We now consider as source measure $\mu_0 = \rho_0\, d\,\mathrm{vol}(x) \in P^{ac}(M)$ and as target measure $\mu_1 = \rho_1\, d\,\mathrm{vol}(x) \in P^{ac}(M)$, and we assume as before that $W_2(\mu_0, \mu_1) < +\infty$. By Proposition 3.1 we have

$$\mu_t = (T_t)_\sharp[\rho_0\, d\,\mathrm{vol}] = \rho_t\, d\,\mathrm{vol} \in P_2^{ac}(M)$$

for a certain $\rho_t \in L^1(M, d\,\mathrm{vol})$.

We now want to consider the behavior of the functional

$$U(\rho) := \int_M A(\rho(x))\, d\,\mathrm{vol}(x)$$

along the path $t \mapsto \rho_t$. In Euclidean spaces, this path is called *displacement interpolation* and the functional U is said to be *displacement convex* if

$$[0,1] \ni t \mapsto U(\rho_t) \quad \text{is convex for every } \rho_0, \rho_1.$$

A sufficient condition for the displacement convexity of U in $\mathbb{R}^n$ is that $A : [0,+\infty) \to \mathbb{R} \cup \{+\infty\}$ satisfy

$$(0,+\infty) \in s \mapsto s^n A(s^{-n}) \text{ is convex and nonincreasing, with } A(0) = 0 \tag{9}$$

(see [7], [9]). Typical examples include the entropy $A(\rho) = \rho \log \rho$ and the L^q-norm $A(\rho) = \frac{1}{q-1}\rho^q$ for $q \geq \frac{n-1}{n}$.

By all the results collected above, arguing as in the proof of [4, Theorem 6.2], we can prove that the displacement convexity of U is still true on Ricci nonnegative manifolds under the assumption (9).

THEOREM 3.7 (displacement convexity on Ricci nonnegative manifolds). *If* Ric $\geq$ 0 *and A satisfies* (9)*, then U is displacement convex.*

Proof. As we remarked above, T_t is the optimal transport map from μ_0 to μ_t. So, by Theorem 3.3 and Proposition 3.5, we get

$$U(\rho_t) = \int_M A(\rho_t(x))\, d\,\mathrm{vol}(x) = \int_{E_t} A\left(\frac{\rho_0(x)}{\left(J_t^{\frac{1}{n}}(x)\right)^n}\right)\left(J_t^{\frac{1}{n}}(x)\right)^n d\,\mathrm{vol}(x), \tag{10}$$

where E_t is the set of full μ_0-measure given by Theorem 3.3 and $J_t(x) \neq 0$ is defined in (8). Since Ric ≥ 0, we know that $v_t(x, y) \geq 1$ for every $x, y \in M$ (see [4, Corollary 2.2]). Thus, for fixed $x \in E_1$, Theorem 3.6 yields the concavity of the map

$$[0, 1] \ni t \mapsto J_t^{\frac{1}{n}}(x).$$

Composing this function with the convex nonincreasing function $s \mapsto s^n A(s^{-n})$ we get the convexity of the integrand in (10). The only problem we run into in trying to conclude the displacement convexity of U is that the domain of integration appears to depend on t. But, since by Theorem 3.3 E_t is a set of full μ_0-measure for any $t \in [0, 1]$, we obtain that, for fixed $t, t', s \in [0, 1]$,

$$U(\rho_{(1-s)t+st'}) \leq (1 - s)U(\rho_t) + sU(\rho_{t'}),$$

simply by computing each of the three integrals above on the full measure set $E_t \cap E_{t'} \cap E_{(1-s)t+st'}$. $\square$

Acknowledgment. I wish to thank Cédric Villani for fruitful discussions on this subject.

REFERENCES

[1] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and in the Wasserstein Space of Probability Measures,* Lectures in Math. ETH Zurich, Birkhäuser Verlag, Basel, 2005.

[2] L. Ambrosio and A. Pratelli, *Existence and stability results in the L^1 theory of optimal transportation*, in Optimal Transportation and Applications (Martina Franca, 2001), Lecture Notes in Math. 1813, Springer-Verlag, Berlin, 2003, pp. 123–160.

[3] V. Bangert, *Analytische Eigenschaften konvexer Funktionen auf Riemannschen Mannigfaltigkeiten*, J. Reine Angew. Math., 30 (1979), pp. 309–324.

[4] D. Cordero-Erasquin, R. J. McCann, and M. Schmuckenschlager, *A Riemannian interpolation inequality á la Borell, Brascamp, and Lieb*, Invent. Math., 146 (2001), pp. 219–257.

[5] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.

[6] A. Fathi and A. Figalli, *Optimal Transportation on Manifolds*, preprint, 2006.

[7] R. J. McCann, *A Convexity Principle for Interacting Gases and Equilibrium Crystals*, Ph.D. thesis, Princeton University, Princeton, NJ, 1994.

[8] R. J. McCann, *Existence and uniqueness of monotone measure-preserving maps*, Duke Math. J., 80 (1995), pp. 309–323.

[9] R. J. McCann, *A convexity principle for interacting gases*, Adv. Math., 128 (1997), pp. 153–179.

[10] R. J. McCann, *Polar factorization of maps on Riemannian manifolds*, Geom. Funct. Anal., 11 (2001), pp. 589–608.

[11] W. Schachermayer and J. Teichmann, *Characterization of optimal transport plans for the Monge–Kantorovich–problem*, Proc. Amer. Math. Soc., to appear.

[12] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, AMS, Providence, RI, 2004.

[13] C. Villani, *Optimal Transport, Old and New*, Lecture Notes, 2005 Saint-Flour Summer School; available online from www.umpa.ens-lyon.fr/˜cvillani.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 138–159

ASYMPTOTIC STABILITY OF TRAVELING WAVE FRONTS IN THE BUFFERED BISTABLE SYSTEM*

JE-CHIANG TSAI†

Abstract. In this paper, we study a model which describes the propagation of increased calcium concentration wave front in excitable systems with the diffusing species being buffered. Our goal is to prove the global exponential stability of the unique traveling wave front. Comparing with the unbuffered system, we conclude that multiple stationary buffers (buffers do not diffuse) cannot prevent the existence of a global asymptotic stable traveling wave front, or cannot eliminate propagated waves in the buffered bistable equation. Concerning the method of the proof, we will present a method in which only the comparison principle and suitably constructed supersolutions (subsolutions) are involved. The feature of the method is to avoid calculating the spectrum of the associated linear operator.

Key words. calcium, reaction-diffusion equations, traveling wave front, bistable equation, FitzHugh–Nagumo equations, asymptotic stability

AMS subject classifications. 34A34, 34A12, 35K57

DOI. 10.1137/050639685

1. Introduction.

Wave propagations have been observed in a wide array of biological and chemical systems (e.g., the FitzHugh–Nagumo model and Belousov–Zhabotinskii reaction), and have been an interesting subject of mathematical studies for many years. Recently, waves of increased calcium concentration that travel within cells have been studied in depth, both by experimentalists and theoreticians [13, 2, 26, 6, 33, 11]. Although their precise physiological function is not always clear, it has been known for some time that they are important ways in which cells can transmit an intracellular signal. Among them, the fertilization calcium (Ca^{2+}) wave in *Xenopus laevis* oocytes (eggs) has attracted much attention. This may be due to the interesting fact that the cell divisions that initiate development of *Xenopus* begin only after the fertilization calcium wave has traveled across the whole cell (see [31, 17]).

Existing experimental evidence suggests the following mechanism for calcium waves in cells [3, 41, 2, 24, 19, 18, 16, 39, 17, 36, 14]. Binding of an agonist (e.g., a hormone) to the receptors in the plasma membrane results in the production of IP_3. Then the IP_3 diffuses rapidly into the interior of the cell, where it can bind to IP_3R (which is the IP_3 receptor and acts as a calcium channel) on the membrane of the internal calcium store ER (endoplasmic reticulum) and activate the IP_3R. It turns out that calcium can be released from ER into cytosol through the IP_3R. Note that the IP_3R is regulated by both IP_3 and calcium. Therefore if the released calcium through the IP_3 receptor diffuses to neighboring IP_3R, then it will initiate further calcium release from there (we assume that the concentration of the released calcium is large enough). This is the so-called calcium-induced calcium release. Repetition of this process can then generate an advancing wave front of high calcium concentration.

*Received by the editors September 6, 2005; accepted for publication (in revised form) December 8, 2006; published electronically April 20, 2007. This work was partially supported by the National Science Council of the Republic of China under the contracts NSC 94-2115-M-019-002 and NSC 95-2115-M-194-015.

http://www.siam.org/journals/sima/39-1/63968.html

†Department of Mathematics, National Chung Cheng University, 168, University Road, Min-Hsiung, Chia-Yi 621, Taiwan (tsaijc@math.ccu.edu.tw).

Before stating the model for calcium waves, we make some comments on ER. First, in addition to the calcium flux through IP_3R, there are two other components of the calcium flux across the ER membrane. More precisely, they are direct leak through the membrane and calcium uptake by molecular pumps. Second, compared with the calcium concentration in cytosol, the calcium concentration in ER is huge and can be treated as a constant. Finally, we will assume that the ER is a homogeneous, continuous medium, and that the dynamics of calcium in cytosol are much slower than the gating variable for calcium inactivation of the IP_3R (see [15, 39, 31]).

Now we can present the model for the calcium waves. Indeed, by taking calcium diffusion into account [1, 15], the above discussion leads to the following equation:

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + f(u), \tag{1.1}$$

where u denotes the concentration of free cytosolic calcium, $D > 0$ is the diffusion coefficient of the free cytosolic calcium, and $f(u) = J_{\text{channel}} + J_{\text{pump}} + J_{\text{leak}}$ is the calcium fluxes through the IP_3R, pumps, and the passive leak, respectively. Regarding the form of the function f, we shall choose the well-known *bistable* nonlinearity. More precisely, f is a function in $C^1[a_0, a_2]$ such that $f(a_0) = f(a_1) = f(a_2) = 0$ for some $0 < a_0 < a_1 < a_2$ and the following condition holds:

$$f(u) < 0 \text{ in } (a_0, a_1),\ f(u) > 0 \text{ in } (a_1, a_2),\ f'(a_0) < 0,\ f'(a_2) < 0.$$

The reason for choosing such a nonlinearity is not only because it can maintain stable self-sustained waves [17, 10], but it is also thought to be essential in the fertilization calcium waves of *Xenopus laevis* oocytes (see [39, 24, 4, 9]). According to the discussion of Smith, Pearson, and Keizer [31], the state a_0 represents a stable resting state which is the basal calcium concentration in cytosol, the state a_2 is a stable resting point at high calcium concentration in cytosol, while the state a_1 is unstable and corresponds to a threshold for calcium-induced calcium release (see [31, 17]). For simplicity, in this paper we shall only consider the typical *bistable* nonlinearity, i.e.,

$$f(u) = u(1-u)(u-a)$$

for some $a \in (0, 1)$. However, since we shall only concern the existence and stability of traveling calcium waves, it will not lose any information that we want.

At first glance, the study of calcium waves is similar to the simplified version of the well-known FitzHugh–Nagumo model, which was originally designed for studying the action potential. However, there are still some crucial differences for the study of calcium waves. Among them, the calcium buffers are the most important. Calcium buffers are large proteins that act as the binding sites for calcium (see [26]). Typical examples are calsequestrin and calbindin. The importance of studying buffers lies in two facts. First, if we want to observe the waves in cells, we need to put the indicator dyes (e.g., Ca^{2+}-green dextran, BAPTA, and fura-2) into the cells. The dye is one kind of (exogeneous) buffer. Therefore, if the buffers cannot preserve the existence of the stable traveling calcium waves, then we may not be able to observe the traveling calcium waves in the laboratory. Second, a large fraction of cytosolic calcium (at least 99%) is bound to calcium (endogenous) buffers. Not only do these buffers restrict the diffusion of free calcium, they also affect the kinetics of calcium release and uptake, and thus they would be expected to have an important effect on the properties of calcium traveling waves (see [35, 17, 31]).

A widely used way to model buffering is to assume that calcium reacts with buffers according to the following reaction scheme (see Wagner and Keizer [38]):

$$\mathrm{Ca}^{2+} + \mathrm{B}_i \underset{k_-^i}{\overset{k_+^i}{\rightleftarrows}} \mathrm{Ca}^{2+}\mathrm{B}_i, \ \ i = 1, \dots, n, \tag{1.2}$$

where B_i denotes the ith buffer in its unbound form, and $\mathrm{Ca}^{2+}\mathrm{B}_i$ denotes the ith buffer that is bound to calcium. Let v_i denote the concentration $[\mathrm{B}_i]$ of the ith buffer and b_0^i denote the total amount of the ith buffer. Also note that $b_0^i = [\mathrm{B}_i] + [\mathrm{Ca}^{2+}\mathrm{B}_i]$. We shall assume that b_0^i is a constant. Then it follows from the law of mass action and (1.2) that the rate of change of u due to buffering is given by

$$\frac{du}{dt} = \sum_{i=1}^{n} [k_-^i(b_0^i - v_i) - k_+^i u v_i],$$

where k_+^i and k_-^i denote the forward and reverse rate constants of the ith reaction (1.2), respectively. Combining this with (1.1), we obtain the following buffered bistable system:

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + f(u) + \sum_{i=1}^{n} [k_-^i(b_0^i - v_i) - k_+^i u v_i], \tag{1.3}$$

$$\frac{\partial v_i}{\partial t} = k_-^i(b_0^i - v_i) - k_+^i u v_i, \ (x,t) \in \mathbf{R} \times (0,\infty), \ i = 1, \dots, n, \tag{1.4}$$

with the initial data

$$u(x,0) = \phi(x), \ v_i(x,0) = \psi_i(x), \ x \in \mathbf{R}, \ i = 1, \dots, n. \tag{1.5}$$

Note that we assume that buffers are stationary (buffers do not diffuse). Hence, mathematically, we also obtain an excitable system with the diffusing species being buffered (see [14, 17]). Thus we will not be restricted to the case for the fertilization calcium traveling wave in *Xenopus* eggs only, it may be applied to other cell types.

There have been a large number of numerical studies on the biological models including calcium buffers (see [8, 15, 23, 27]). Regarding the analytical studies, the assumption that the buffer has fast kinetics with respect to the other reactions in the model is always made. Under this assumption, the well-known *rapid buffering approximation* (see Wagner and Keizer [38]) can be applied, and the full model can be reduced to a single quasilinear equation, in which the effective diffusion coefficient of calcium now depends on the calcium concentration (see [20, 21, 22, 29, 30]). Using this reduction and assuming that there is only one buffer, Sneyd, Dale, and Duffy [34] have proved the existence of traveling waves of the reduced system and Tsai and Sneyd [36] have considered the uniqueness and stability of waves. Both Sneyd, Dale, and Duffy [34] and Slepchenko, Schaff, and Choi [28] also did many interesting simulations on the relation between the wave speed and the parameters k_-, k_+, b_0, and D. However, these previous studies left open the question of whether or not multiple buffers, not necessarily having fast kinetics, could eliminate wave propagation, and the stability is also unknown. Thus, in order to get a deeper understanding of how buffers affect the calcium traveling waves in the model (1.1), we shall not assume that the buffers have lower affinity and fast kinetics with respect to the other reactions in the model (1.3)–(1.4). Hence we need to consider the whole system (1.3)–(1.4) instead of the reduced equation.

Since we are concerned with how stationary buffers (buffers do not diffuse) affect wave activity, we briefly review the results of traveling wave solutions for the case of (1.3)–(1.4) without buffers, i.e., we consider the well-known *bistable* equation (1.1),

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + f(u).$$

A nonnegative function $U(\xi) \in C^2(\mathbf{R})$ is said to be a traveling wave solution of (1.1), if for some $c \in \mathbf{R}$, (U, c) satisfies that $u(x,t) = U(x - ct)$ is a solution of (1.1) and

$$U(+\infty) = 1, \;\; U(-\infty) = 0.$$

Note that c is the so-called wave speed associated with the profile of the traveling wave front U. Fife and Mcleod [10] have shown that for any given $a \in (0, 1/2)$, there exists a unique $c := c(a) < 0$ such that there is a unique (up to a translation) traveling wave solution U of (1.1) with wave speed c. Moreover, the profile of the front U is monotone increasing on $\mathbf{R}$. Furthermore, they have also considered the asymptotic stability for this traveling wave front $U(\xi)$. More precisely, let $u(x,t)$ be the solution of (1.1) with the initial data $u(x, 0) = \phi(x)$ satisfying $\phi \in [0, 1]$ on $\mathbf{R}$. Then $u(x,t)$ approaches a translate of U uniformly in x and exponentially fast in time if $\limsup_{x\to-\infty} \phi(x) > 0$ and $1 - \liminf_{x\to+\infty} \phi(x) > 0$ are sufficiently small.

Next we turn to the system (1.3)–(1.4). First, we set up some notation:

$$\begin{aligned}
&\kappa_i(u) := k_-^i b_0^i/(k_+^i u + k_-^i) \;\text{ for }\; i = 1, \ldots, n,\\
&\mathbf{b}_0 = (b_0^1, \ldots, b_0^n) := (\kappa_1(0), \ldots, \kappa_n(0)),\\
&\mathbf{b}_1 = (b_1^1, \ldots, b_1^n) := (\kappa_1(a), \ldots, \kappa_n(a)),\\
&\mathbf{b}_2 = (b_2^1, \ldots, b_2^n) := (\kappa_1(1), \ldots, \kappa_n(1)),\\
&\mathbf{v}(x,t) := (v_1(x,t), \ldots, v_n(x,t)).
\end{aligned}$$

We are interested in traveling wave solutions of (1.3)–(1.4) that connect the two equilibria $(0, \mathbf{b}_0)$ and $(1, \mathbf{b}_2)$. More precisely, a set of nonnegative functions $(\mathcal{U}(\xi), \mathbf{\Pi}(\xi)) = (\mathcal{U}(\xi), \Pi_1(\xi), \ldots, \Pi_n(\xi)) \in \mathbf{C}^2(\mathbf{R}) \times \mathbf{C}^1(\mathbf{R}) \times \cdots \times \mathbf{C}^1(\mathbf{R})$ are said to be a traveling wave solution of (1.3)–(1.4), if for some $c < 0$, $(\mathcal{U}(\xi), \mathbf{\Pi}(\xi), c)$ satisfies that $(u(x,t), \mathbf{v}(x,t)) = (\mathcal{U}(x - ct), \mathbf{\Pi}(x - ct))$ is a solution of (1.3)–(1.4) and

$$\mathcal{U}(+\infty) = 1, \; \mathbf{\Pi}(+\infty) = \mathbf{b}_2, \; \mathcal{U}(-\infty) = 0, \; \mathbf{\Pi}(-\infty) = \mathbf{b}_0. \tag{1.6}$$

As before, c is the so-called wave speed associated with the profile of the traveling wave front $(\mathcal{U}, \mathbf{\Pi})$. Therefore, $(\mathcal{U}(\xi), \mathbf{\Pi}(\xi))$ satisfies the following ordinary differential equations:

$$\begin{aligned}
&D\ddot{\mathcal{U}} + c\,\dot{\mathcal{U}} + f(\mathcal{U}) + \sum_{i=1}^{n}\Big[k_-^i(b_0^i - \Pi_i) - k_+^i \mathcal{U}\Pi_i\Big] = 0,\\
&c\dot{\Pi}_i + k_-^i(b_0^i - \Pi_i) - k_+^i \mathcal{U}\Pi_i = 0, \; \xi = x - ct \in \mathbf{R}, \; i = 1, \ldots, n,
\end{aligned} \tag{1.7}$$

where $\cdot$ denotes $d/d\xi$. In [35], we have proved that for any given $a \in (0, 1/2)$, there exists a unique $c := c(a) < 0$ such that there exists a unique (up to a translation) traveling wave solution $(\mathcal{U}, \mathbf{\Pi})$ of the buffered bistable equations (1.3)–(1.4) with wave speed c. Moreover, $(\mathcal{U}, \mathbf{\Pi})$ satisfies that $\dot{\mathcal{U}} > 0$ and $\dot{\Pi}_i < 0$, $i = 1, \ldots, n$, on $\mathbf{R}$. Throughout the remainder of this paper, $(\mathcal{U}, \mathbf{\Pi})$ will denote the unique traveling wave solution of (1.3)–(1.4) with speed c and $\mathcal{U}(0) = 1/2$.

In [35], the stability of this traveling wave solution of (1.3)–(1.4) was established under the following technical constraint on the initial values. Let $a \in (0, 1/2)$ and $(u(x,t), \mathbf{v}(x,t))$ be the solution of (1.3)–(1.4) with the initial data $(u(\cdot,0), \mathbf{v}(\cdot,0))$ satisfying the following conditions:

(1) $u(\cdot,0)$, $v_i(\cdot,0)$ are sufficiently smooth and $u_x(x,0) \geq 0$, $v_{i,x}(\cdot,0) \leq 0$ on $\mathbf{R}$ for $i = 1,\ldots,n$;
(2) $\tilde{u}_t(x,0) \geq 0$ and $\tilde{v}_{i,t}(x,0) \leq 0$ for all $x \in \mathbf{R}$ and $i = 1,\ldots,n$;
(3) $\sup_{x\in\mathbf{R}} |u(x,0)| + \sup_{x\in\mathbf{R}} |u_x(x,0)| + \sup_{x\in\mathbf{R}} |u_{xx}(x,0)| + \sup_{x\in\mathbf{R}} |u_{xxx}(x,0)| < +\infty$;
(4) $u(\cdot,0)$, $v_i(\cdot,0)$, $i = 1,\ldots,n$, are in $[0,1]$ and $[b_2^i, b_0^i]$, respectively;
(5) $1 - \phi_2 > 0$, $\psi_{2i} - b_2^i > 0$, $\phi_0 > 0$ and $b_0^i - \psi_{0i} > 0$ are sufficiently small for $i = 1,\ldots,n$,

where $v_{i,x} = \partial v_i/\partial x$, $\tilde{v}_{i,t} = \partial \tilde{v}_i/\partial t$,

$$\phi_2 = \lim_{x\to+\infty} u(x,0),\ \psi_{2i} = \lim_{x\to+\infty} v_i(x,0),\ \phi_0 = \lim_{x\to-\infty} u(x,0),$$
$$\text{and}\ \ \psi_{0i} = \lim_{x\to-\infty} v_i(x,0),$$

and $(\tilde{u}(x,t), \tilde{\mathbf{v}}(x,t)) = (u(x+ct,t), \mathbf{v}(x+ct,t))$ on $\mathbf{R}\times\mathbf{R}^+$. Then there exists $x_0 \in \mathbf{R}$ such that

$$\lim_{t\to+\infty} |u(x,t) - \mathcal{U}(x-ct+x_0)| = 0,\ \ \lim_{t\to+\infty} |v_i(x,t) - \Pi_i(x-ct+x_0)| = 0,\ i = 1,\ldots,n,$$

uniformly with respect to $x \in \mathbf{R}$. Roughly speaking, this implies that a solution of (1.3)–(1.4) which vaguely resembles a traveling front $(\mathcal{U}, \mathbf{\Pi})$ at initial time will develop into a translation of such a traveling front as $t \to +\infty$.

It is obvious that the conditions (1), (2), and (3) in our previous stability result are technical. Moreover, the condition (5) also implies that the previous result is only "local stability." Therefore, the goal of this paper is to remove these conditions and our main result is the following theorem which is global exponential stability.

THEOREM 1. *Let $a \in (0,1/2)$ and $(u(x,t), \mathbf{v}(x,t))$ be the solution of* (1.3)–(1.4) *with the initial data $(u(\cdot,0), \mathbf{v}(\cdot,0))$ satisfying the following conditions:*

(1) *$u(\cdot,0)$, $v_i(\cdot,0)$, are uniformly Hölder continuous in $\mathbf{R}$ with exponent α for some $\alpha \in (0,1)$ and for $i = 1,\ldots,n$;*
(2) *$u(\cdot,0)$, $v_i(\cdot,0)$, $i = 1,\ldots,n$, are in $[0,1]$ and $[b_2^i, b_0^i]$, respectively;*
(3)

$$\liminf_{x\to+\infty} u(x,0) > a,\ \limsup_{x\to+\infty} v_i(x,0) < b_1^i,\ \limsup_{x\to-\infty} u(x,0) < a,\ \liminf_{x\to-\infty} v_i(x,0) > b_1^i$$

for $i = 1,\ldots,n$.

Then there exists a positive constant κ which is independent of the initial data $(u(\cdot,0), \mathbf{v}(\cdot,0))$ such that

$$\sup_{x\in\mathbf{R}} |u(x,t) - \mathcal{U}(x-ct+\xi^*)| \leq Ke^{-\kappa t},\quad \sup_{x\in\mathbf{R}} |v_i(x,t) - \Pi_i(x-ct+\xi^*)| \leq Ke^{-\kappa t},$$

for $i = 1,\ldots,n$, and for some constants ξ^* and K which may depend on $(u(\cdot,0), \mathbf{v}(\cdot,0))$.

Note that the condition (3) is weaker than the previous one.

Comparing this result with the one of the unbuffered equation (1.1), we may conclude that *physiologically, a unique asymptotic stable traveling wave front exists as long as it exists in the absence of buffers.* Hence, our results complete the picture of how stationary buffers affect wave activity in the bistable equation. This may suggest that the same result is true for more complex and realistic models of calcium wave propagation (see [17, 31, 35]).

We conclude the introduction with some comments on the method of our proof. In general, the key step of the commonly used method for proving the asymptotic exponential stability of a traveling wave solution is to study the spectrum of the linear operator associated with the traveling wave solution under study, which is sometimes bothersome (see [10, 25, 40, 37] and references therein). Here, to avoid the spectrum analysis, we shall use a method involving only comparison principle and suitably constructed supersolutions (subsolutions) to establish our result. In fact, we shall extend the method of Chen [5] to prove our result, where the author studied the existence, uniqueness, and stability of a single nonlocal equation. However, since our problem is a system of differential equations (not a single equation), this method cannot be carried along the whole way and much modification is needed. In fact, for each crucial part (except the very final argument) in [5], we need to find the counterpart for our problem.

This paper is organized as follows. In section 2, we shall state some properties of (1.3)–(1.4) and derive the comparison principle for the subsolution and supersolution of (1.3)–(1.4). Then in section 3, we will prove Theorem 1.

2. Preliminaries. First, for simplicity we set up some notation. For each $i = 1, \ldots, n$, we set

$$F(u, \mathbf{v}) := f(u) + \sum_{i=1}^{n} G_i(u, \mathbf{v}),$$

$$G_i(u, \mathbf{v}) := k_-^i b_0^i - (k_+^i u + k_-^i) v_i, \ i = 1, \ldots, n.$$

Then we can rewrite (1.3)–(1.4) as the following system:

$$\frac{\partial u}{\partial t} = L_1[u, \mathbf{v}] := D\frac{\partial^2 u}{\partial x^2} + F(u, \mathbf{v}), \tag{2.1}$$

$$\frac{\partial v_i}{\partial t} = L_{2i}[u, \mathbf{v}] := G_i(u, \mathbf{v}), \ (x, t) \in \mathbf{R} \times \mathbf{R}^+ \ (\mathbf{R}^+ := (0, \infty)), \ i = 1, \ldots, n. \tag{2.2}$$

Note that $G_i(u, \mathbf{v}) = 0$ if and only if $v_i = \kappa_i(u)$. Furthermore, we have that

$$F_u(u, \mathbf{v}) = f'(u) - \sum_{i=1}^{n} k_+^i v_i, \quad F_{v_i}(u, \mathbf{v}) = -(k_+^i u + k_-^i) < 0,$$

$$G_{i,u}(u, \mathbf{v}) = -k_+^i v_i \le 0, \quad G_{i,v_i}(u, \mathbf{v}) = -(k_+^i u + k_-^i) < 0$$

for all $u \in [-\min_{j=1,\ldots,n}\{1, k_-^j/(2k_+^j)\}, 2]$ and $v_i \in [0, \infty)$, $i = 1, \ldots, n$. Finally, for two vectors $\mathbf{c} = (c_1, \ldots, c_n)$ and $\mathbf{d} = (d_1, \ldots, d_n)$, the symbol $\mathbf{c} < \mathbf{d}$ means $c_i < d_i$ for $i = 1, \ldots, n$, and $\mathbf{c} \le \mathbf{d}$ means $c_i \le d_i$ for $i = 1, \ldots, n$.

We shall investigate the asymptotic behavior as $t \to +\infty$ of the solution of (2.1)–(2.2) with the initial data $(u(\cdot, 0), \mathbf{v}(\cdot, 0))$. For this we briefly discuss the existence of the global solution $(u, \mathbf{v})$ of (2.1)–(2.2) with the initial data $(u(\cdot, 0), \mathbf{v}(\cdot, 0))$ (also see [14]). Suppose that $u(\cdot, 0)$ and $v_i(\cdot, 0)$, $i = 1, \ldots, n$, are uniformly Hölder continuous in $\mathbf{R}$ with exponent α for some $\alpha \in (0, 1)$ and that $0 \le u(x, 0) \le 1$ and $b_2^i \le v_i(x, 0) \le b_0^i$ for all $x \in \mathbf{R}$. Then by using an iteration method of Evans and Shenk [7], a regularity theory of parabolic equations (cf. [12]), and an invariance region theory [32, Theorem 14.11 on p. 203], we can prove that there is a unique solution $(u,\mathbf{v})$ defined for all $t>0$ such that $u\in C^{2,1}(\mathbf{R}\times\mathbf{R}^+)\cap C^0(\mathbf{R}\times[0,\infty))$, $v_i\in C^0(\mathbf{R}\times\mathbf{R}^+)$, and $v_{i,t}\in C^0(\mathbf{R}\times\mathbf{R}^+)$ for $i=1,\dots,n$, where the set $C^{2,1}(\mathbf{R}\times\mathbf{R}^+)$ consists of all functions that are once continuously differentiable in t and twice continuously differentiable in x for all $(x,t)\in\mathbf{R}\times\mathbf{R}^+$, and the set $C^0(\mathbf{R}\times[0,\infty))$ consists of continuous functions in $\mathbf{R}\times[0,\infty)$. Moreover, we recall from [35, Lemma 4.1] the following proposition of invariance regions.

PROPOSITION 1 (invariance region). *Let $(u,\mathbf{v})$ be a global solution of* (2.1)–(2.2) *with $0\le u(x,0)\le 1$ and $\mathbf{b}_2\le\mathbf{v}(x,0)\le\mathbf{b}_0$ for all $x\in\mathbf{R}$. Then we have $0\le u(x,t)\le 1$ and $\mathbf{b}_2\le\mathbf{v}(x,t)\le\mathbf{b}_0$ for all $(x,t)\in\mathbf{R}\times\mathbf{R}^+$.*

Therefore, except otherwise stated, we shall assume that the initial data $u(\cdot,0)$ and $v_i(\cdot,0)$, $i=1,\dots,n$, are uniformly Hölder continuous in $\mathbf{R}$ with exponent α for some $\alpha\in(0,1)$ and satisfy that $u(\cdot,0)$ and $v_i(\cdot,0)$ are in $[0,1]$ and $[b_2^i,b_0^i]$, respectively.

Since we will frequently use the notion of supersolutions (subsolutions) of (2.1)–(2.2) and the comparison principle, we state the definition of supersolutions (subsolutions) as follows.

DEFINITION 1. *A set of functions $(u,\mathbf{v})$ is called a subsolution of* (2.1)–(2.2) *in $\mathbf{R}\times\mathbf{R}^+$ if $u,v_i\in C^{2,1}(\mathbf{R}\times\mathbf{R}^+)$ satisfies that $u_t\le L_1[u,\mathbf{v}]$, $v_{i,t}\ge L_{2i}[u,\mathbf{v}]$, that $|u_t-L_1[u,\mathbf{v}]|$, $|v_{i,t}-L_{2i}[u,\mathbf{v}]|$ are bounded on $\mathbf{R}\times\mathbf{R}^+$, and that*

$$-\min_{j=1,\dots,n}\{1,k_-^j/(2k_+^j)\}\le u(x,t)\le 2,\quad b_2^i/2\le v_i(x,t)\le 2b_0^i \tag{2.3}$$

on $\mathbf{R}\times\mathbf{R}^+$ for $i=1,\dots,n$. Supersolution is defined by reversing the inequalities with (2.3) *held.*

The next proposition is a comparison theorem for the system (2.1)–(2.2) and the idea of the proof is based on [35, Lemma 5.3]. We remark that Proposition 2 essentially is a comparison principle for cooperative systems. Note that (2.1)–(2.2) is a competitive type. In general, a competition system with three or more equations cannot be transformed into cooperative ones, but it works for (2.1)–(2.2) due to its special structure.

PROPOSITION 2 (comparison principle). *Let $(u_1,\mathbf{v}_1)$ and $(u_2,\mathbf{v}_2)$ be a subsolution and a supersolution of* (2.1)–(2.2) *on $\mathbf{R}\times\mathbf{R}^+$, respectively, with $u_1(x,0)\le u_2(x,0)$ and $\mathbf{v}_2(x,0)\le\mathbf{v}_1(x,0)$ for all $x\in\mathbf{R}$. Then we have $u_1(x,t)\le u_2(x,t)$ and $\mathbf{v}_2(x,t)\le\mathbf{v}_1(x,t)$ for all $(x,t)\in\mathbf{R}\times\mathbf{R}^+$.*

Proof. Set $\mathbf{v}_1=(v_{11},\dots,v_{1n})$, $\mathbf{v}_2=(v_{21},\dots,v_{2n})$, and $(\tilde u,\tilde{\mathbf{v}})=(u_1-u_2,\mathbf{v}_2-\mathbf{v}_1)$. Then $(\tilde u,\tilde{\mathbf{v}})$ satisfies

$$\begin{aligned} u_t-Du_{xx} &= F(u_1,\mathbf{v}_1)-F(u_2,\mathbf{v}_2)+N_1(z,t)\\ &:=\left[f_1(x,t)\tilde u-\sum_{i=1}^n f_{1i}(x,t)\tilde v_i\right]+N_1(z,t), \end{aligned}\tag{2.4}$$

$$\begin{aligned} v_{i,t} &= G_i(u_2,\mathbf{v}_2)-G_i(u_1,\mathbf{v}_1)+N_{2i}(z,t)\\ &:=[-g_{2i}(x,t)\tilde u+\tilde g_{2i}(x,t)\tilde v_i]+N_{2i}(z,t) \end{aligned}\tag{2.5}$$

for $i=1,\dots,n$, together with the initial data

$$\tilde u(x,0)=u_1(x,0)-u_2(x,0),\ \tilde v_i(x,0)=v_{2i}(x,0)-v_{1i}(x,0),\ i=1,\dots,n,$$

where $N_1(z,t)$ and $N_{2i}(z,t)$ are nonpositive bounded smooth functions on $\mathbf{R}\times[0,+\infty)$,

and

$$
\begin{aligned}
f_1(x,t) &= F_u(\theta_1 u_1 + (1-\theta_1)u_2, \theta_1 \mathbf{v}_1 + (1-\theta_1)\mathbf{v}_2)(x,t),\\
f_{1i}(x,t) &= F_{v_i}(\theta_1 u_1 + (1-\theta_1)u_2, \theta_1 \mathbf{v}_1 + (1-\theta_1)\mathbf{v}_2)(x,t),\\
g_{2i}(x,t) &= G_{i,u}(\theta_{2i} u_1 + (1-\theta_{2i})u_2, \theta_{2i} \mathbf{v}_1 + (1-\theta_{2i})\mathbf{v}_2)(x,t),\\
\tilde{g}_{2i}(x,t) &= G_{i,v_i}(\theta_{2i} u_1 + (1-\theta_{2i})u_2, \theta_{2i} \mathbf{v}_1 + (1-\theta_{2i})\mathbf{v}_2)(x,t),
\end{aligned}
$$

for some $\theta_1 = \theta_1(u_1, \mathbf{v}_1, u_2, \mathbf{v}_2) \in (0,1)$, $\theta_{2i} = \theta_{2i}(u_1, \mathbf{v}_1, u_2, \mathbf{v}_2) \in (0,1)$ and $i = 1, \ldots, n$.

We claim that the region $\{\tilde{u} \le 0, \tilde{v}_i \le 0, i = 1, \ldots, n\}$ is invariant under the flow (2.4)–(2.5). Indeed, from (2.3) and the definitions of F and G_i, it follows that

$$f_{1i}(x,t) < 0 \text{ and } g_{2i}(x,t) < 0 \text{ for all } (x,t) \in \mathbf{R} \times [0, +\infty) \text{ and } i = 1, \ldots, n.$$

Note that $\tilde{u}(x,0) \le 0$ and $\tilde{\mathbf{v}}(x,0) \le \mathbf{0}$ for all $x \in \mathbf{R}$. Therefore, by [32, Theorem 14.11 on p. 203], we have $\tilde{u}(x,t) \le 0$ and $\tilde{\mathbf{v}}(x,t) \le \mathbf{0}$ for all $(x,t) \in \mathbf{R} \times \mathbf{R}^+$. The proof is completed. ◻

3. Asymptotic behavior of traveling wave fronts.

3.1. Several auxiliary lemmas. Before proving our main results, we need some auxiliary lemmas. First, we use the traveling waves $(\mathcal{U}, \mathbf{\Pi})$ to construct supersolutions (subsolutions) of (2.1)–(2.2) whose proof can be found in [35, Lemma 3.7].

LEMMA 3.1. *There exist positive constants d_0, μ_0, and k_{0i}, $i = 1, \ldots, n$, which are independent of $(\mathcal{U}, \mathbf{\Pi}, c)$, and a positive constant ν, which depends on $(\mathcal{U}, \mathbf{\Pi}, c)$, such that, for any $d \in (0, d_0]$ and $\xi_0 \in \mathbf{R}$, the functions $(w^+, \mathbf{p}^+)$ and $(w^-, \mathbf{p}^-)$ defined by*

$$
\begin{aligned}
w^{\pm}(x,t) &:= \mathcal{U}(x - ct + \xi_0 \pm \nu d(1 - e^{-\mu_0 t})) \pm d e^{-\mu_0 t},\\
p_i^{\pm}(x,t) &:= \Pi_i(x - ct + \xi_0 \pm \nu d(1 - e^{-\mu_0 t})) \mp d k_{0i} e^{-\mu_0 t},\ i = 1, \ldots, n,
\end{aligned}
\tag{3.1}
$$

are a supersolution and a subsolution of (2.1)–(2.2)*, respectively. In what follows, we will retain the notation: d_0, μ_0, ν, and k_{0i}.*

Next, we will show that under the assumption of Theorem 1, the solution $(u, \mathbf{v})$ of (2.1)–(2.2) will eventually get trapped between two translates of the traveling wave. For this, we will construct a set of supersolution and subsolution of (2.1)–(2.2). First, we define $\zeta \in C^\infty(\mathbf{R})$ as follows:

$$
\begin{aligned}
&\zeta(s) = 0 \ \text{ if } \ s \le 0, \quad \zeta(s) = 1 \ \text{ if } \ s \ge 2,\\
&0 < \zeta'(s) < 1,\ |\zeta''(s)| \le 1 \ \text{ if } \ s \in (0,2).
\end{aligned}
$$

LEMMA 3.2. *There exist a small positive constant $\hat{d} \in (0, \min\{a/2, (1-a)/2\})$, and functions $\epsilon(\cdot)$, $\sigma(\cdot)$, and $C(\cdot)$ defined on $[0, \hat{d}]$, such that for every $d \in (0, \hat{d}]$ and $\xi \in \mathbf{R}$, the functions $(\hat{w}^+(x,t), \hat{\mathbf{p}}^+(x,t))$ and $(\hat{w}^-(x,t), \hat{\mathbf{p}}^-(x,t))$ defined by*

$$
\begin{aligned}
\hat{w}^+(x,t) = \hat{w}^+(x,t;\xi) &:= (1+d) - [1 - (a - 2d)e^{-\epsilon t}]\zeta(-\epsilon(x - \xi + Ct)),\\
\hat{p}_i^+(x,t) = \hat{p}_i^+(x,t;\xi) &:= \kappa_i(\hat{w}^+(x,t)) - \sigma,\ i = 1, \ldots, n,
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{w}^-(x,t) = \hat{w}^-(x,t;\xi) &:= -d + [1 - (1 - a - 2d)e^{-\epsilon t}]\zeta(\epsilon(x - \xi - Ct)),\\
\hat{p}_i^-(x,t) = \hat{p}_i^-(x,t;\xi) &:= \kappa_i(\hat{w}^-(x,t)) + \sigma,\ i = 1, \ldots, n,
\end{aligned}
$$

are a supersolution and a subsolution of (2.1)–(2.2)*, respectively.*

Proof. Let (and then fix) $\hat{d} \in (0, \min\{a/2, (1-a)/2\})$ be sufficiently small such that $(\hat{w}^+, \hat{\mathbf{p}}^+)$ and $(\hat{w}^-, \hat{\mathbf{p}}^-)$ satisfy (2.3) on $\mathbf{R} \times \mathbf{R}^+$. Next, for each fixed $d \in (0, \hat{d}]$, we will only prove that the conclusion of this lemma holds for $(\hat{w}^+, \hat{\mathbf{p}}^+)$ since the case for $(\hat{w}^-, \hat{\mathbf{p}}^-)$ follows by analogous arguments. Without loss of generality, we may assume that $\xi = 0$. Also by (2.3), there exist positive constants C_1 and C_2, which depend only on k_+^i, k_-^i, b_0^i, and D, such that for all $C > 0$, the inequalities

$$|D[1-(a-2d)e^{-\epsilon t}]\zeta''(-\epsilon(x+Ct))| \le C_1$$

and

$$C_2 \le \sum_{i=1}^n (k_+^i \hat{w}^+(x,t) + k_-^i),\ -\kappa_i'(\hat{w}^+(x,t)) \le C_1$$

hold for all $(x,t) \in \mathbf{R} \times \mathbf{R}^+$. Finally, we set

$$\sigma = \sigma(d) := \min\left\{ d/2, -\frac{f(1+d/2)}{2C_1}, \frac{\min_{u\in[d,a-\frac{d}{2}]}(-f(u))}{2C_1} \right\}.$$

Recalling that $\hat{p}_i^+(x,t) = \kappa_i(\hat{w}^+(x,t)) - \sigma$ and $\kappa_i(u) := k_-^i b_0^i/(k_+^i u + k_-^i)$, we have

$$\begin{aligned}
&k_-^i b_0^i - (k_+^i \hat{w}^+(x,t) + k_-^i)\hat{p}_i^+(x,t) \\
&= \left[k_-^i b_0^i - (k_+^i \hat{w}^+(x,t) + k_-^i)\kappa_i(\hat{w}^+(x,t))\right] + (k_+^i \hat{w}^+(x,t) + k_-^i)\sigma \\
&= (k_+^i \hat{w}^+(x,t) + k_-^i)\sigma
\end{aligned}$$

for $i = 1, \ldots, n$ and $(x,t) \in \mathbf{R} \times \mathbf{R}^+$. Hence for each $i = 1, \ldots, n$, and all $(x,t) \in \mathbf{R} \times \mathbf{R}^+$, we have the following inequalities:

$$\begin{aligned}
\hat{w}_t^+(x,t) - L_1[\hat{w}^+, \hat{\mathbf{p}}^+](x,t) &= C\epsilon[1-(a-2d)e^{-\epsilon t}]\zeta' - \epsilon(a-2d)e^{-\epsilon t}\zeta \\
&\quad + D\epsilon^2[1-(a-2d)e^{-\epsilon t}]\zeta'' \\
&\quad - f(\hat{w}^+(x,t)) - \sum_{i=1}^n \left[k_-^i b_0^i - (k_+^i \hat{w}^+(x,t) + k_-^i)\hat{p}_i^+(x,t)\right] \\
&\ge C\epsilon(1-a)\zeta' - a\epsilon - C_1\epsilon^2 - f(\hat{w}^+(x,t)) \\
&\quad - \sum_{i=1}^n (k_+^i \hat{w}^+(x,t) + k_-^i)\sigma \\
&\ge C\epsilon(1-a)\zeta' - a\epsilon - C_1\epsilon^2 - f(\hat{w}^+(x,t)) - C_1\sigma,
\end{aligned} \tag{3.2}$$

and

$$\begin{aligned}
\hat{p}_{i,t}^+(x,t) - L_{2i}[\hat{w}^+, \hat{\mathbf{p}}^+](x,t) &= \kappa_i'(\hat{w}^+(x,t))\left[C\epsilon(1-(a-2d)e^{-\epsilon t})\zeta' - \epsilon(a-2d)e^{-\epsilon t}\zeta\right] \\
&\quad - \left[k_-^i b_0^i - (k_+^i \hat{w}^+(x,t) + k_-^i)\hat{p}_i^+(x,t)\right] \\
&\le -CC_2\epsilon(1-a)\zeta' + C_1 a\epsilon - C_2\sigma,
\end{aligned} \tag{3.3}$$

where, for simplicity, we have denoted $\zeta(-\epsilon(x-\xi+Ct))$ by ζ. In order to find ϵ and C satisfying that the right-hand sides of (3.2) and (3.3) are nonnegative and nonpositive, respectively, we consider the following three cases:

(i) $\zeta < d/2$,
(ii) $\zeta > 1 - d/2$,
(iii) $\zeta \in [d/2, 1-d/2]$.

For case (i), we have $\hat{w}^+ > 1 + d/2$, and so $-f(\hat{w}^+) > -f(1 + d/2)$. Choose a sufficiently small ϵ satisfying that

$$\epsilon \le \min\left\{d, \frac{C_2\sigma}{C_1 a}\right\} \quad \text{and} \quad -a\epsilon - C_1\epsilon^2 - \frac{f(1 + d/2)}{2} > 0. \tag{3.4}$$

Noting that $\zeta' \ge 0$ and using (3.4) and the definition of σ, it follows that the right-hand sides of (3.2) and (3.3) are nonnegative and nonpositive, respectively.

For case (ii), we have

$$d \le \hat{w}^+ \le (1 + d) - [1 - (a - 2d)](1 - d/2) < a - d/2.$$

Hence if we choose ϵ satisfying both (3.4) and the inequality

$$-a\epsilon - C_1\epsilon^2 + \frac{\min_{u\in[d,a-d/2]}(-f(u))}{2} > 0,$$

then by the definition of σ, the right-hand sides of (3.2) and (3.3) in both cases (i) and (ii) are nonnegative and nonpositive, respectively. In the remaining proof, we fix such ϵ.

Finally, for case (iii), we have a positive lower bound for ζ'. Note that $|f(\hat{w}^+)|$ has an upper bound which depends only on k_+^i, k_-^i, b_0^i, and D. Hence we can choose a sufficiently large $C = C(d, \epsilon)$ to make the right-hand sides of (3.2) and (3.3) nonnegative and nonpositive, respectively. This completes the proof. ◻

With the use of $(\hat{w}^\pm, \hat{\mathbf{p}}^\pm)$, we can obtain the long time behavior of $(u(x,t), \mathbf{v}(x,t))$ for sufficiently large $|x|$, which implies that $(u(\cdot,t), \mathbf{v}(\cdot,t))$ will eventually get trapped between two translates of the traveling wave. We have the following lemma.

Lemma 3.3. *Under the assumption of Theorem* 1, *the solution* $(u, \mathbf{v})$ *of* (2.1)–(2.2) *satisfies that*

$$\lim_{t\to+\infty}\left(\liminf_{x\to+\infty} u(x,t)\right) = 1, \quad \lim_{t\to+\infty}\left(\limsup_{x\to+\infty} v_i(x,t)\right) = b_2^i,$$
$$\lim_{t\to+\infty}\left(\limsup_{x\to-\infty} u(x,t)\right) = 0, \quad \lim_{t\to+\infty}\left(\liminf_{x\to-\infty} v_i(x,t)\right) = b_0^i,\ i = 1, \dots, n.$$

Proof. We only prove that $\lim_{t\to+\infty}(\liminf_{x\to+\infty} u(x,t)) = 1, \lim_{t\to+\infty}(\limsup_{x\to+\infty} v_i(x,t)) = b_2^i$ for $i = 1, \dots, n$, since the remaining part follows in a similar way. Let $\hat{d}$ be defined in Lemma 3.2. By the assumption of Theorem 1, we may choose $\bar{d} \in (0, \hat{d})$ and M_1 such that

$$u(x,0) \ge a + \bar{d} \quad \text{and} \quad v_i(x,0) \le \kappa_i(a) - \frac{k_+^i k_-^i b_0^i}{(k_+^i a + k_-^i)^2}\bar{d} \quad \text{for all} \quad x \ge M_1 \tag{3.5}$$

and $i = 1, \dots, n$. Now for any fixed $d \in (0, \bar{d})$, recall from Lemma 3.2 that $(\hat{w}^-, \hat{\mathbf{p}}^-)$ satisfies

$$\hat{w}^-(x,0) = -d \ \text{ if } \ x \le \xi, \ \ \hat{w}^-(x,0) \le a + d \ \text{ for all } x \in \mathbf{R}, \ \text{ and}$$
$$\hat{p}_i^-(x,0) \ge \kappa_i(-d) \ \text{ if } \ x \le \xi, \ \ \hat{p}_i^-(x,0) \ge \kappa_i(a + d) \ \text{ for all } x \in \mathbf{R}$$

for any $\xi \in \mathbf{R}$ and $i = 1, \dots, n$. Set $\xi = M_1$. *Then we claim that* $u(x,0) \ge \hat{w}^-(x,0)$ *and* $v_i(x,0) \le \hat{p}_i^-(x,0)$ *for all* $x \in \mathbf{R}$. Indeed, it follows from (3.5) that for all $x \ge M_1$

and $i = 1, \ldots, n$, $u(x, 0) \geq \hat{w}^-(x, 0)$ and

$$\begin{aligned} v_i(x,0) &\leq \kappa_i(a) - \frac{k_+^i k_-^i b_0^i}{(k_+^i a + k_-^i)^2} d \\ &\leq \kappa_i(a+d) \\ &\leq \hat{p}_i^-(x,0), \end{aligned}$$

where, in the second inequality, we have used the mean-value theorem and the fact that $\kappa_i'(u) = -k_+^i k_-^i b_0^i/(k_+^i u + k_-^i)^2$. Now we turn to the case $x \leq M_1$. Since $u(\cdot, t) \geq 0$ and $v_i(\cdot, t) \leq b_0^i$, $i = 1, \ldots, n$, it follows from the definition of $(\hat{w}^-, \hat{\mathbf{p}}^-)$ that $u(x, 0) \geq \hat{w}^-(x, 0)$ and $v_i(x, 0) \leq \hat{p}_i^-(x, 0)$ for all $x \leq M_1$. Hence the assertion of this claim follows.

From the definition of $(\hat{w}^-, \hat{\mathbf{p}}^-)$ and the above claim, it follows that for each $i = 1, \ldots, n$, and $d \in (0, \bar{d})$, we have

$$u(x,t) \geq 1 - d - (1 - a - 2d)e^{-\epsilon(d)t}$$

and

$$v_i(x,t) \leq \kappa_i(1 - d - (1 - a - 2d)e^{-\epsilon(d)t}) + \sigma(d)$$

for all $x \geq M_1 + C(d)t + 2/\epsilon(d)$ and $t > 0$. Since d can be taken arbitrary small and $\epsilon(d), \sigma(d) < d$, the assertion of the lemma follows from taking the limit in the above inequalities. This completes the proof. ∎

The following technical lemma is one of the crucial tools, with which we can extend Chen's method [5] to the system (1.3)–(1.4), or, more generally, a system of parabolic equations.

Lemma 3.4. *There exists a positive function $\eta(m)$ defined on $[1, +\infty)$ such that if $(u_1, \mathbf{v}_1)$, $(u_2, \mathbf{v}_2)$ are the supersolution and subsolution of* (2.1)–(2.2)*, respectively, with $u_1(\cdot, 0) \geq u_2(\cdot, 0)$ and $\mathbf{v}_1(\cdot, 0) \leq \mathbf{v}_2(\cdot, 0)$ on $\mathbf{R}$, then for all $m \geq 1$ and $i = 1, \ldots, n$, the following hold:*

$$\min_{x\in[-m,m]} \{u_1(x,1) - u_2(x,1)\} \geq \eta(m) \left[\int_0^1 (u_1(\xi,0) - u_2(\xi,0))d\xi + \sum_{i=1}^n \int_0^1 (v_{2i}(\xi,0) - v_{1i}(\xi,0))d\xi \right],$$

$$\min_{x\in[-m,m]} \{v_{2i}(x,1) - v_{1i}(x,1)\} \geq \eta(m) \left[\int_0^1 (u_1(\xi,0) - u_2(\xi,0))d\xi + \sum_{i=1}^n \int_0^1 (v_{2i}(\xi,0) - v_{1i}(\xi,0))d\xi \right].$$

Proof. First, we set

$$\hat{u} := u_1 - u_2, \quad \hat{\mathbf{v}} = (\hat{v}_1, \ldots, \hat{v}_n) := (v_{21} - v_{11}, \ldots, v_{2n} - v_{1n}),$$

and

$$\begin{aligned} r(x,t) &:= (u_{1,t}(x,t) - L_1[u_1, \mathbf{v}_1](x,t)) - (u_{2,t}(x,t) - L_1[u_2, \mathbf{v}_2](x,t)), \\ s_i(x,t) &:= (v_{2i,t}(x,t) - L_{2i}[u_2, \mathbf{v}_2](x,t)) - (v_{1i,t}(x,t) - L_{2i}[u_1, \mathbf{v}_1](x,t)) \end{aligned}$$

for $i = 1, \ldots, n$. Then $r(x,t)$, $s_i(x,t)$ are nonnegative bounded continuous functions and the following hold:

$$\begin{aligned}\hat{u}_t &= L_1[u_1, \mathbf{v}_1] - L_1[u_2, \mathbf{v}_2] + r(x,t)\\ &= D\hat{u}_{xx} + \hat{b}(x,t)\hat{u} + \sum_{i=1}^{n}[(k_+^i u_2 + k_-^i)\hat{v}_i] + r(x,t)\\ &:= D\hat{u}_{xx} + \hat{b}(x,t)\hat{u} + \sum_{i=1}^{n}(k_-^i \hat{v}_i) + \bar{r}(x,t), \end{aligned}\tag{3.6}$$

$$\hat{v}_{i,t} = -(k_+^i u_2 + k_-^i)\hat{v}_i + (k_+^i \hat{u} v_{1i} + s_i)(x,t),\tag{3.7}$$

where

$$\hat{b}(x,t) := \int_0^1 \frac{df}{du}(\tau u_1(x,t) + (1-\tau)u_2(x,t))d\tau - \sum_{i=1}^{n} k_+^i v_{1i}(x,t)$$

and $\bar{r}(x,t)$ is a nonnegative bounded continuous function. Note that $\hat{u} \geq 0$ and $\hat{\mathbf{v}} \geq \mathbf{0}$ on $\mathbf{R} \times \mathbf{R}^+$, by Proposition 2.

Now we will prove that the conclusion holds for $u_1 - u_2$. Indeed, by the definitions of supersolution and subsolution of (2.1)–(2.2), there exists a positive constant M, which is independent of u_1, u_2, and $\mathbf{v}_1$, such that for all $(x,t) \in \mathbf{R} \times [0, +\infty)$, we have

$$|\hat{b}(x,t)| \leq M.$$

Let $\bar{u}(x,t) := e^{Mt}\hat{u}(x,t)$ on $\mathbf{R} \times [0, +\infty)$. Then $\bar{u}$ satisfies the following equation:

$$\bar{u}_t = D\bar{u}_{xx} + (M + \hat{b}(x,t))\bar{u} + e^{Mt}\sum_{i=1}^{n}(k_-^i \hat{v}_i) + e^{Mt}\bar{r}(x,t),$$

and so $\bar{u}$ can be represented by

$$\begin{aligned}\bar{u}(x,t) = &\int_{\mathbf{R}} \Gamma(x,t,\xi,0)\bar{u}(\xi,0)d\xi\\ &+ \int_0^t \int_{\mathbf{R}} \Gamma(x,t,\xi,\tau)\left[e^{M\tau}\sum_{i=1}^{n}(k_-^i \hat{v}_i(\xi,\tau))\right] d\xi d\tau\\ &+ \int_0^t \int_{\mathbf{R}} \Gamma(x,t,\xi,\tau)\left[(M + \hat{b}(\xi,\tau))\bar{u}(\xi,\tau) + e^{M\tau}\bar{r}(\xi,\tau)\right] d\xi d\tau,\end{aligned}\tag{3.8}$$

where

$$\Gamma(x,t,\xi,\tau) = \frac{\exp[-\frac{(x-\xi)^2}{4D(t-\tau)}]}{2\sqrt{\pi D(t-\tau)}}.$$

Solving (3.7), we obtain

$$\begin{aligned}\hat{v}_i(x,t) &= \hat{v}_i(x,0)\exp\left[-\int_0^t (k_+^i u_2(x,\sigma) + k_-^i)d\sigma\right]\\ &\quad + \int_0^t \exp\left[-\int_\sigma^t (k_+^i u_2(x,\tau) + k_-^i)d\tau\right](k_+^i \hat{u} v_{1i} + s_i)(x,\sigma)d\sigma\\ &\geq \hat{v}_i(x,0)\exp\left[-\int_0^t (k_+^i u_2(x,\sigma) + k_-^i)d\sigma\right]\\ &\geq \hat{v}_i(x,0)\exp[-\tilde{C}_1 t],\end{aligned}\tag{3.9}$$

where $\tilde{C}_1$ is a positive constant depending only on k_+^i, k_-^i, and b_0^i. Note that by the definitions of M, $\hat{b}$, and $\bar{r}$, the third term in (3.8) is nonnegative. Combining this with (3.8)–(3.9), there exist positive constants C_1 and $\hat{C}_1$, which depend only on D, k_+^i, k_-^i, and b_0^i, such that for all $(x,t) \in \mathbf{R} \times \mathbf{R}^+$, the following holds:

$$
\begin{aligned}
\bar{u}(x,t) \geq\ & \frac{C_1}{\sqrt{t}} \int_{\mathbf{R}} \exp\Big[-\hat{C}_1 \frac{(x-\xi)^2}{t} \Big] \bar{u}(\xi,0) d\xi \\
& + C_1 \sum_{i=1}^n \int_0^t \int_{\mathbf{R}} \frac{\exp(-\hat{C}_1 \frac{(x-\xi)^2}{t-\tau})}{\sqrt{t-\tau}} \Big[\exp(M\tau) \hat{v}_i(\xi,\tau) \Big] d\xi d\tau \\
\geq\ & \frac{C_1}{\sqrt{t}} \int_{\mathbf{R}} \exp\Big[-\hat{C}_1 \frac{(x-\xi)^2}{t} \Big] \bar{u}(\xi,0) d\xi \\
& + C_1 \sum_{i=1}^n \int_0^t \int_{\mathbf{R}} \frac{\exp(-\hat{C}_1 \frac{(x-\xi)^2}{t-\tau})}{\sqrt{t-\tau}} \Big[\exp((M-\tilde{C}_1)\tau) \hat{v}_i(\xi,0) \Big] d\xi d\tau.
\end{aligned}
$$

Hence for each $m \geq 1$ and $t > 0$, we have

(3.10)

$$
\begin{aligned}
\min_{x \in [-m,m]} \hat{u}(x,t) \geq\ & C_1 \frac{\exp[-Mt - \hat{C}_1 (m+1)^2/t]}{\sqrt{t}} \int_0^1 \hat{u}(\xi,0) d\xi \\
& + C_1 \sum_{i=1}^n \left(\int_0^t \frac{\exp[-Mt - \hat{C}_1 \frac{(m+1)^2}{t-\tau} + (M-\tilde{C}_1)\tau]}{\sqrt{t-\tau}} d\tau \right) \int_0^1 \hat{v}_i(\xi,0) d\xi \\
:=\ & A(t) \int_0^1 \hat{u}(\xi,0) d\xi + B(t) \sum_{i=1}^n \int_0^1 \hat{v}_i(\xi,0) d\xi.
\end{aligned}
$$

Finally, by setting

$$\eta_1(m) := \min\{A(1), B(1)\}$$

and using (3.10), our conclusion for $u_1 - u_2$ follows. However, we will replace $\eta_1(m)$ by a smaller quantity, $\eta(m)$, later.

Next, we will show that the conclusion holds for $v_2 - v_1$. Indeed, by the definitions of supersolution and subsolution of (2.1)–(2.2), there exist positive constants C_2 and $\hat{C}_2$, which depend only on k_+^i, k_-^i, and b_0^i, such that for each $i = 1, \ldots, n$, the following holds:

$$k_+^i u_2 + k_-^i \leq \hat{C}_2, \quad k_+^i v_{1i} \geq C_2 \quad \text{on} \quad \mathbf{R} \times [0, +\infty).$$

Using this inequality and the fact that $\hat{v}_i, s_i \geq 0$ on $\mathbf{R} \times [0, +\infty)$, and solving (3.7), we obtain that for $t = 1$ and $x \in [-m, m]$ with $m \geq 1$, there holds

$$
\begin{aligned}
\hat{v}_i(x,1) =\ & \hat{v}_i(x,0) \exp\left[-\int_0^1 (k_+^i u_2(x,\sigma) + k_-^i) d\sigma \right] \\
& + \int_0^1 \exp\left[-\int_\sigma^1 (k_+^i u_2(x,\tau) + k_-^i) d\tau \right] (k_+^i \hat{u} v_{1i} + s_i)(x,\sigma) d\sigma \\
\geq\ & \int_0^1 \exp\left[-\int_\sigma^1 (k_+^i u_2(x,\tau) + k_-^i) d\tau \right] (k_+^i \hat{u}(x,\sigma) v_{1i}(x,\sigma)) d\sigma \\
\geq\ & C_2 \int_0^1 \exp\left[-\int_\sigma^1 \hat{C}_2 d\tau \right] A(\sigma) d\sigma \cdot \int_0^1 \hat{u}(\xi,0) d\xi \\
& + C_2 \int_0^1 \exp\left[-\int_\sigma^1 \hat{C}_2 d\tau \right] B(\sigma) d\sigma \cdot \sum_{i=1}^n \int_0^1 \hat{v}_i(\xi,0) d\xi,
\end{aligned}
$$

where, in the last inequality, we have used (3.10). Finally, by setting

$$\eta(m) := \min\left\{\eta_1(m), C_2\int_0^1 \exp\left[-\int_\sigma^1 \hat{C}_2 d\tau\right] A(\sigma)d\sigma, C_2\int_0^1 \exp\left[-\int_\sigma^1 \hat{C}_2 d\tau\right] B(\sigma)d\sigma\right\},$$

and using the above inequality and (3.10), the assertion of this lemma follows. This completes the proof. ∎

The following lemma is the main tool for the iteration process in the proof of Theorem 1. We will modify the method of [5, Lemma 3.3] to prove it (see also [36]).

LEMMA 3.5. *Suppose that the assumptions of Theorem* 1 *hold. Then for each* $h_0 > 0$*, there exists a small positive constant* $\epsilon^* = \epsilon^*(h_0) \in (0, \min\{1/(2\nu), d_0/2\})$*, which is independent of* $(u(\cdot,0), \mathbf{v}(\cdot,0))$*, such that if, for some* $\tau \geq 0$*,* $\xi \in \mathbf{R}$*,* $d \in (0, \min\{1/\nu, d_0/2\}]$*, and* $h \in [0, h_0]$*, the following holds:*

$$\mathcal{U}(x - c\tau + \xi) - d \leq u(x,t) \leq \mathcal{U}(x - c\tau + \xi + h) + d, \tag{3.11}$$

$$\Pi_i(x - c\tau + \xi + h) - dk_{0i} \leq v_i(x,t) \leq \Pi_i(x - c\tau + \xi) + dk_{0i} \tag{3.12}$$

for all $x \in \mathbf{R}$ *and* $i = 1, \ldots, n$*, then for every* $t > \tau + 1$*, there exist* $\hat{\xi}(t)$*,* $\hat{d}(t)$*, and* $\hat{h}(t)$ *satisfying*

$$\begin{aligned}
&\hat{\xi}(t) \in [\xi - \nu d, \xi + \min\{h, 1\} + \nu d]\\
&\hat{d}(t) \leq e^{-\mu_0(t-\tau-1)}[d + \epsilon^*(h_0)\min\{h, 1\}],\\
&\hat{h}(t) \leq [h - \nu\epsilon^*(h_0)\min\{h, 1\}] + 2\nu d,
\end{aligned}$$

such that (3.11)–(3.12) *hold with* (τ, ξ, d, h) *replaced by* $(t, \hat{\xi}(t), \hat{d}(t), \hat{h}(t))$*. In the remainder of this paper, we will retain the notation* $\epsilon^*(h_0)$*.*

Proof. If necessary, by a shift of time, we may assume that $\tau = 0$. Now by comparing $(u, \mathbf{v})$ with $(w^\pm, \mathbf{p}^\pm)$ in Lemma 3.1 (with $\xi_0 = \xi$ for $(w^-, \mathbf{p}^-)$ and $\xi_0 = \xi + h$ for $(w^+, \mathbf{p}^+)$), we have that for all $(x,t) \in \mathbf{R} \times \mathbf{R}^+$ and $i = 1, \ldots, n$, there holds

$$\begin{aligned}
\mathcal{U}(x - ct + \xi - \nu d(1 - e^{-\mu_0 t})) - de^{-\mu_0 t} &\leq u(x,t)\\
&\leq \mathcal{U}(x - ct + \xi + h + \nu d(1 - e^{-\mu_0 t})) + de^{-\mu_0 t},
\end{aligned} \tag{3.13}$$

$$\begin{aligned}
\Pi_i(x - ct + \xi + h + \nu d(1 - e^{-\mu_0 t})) - dk_{0i}e^{-\mu_0 t} &\leq v_i(x,t)\\
&\leq \Pi_i(x - ct + \xi - \nu d(1 - e^{-\mu_0 t})) + dk_{0i}e^{-\mu_0 t}.
\end{aligned} \tag{3.14}$$

Fix a number x_0 with the property

$$0 \leq x_0 + \xi < \frac{1}{2}.$$

Since $\lim_{|s|\to\infty}(\mathcal{U}'(s), \mathbf{\Pi}'(s)) = (0, \mathbf{0})$, we can fix a sufficiently large constant $M > h_0$ such that for all $|s| \geq M$ and $i = 1, \ldots, n$, we have

$$\mathcal{U}'(s) \leq \frac{1}{2\nu} \quad \text{and} \quad -\Pi_i'(s) \leq \frac{\min\{1, k_{0i}\}}{2\nu}. \tag{3.15}$$

Set

$$\bar{h} = \min\{h, 1\} \ \text{ and } \ \epsilon_1 := \frac{(n+1)}{2} \min_{s\in[0,3], 1\leq i\leq n} \left\{\mathcal{U}'(s), -\Pi_i'(s)\right\}.$$

Then by using the mean-value theorem, we have

$$\int_0^1 \Big(\mathcal{U}(y+x_0+\xi+\bar{h}) - \mathcal{U}(y+x_0+\xi)\Big)dy + \sum_{i=1}^n \int_0^1 \Big(\Pi_i(y+x_0+\xi) - \Pi_i(y+x_0+\xi+\bar{h})\Big)dy \geq 2\epsilon_1\bar{h},$$

and so, at least one of the following holds:

$$\text{(i)} \int_0^1 \Big(\mathcal{U}(y+x_0+\xi+\bar{h}) - u(y+x_0,0)\Big)dy + \sum_{i=1}^n \int_0^1 \Big(v_i(y+x_0,0) - \Pi_i(y+x_0+\xi+\bar{h})\Big)dy \geq \epsilon_1\bar{h},$$

$$\text{(ii)} \int_0^1 \Big(u(y+x_0,0) - \mathcal{U}(y+x_0+\xi)\Big)dy + \sum_{i=1}^n \int_0^1 \Big(\Pi_i(y+x_0+\xi) - v_i(y+x_0,0)\Big)dy \geq \epsilon_1\bar{h}.$$

In what follows, we will consider case (i) since case (ii) follows by analogous arguments. Set $\xi_1 := c - \nu d(1-e^{-\mu_0})$. Note that $|\xi_1| \leq |c|+1$, since $\nu d \leq 1$. By using (3.13)–(3.14) and Lemma 3.4, we obtain that for $\eta = \eta(2(M+|c|+1))$, every x with $x+x_0+\xi+h-\xi_1 \in [-2(M+|c|+1), 2(M+|c|+1)]$, and each $i = 1,\ldots,n$, there holds

$$\begin{aligned}
&[\mathcal{U}(x+x_0+\xi+h-\xi_1) + de^{-\mu_0}] - u(x+x_0,1) \\
&\geq \eta\left[\int_0^1 \Big([\mathcal{U}(y+x_0+\xi+\bar{h}) + d] - u(y+x_0,0)\Big)dy \right.\\
&\quad \left. + \sum_{i=1}^n \int_0^1 \Big(v_i(y+x_0,0) - [\Pi_i(y+x_0+\xi+\bar{h}) - dk_{0i}]\Big)dy\right] \\
&\geq \eta\epsilon_1\bar{h}, \\
&v_i(x+x_0,1) - [\Pi_i(x+x_0+\xi+h-\xi_1) - dk_{0i}e^{-\mu_0}] \\
&\geq \eta\left[\int_0^1 \Big([\mathcal{U}(y+x_0+\xi+\bar{h}) + d] - u(y+x_0,0)\Big)dy \right.\\
&\quad \left. + \sum_{i=1}^n \int_0^1 \Big(v_i(y+x_0,0) - [\Pi_i(y+x_0+\xi+\bar{h}) - dk_{0i}]\Big)dy\right] \\
&\geq \eta\epsilon_1\bar{h},
\end{aligned} \tag{3.16}$$

where we have used the facts that $\mathcal{U}$, $-\Pi_i$ are monotone increasing and $h \geq \bar{h}$. Define

$$\epsilon^* = \epsilon^*(h_0) := \min_{i=1,\ldots,n}\Big\{\frac{d_0}{2}, \frac{1}{2\nu}, \min_{|x|\leq 2(M+|c|+1)} \frac{\eta\epsilon_1}{2\nu\mathcal{U}'(x)}, \min_{|x|\leq 2(M+|c|+1)} \frac{\eta\epsilon_1}{-2\nu\Pi_i'(x)}\Big\}.$$

Let the set A_1 defined by

$$A_1 = \{x \in \mathbf{R} \mid -2(M+|c|+1) \leq x+x_0+\xi-\xi_1 \leq (M+|c|+1)\}.$$

Recall that $h \in (0, M)$ and $2\nu\epsilon^*(h_0) \in (0,1)$. Therefore for all $x \in A_1$ and $\theta \in (0,1)$, there holds

$$|x + x_0 + \xi + h - \xi_1| \leq 2(M + |c| + 1),$$
$$|x + x_0 + \xi + h - \xi_1 - 2\nu\theta\epsilon^*(h_0)\bar{h})| \leq 2(M + |c| + 1).$$

Hence by the mean-value theorem and the choice of $\epsilon^*(h_0)$, we can choose $\theta_i \in (0,1)$, $i = 0, 1, \dots, n$, such that for all $x \in A_1$ and $i = 1, \dots, n$, the following holds:

$$\begin{aligned}
&\mathcal{U}(x + x_0 + \xi + h - \xi_1 - 2\nu\epsilon^*(h_0)\bar{h}) - \mathcal{U}(x + x_0 + \xi + h - \xi_1)\\
&= \mathcal{U}'(x + x_0 + \xi + h - \xi_1 - 2\theta_0\nu\epsilon^*(h_0)\bar{h})(-2\nu\epsilon^*(h_0)\bar{h})\\
&\geq -\eta\epsilon_1\bar{h},\\
&\Pi_i(x + x_0 + \xi + h - \xi_1 - 2\nu\epsilon^*(h_0)\bar{h}) - \Pi_i(x + x_0 + \xi + h - \xi_1)\\
&= \Pi_i'(x + x_0 + \xi + h - \xi_1 - 2\theta_i\nu\epsilon^*(h_0)\bar{h})(-2\nu\epsilon^*(h_0)\bar{h})\\
&\leq \eta\epsilon_1\bar{h}.
\end{aligned}$$

Combining these two inequalities with (3.16), it follows that for all $x \in A_1$ and $i = 1, \dots, n$, we have

$$\begin{aligned}
u(x + x_0, 1) &\leq \mathcal{U}(x + x_0 + \xi + h - \xi_1 - 2\nu\epsilon^*(h_0)\bar{h}) + de^{-\mu_0},\\
v_i(x + x_0, 1) &\geq \Pi_i(x + x_0 + \xi + h - \xi_1 - 2\nu\epsilon^*(h_0)\bar{h}) - dk_{0i}e^{-\mu_0}.
\end{aligned} \tag{3.17}$$

Next, we turn to consider the set $A_2 := \mathbf{R} \setminus A_1$. By using the properties of h_0 and $\epsilon^*(h_0)$ again, we can conclude that for all $x \in A_2$ and $\theta \in (0,1)$, there holds

$$|x + x_0 + \xi + h - \xi_1| \geq M + |c| + 1,$$
$$|x + x_0 + \xi + h - \xi_1 - 2\nu\theta\epsilon^*(h_0)\bar{h})| \geq M + |c| + 1.$$

Therefore by using the mean-value theorem, the last inequalities, (3.15), and the second inequality of (3.13) and the first inequality of (3.14) with $t = 1$, we can conclude that there exist $\theta_i \in (0,1)$, $i = 0, 1, \dots, n$, such that for all $x \in A_2$ and $i = 1, \dots, n$, there holds

$$\begin{aligned}
&\mathcal{U}(x + x_0 + \xi + h - \xi_1 - 2\nu\epsilon^*(h_0)\bar{h})\\
&\quad = \mathcal{U}(x + x_0 + \xi + h - \xi_1)\\
&\quad\quad + \mathcal{U}'(x + x_0 + \xi + h - \xi_1 - 2\theta_0\nu\epsilon^*(h_0)\bar{h})(-2\nu\epsilon^*(h_0)\bar{h})\\
&\quad \geq \mathcal{U}(x + x_0 + \xi + h - \xi_1) - \epsilon^*(h_0)\bar{h}\\
&\quad \geq u(x + x_0, 1) - [de^{-\mu_0} + \epsilon^*(h_0)\bar{h}],\\
&\Pi_i(x + x_0 + \xi + h - \xi_1 - 2\nu\epsilon^*(h_0)\bar{h})\\
&\quad = \Pi(x + x_0 + \xi + h - \xi_1)\\
&\quad\quad + \Pi'(x + x_0 + \xi + h - \xi_1 - 2\theta_i\nu\epsilon^*(h_0)\bar{h})(-2\nu\epsilon^*(h_0)\bar{h})\\
&\quad \leq \Pi(x + x_0 + \xi + h - \xi_1) + \epsilon^*(h_0)\bar{h}k_{0i}\\
&\quad \leq v(x + x_0, 1) + [de^{-\mu_0} + \epsilon^*(h_0)\bar{h}]k_{0i},
\end{aligned}$$

which together with (3.17), implies that for all $x \in \mathbf{R}$ and $i = 1, \dots, n$, we have

$$\begin{aligned}
u(x, 1) &\leq \mathcal{U}(x + \xi + h - \xi_1 - 2\nu\epsilon^*(h_0)\bar{h}) + [de^{-\mu_0} + \epsilon^*(h_0)\bar{h}],\\
v_i(x, 1) &\geq \Pi_i(x + \xi + h - \xi_1 - 2\nu\epsilon^*(h_0)\bar{h}) - [de^{-\mu_0} + \epsilon^*(h_0)\bar{h}]k_{0i}.
\end{aligned}$$

Now set $q := de^{-\mu_0} + \epsilon^*(h_0)\bar{h}$. Note that q is not bigger than d_0 by the choice of $\epsilon^*(h_0)$. Also recall that $\xi_1 = c - \nu d(1 - e^{-\mu_0})$. Hence by comparing $(u(x, 1 + t'), \mathbf{v}(x, 1 + t'))$ with $(\mathcal{U}(\xi_2) + qe^{-\mu_0 t'}, \mathbf{\Pi}(\xi_2) - \mathbf{k}_0 qe^{-\mu_0 t'})$ (with $\xi_2 = x - ct' + \xi + h - \xi_1 - 2\nu\epsilon^*(h_0)\bar{h} + \nu q(1 - e^{-\mu_0 t'})$ and $\mathbf{k}_0 = (k_{01}, \ldots, k_{0n})$), we can conclude that for all $(x, t') \in \mathbf{R} \times \mathbf{R}^+$ and $i = 1, \ldots, n$, there holds

$$\begin{aligned} u(x, 1 + t') &\leq \mathcal{U}(x - ct' + \xi + h - \xi_1 - 2\nu\epsilon^*(h_0)\bar{h} + \nu q(1 - e^{-\mu_0 t'})) + qe^{-\mu_0 t'} \\ &\leq \mathcal{U}(x - c(1 + t') + \xi - \nu d + [h - \nu\epsilon^*(h_0)\bar{h} + 2\nu d]) + (d + \epsilon^*(h_0)\bar{h})e^{-\mu_0 t'} \end{aligned}$$

and

$$\begin{aligned} &v_i(x, 1 + t') \\ &\quad \geq \Pi_i(x - ct' + \xi + h - \xi_1 - 2\nu\epsilon^*(h_0)\bar{h} + \nu q(1 - e^{-\mu_0 t'})) - qk_{0i}e^{-\mu_0 t'} \\ &\quad \geq \Pi_i(x - c(1 + t') + \xi - \nu d + [h - \nu\epsilon^*(h_0)\bar{h} + 2\nu d]) - (d + \epsilon^*(h_0)\bar{h})k_{0i}e^{-\mu_0 t'}. \end{aligned}$$

Set

$$t = 1 + t', \ \hat{\xi}(t) = \xi - \nu d, \ \hat{h}(t) = h - \nu\epsilon^*(h_0)\bar{h} + 2\nu d, \ \text{and} \ \hat{d}(t) = (d + \epsilon^*(h_0)\bar{h})e^{-\mu_0(t-1)}.$$

It then follows that for all $(x, t) \in \mathbf{R} \times \mathbf{R}^+$ with $t > 1$ and $i = 1, \ldots, n$, we have

$$\begin{aligned} u(x, t) &\leq \mathcal{U}(x - ct + \hat{\xi}(t) + \hat{h}(t)) + \hat{d}(t), \\ v_i(x, t) &\geq \Pi_i(x - ct + \hat{\xi}(t) + \hat{h}(t)) - \hat{d}(t)k_{0i}. \end{aligned} \tag{3.18}$$

Finally, it remains to be seen whether the first inequality of (3.11) and the second inequality of (3.12) hold with (τ, ξ, d, h) replaced by $(t, \hat{\xi}(t), \hat{d}(t), \hat{h}(t))$. In fact, by using (3.13)–(3.14) and the monotonicity of $(\mathcal{U}, \mathbf{\Pi})$, we can conclude that for all $(x, t) \in \mathbf{R} \times \mathbf{R}^+$ and $i = 1, \ldots, n$, there holds

$$\begin{aligned} u(x, t) &\geq \mathcal{U}(x - ct + \xi - \nu d(1 - e^{-\mu_0 t})) - de^{-\mu_0 t} \\ &\geq \mathcal{U}(x - ct + \xi - \nu d) - (d + \epsilon^*(h_0)\bar{h})e^{-\mu_0(t-1)} \\ &= \mathcal{U}(x - ct + \hat{\xi}(t)) - \hat{d}(t), \end{aligned} \tag{3.19}$$

and

$$\begin{aligned} v_i(x, t) &\leq \Pi_i(x - ct + \xi - \nu d(1 - e^{-\mu_0 t})) + dk_{0i}e^{-\mu_0 t} \\ &\leq \Pi_i(x - ct + \xi - \nu d) + (d + \epsilon^*(h_0)\bar{h})k_{0i}e^{-\mu_0(t-1)} \\ &= \Pi_i(x - ct + \hat{\xi}(t)) + \hat{d}(t)k_{0i}. \end{aligned} \tag{3.20}$$

Therefore (3.18), (3.19), and (3.20) imply the assertion of this lemma. The proof is completed. □

3.2. The proof of Theorem 1. Now we have every necessary tool at hand to prove our main theorem. Note that it may take a long time for the solution $(u, \mathbf{v})$ of (2.1)–(2.2) with the initial data satisfying the condition of Theorem 1 to be trapped between two translates of the wave whose difference of phase shifts is less than 1. Therefore, the proof will consist of two main steps. First, we will use the squeezing method [5, pp. 139–140] with minor modification to prove that the assertion of Theorem 1 holds, but the rate constant of convergence $\widehat{\kappa}$ may depend on the initial data $(u(\cdot, 0), \mathbf{v}(\cdot, 0))$ (see Lemma 3.6). Next, we will show that the exponent constant

$\widehat{\kappa}$ can be replaced with a universal constant κ. We remark that the presentation of this subsection is similar to that in [36] where we study a single quasilinear diffusion-advection equation. But for completeness and reader's convenience, we would like to include it here.

LEMMA 3.6. *Under the assumption of Theorem* 1, *there exists a positive constant* $\widehat{\kappa}$ *which may depend on the initial data* $(u(\cdot,0),\mathbf{v}(\cdot,0))$ *such that*

$$\sup_{x\in\mathbf{R}}|u(x,t)-\mathcal{U}(x-ct+\xi^*)|\le Ke^{-\widehat{\kappa}t},\quad \sup_{x\in\mathbf{R}}|v_i(x,t)-\Pi_i(x-ct+\xi^*)|\le Ke^{-\widehat{\kappa}t},$$

for $i=1,\dots,n$, *and for some constants* ξ^* *and* K *which may depend on* $(u(\cdot,0),\mathbf{v}(\cdot,0))$.

Proof. The proof consists of four steps.

Step 1. Define the quantity

$$\tilde{d}^*:=\min\{1/(2\nu),d_0/2\}.$$

For this $\tilde{d}^*$, by Lemma 3.3 and the fact that $(\mathcal{U}(+\infty),\mathbf{\Pi}(+\infty))=(1,\mathbf{b}_2)$, $(\mathcal{U}(-\infty),\mathbf{\Pi}(-\infty))=(0,\mathbf{b}_0)$, there exist sufficiently large constants $h_0>2$ and $\tilde{T}_0$ such that the following holds:

$$\mathcal{U}(x-c\tilde{T}_0-h_0/4)-\tilde{d}^*\le u(x,\tilde{T}_0)\le\mathcal{U}(x-c\tilde{T}_0-h_0/4+h_0/2)+\tilde{d}^*, \tag{3.21}$$
$$\Pi(x-c\tilde{T}_0-h_0/4+h_0/2)-\tilde{d}^*k_{0i}\le v(x,\tilde{T}_0)\le\Pi(x-c\tilde{T}_0-h_0/4)+\tilde{d}^*k_{0i} \tag{3.22}$$

for all $x\in\mathbf{R}$ and $i=1,\dots,n$. Therefore (3.11)–(3.12) holds with $\tau=\tilde{T}_0$, $\xi=-h_0/4$, $h=h_0/2$, and $d=\tilde{d}^*$. Let $\epsilon^*(h_0)$ be defined as in the proof of Lemma 3.5, and set

$$d_0^*:=\epsilon^*(h_0)/4,\quad \kappa_0^*:=\nu\epsilon^*(h_0)/2<1/4.$$

Note that $\epsilon^*(h_0)\le\tilde{d}^*$ by the definition of $\epsilon^*(h_0)$. Therefore we can fix a sufficiently large constant $t_0^*\ge 2$ with the property that

$$e^{-\mu_0(t_0^*-1)}(\tilde{d}^*+\epsilon^*(h_0))\le d_0^*\ \text{ and }\ e^{-\mu_0(t_0^*-1)}(1+\epsilon^*(h_0)/d_0^*)\le 1-\kappa_0^*.$$

Now by (3.21)–(3.22) and Lemma 3.5, we can conclude that (3.11)–(3.12) hold for $\tau=\tilde{T}_0+t_0^*$, some $\hat{\xi}\in[-h_0/4-\nu\tilde{d}^*,-h_0/4+1+\nu\tilde{d}^*]$, $h=\hat{h}(\tilde{T}_0+t_0^*)$, and $d=\hat{d}(\tilde{T}_0+t_0^*)$. By the associated definitions, we can estimate $\hat{h}(\tilde{T}_0+t_0^*)$ and $\hat{d}(\tilde{T}_0+t_0^*)$ as follows:

$$\begin{aligned}\hat{h}(\tilde{T}_0+t_0^*)&\le\frac{h_0}{2}-\nu\epsilon^*(h_0)+2\nu\tilde{d}^*<\frac{h_0}{2}+1<h_0,\\ \hat{d}(\tilde{T}_0+t_0^*)&\le e^{-\mu_0(t_0^*-1)}(\tilde{d}^*+\epsilon^*(h_0))\le d_0^*.\end{aligned}$$

Together with the monotonicity of $(\mathcal{U},\mathbf{\Pi})$, the above discussion leads to

$$\mathcal{U}(x-c(\tilde{T}_0+t_0^*)+\hat{\xi})-d_0^*\le u(x,\tilde{T}_0+t_0^*)\le\mathcal{U}(x-c(\tilde{T}_0+t_0^*)+\hat{\xi}+h_0)+d_0^*, \tag{3.23}$$
$$\begin{aligned}\Pi_i(x-c(\tilde{T}_0+t_0^*)+\hat{\xi}+h_0)-d_0^*k_{0i}&\le v_i(x,\tilde{T}_0+t_0^*)\\ &\le\Pi_i(x-c(\tilde{T}_0+t_0^*)+\hat{\xi})+d_0^*k_{0i}\end{aligned} \tag{3.24}$$

for all $x\in\mathbf{R}$ and $i=1,\dots,n$. Therefore (3.11)–(3.12) hold for $\tau=T_0:=\tilde{T}_0+t_0^*$, some $\hat{\xi}\in[-h_0/4-\nu\tilde{d}^*,-h_0/4+1+\nu\tilde{d}^*]$, $h=h_0$, and $d=d_0^*$. Although the remainder of the proof is almost identical to [5, pp. 139–140], for completeness and the later use, we include it here.

Step 2. *Now we claim that there exists a finite time $T_1 > T_0$ such that* (3.11)–(3.12) *hold for $\tau = T_1$, $h = 1$, $d = d_0^*$, and some $\xi \in \mathbf{R}$.* Recall that $h_0 > 2$. Let N be the unique positive integer such that $h_0 \in [1 + (N-1)\kappa_0^*, 1 + N\kappa_0^*)$. Now by (3.23)–(3.24) and Lemma 3.5, it follows that (3.11)–(3.12) hold for $\tau = T_0 + t_0^*$, some $\xi \in [\hat{\xi} - \nu d_0^*, \hat{\xi} + 1 + \nu d_0^*]$, $h = \hat{h}(T_0 + t_0^*)$, and $d = \hat{d}(T_0 + t_0^*)$. As before, $\hat{h}(T_0 + t_0^*)$ and $\hat{d}(T_0 + t_0^*)$ can be estimated as follows:

$$\hat{h}(T_0+t_0^*) \le h_0 - \nu\epsilon^*(h_0) + 2\nu d_0^* = h_0 - \kappa_0^*, \quad \hat{d}(T_0+t_0^*) \le e^{-\mu_0(t_0^*-1)}(d_0^* + \epsilon^*(h_0)) \le d_0^*.$$

Combined with the monotonicity of $(\mathcal{U}, \mathbf{\Pi})$, it follows that (3.11)–(3.12) hold for $\tau = T_0 + t_0^*$, some $\xi \in [\hat{\xi} - \nu d_0^*, \hat{\xi} + 1 + \nu d_0^*]$, $h = h_0 - \kappa_0^*$, and $d = d_0^*$. Repeating the same process N times, it yields that (3.11)–(3.12) hold for $\tau = T_1 := T_0 + Nt_0^*$, some $\xi^0 \in \mathbf{R}$, $h = h_0 - N\kappa_0^*$, and $d = d_0^*$. Finally, by the monotonicity of $(\mathcal{U}, \mathbf{\Pi})$ again, the assertion of this claim follows.

Step 3. *Now we claim that for each $k \in \mathbf{N} \cup \{0\}$,* (3.11)–(3.12) *hold for some $\xi = \xi^k \in \mathbf{R}$ and*

$$\tau = T^k := T_1 + kt_0^*, \quad h = h^k := (1 - \kappa_0^*)^k, \quad d = d^k := (1-\kappa_0^*)^k d_0^*.$$

Moreover, we have

$$|\xi^{k+1} - \xi^k| \le (1 + \nu d_0^*)(1-\kappa_0^*)^k \text{ for all } k \in \mathbf{N} \cup \{0\}. \tag{3.25}$$

Indeed, *Step* 2 implies that the claim holds for $k = 0$. Now we assume that the claim holds for some $k = m \ge 0$. Applying Lemma 3.5 with $(\tau, \xi, h, d) = (T^m, \xi^m, h^m, d^m)$, we then obtain that (3.11)–(3.12) hold for $(\tau, \xi, h, d) = (T^m + t_0^*, \hat{\xi}(T^m + t_0^*), \hat{h}(T^m + t_0^*), \hat{d}(T^m + t_0^*))$ which satisfies the following:

$$\begin{aligned}
\xi^{m+1} &:= \hat{\xi}(T^m + t_0^*) \in [\xi^m - \nu d^m, \xi^m + \nu d^m + h^m],\\
\hat{h}(T^m + t_0^*) &\le h^m - \nu\epsilon^*(h_0)h^m + 2\nu d^m\\
&= (1-\kappa_0^*)^{m+1},\\
\hat{d}(T^m + t_0^*) &\le e^{-\mu_0(t_0^*-1)}(d^m + \epsilon^*(h_0)h^m)\\
&= (1-\kappa_0^*)^m d_0^* e^{-\mu_0(t_0^*-1)}(1 + \epsilon^*(h_0)/d_0^*)\\
&\le (1-\kappa_0^*)^{m+1} d_0^*.
\end{aligned}$$

Using this and the monotonicity of $(\mathcal{U}, \mathbf{\Pi})$, we can conclude that (3.11)–(3.12) hold for $(\tau, \xi, h, d) = (T^{m+1}, \xi^{m+1}, h^{m+1}, d^{m+1})$. This completes the proof of the first part of this claim.

Finally, noting that $|\xi^{k+1} - \xi^k| \le \nu d^k + h^k$ and using the definitions of h^k and d^k, (3.25) follows.

Step 4. *Finally, we will prove the assertion of the lemma.* For each $k \in \mathbf{N} \cup \{0\}$, recall that (3.11)–(3.12) hold for $(\tau, \xi, h, d) = (T^k, \xi^k, h^k, d^k)$. Hence from (3.13)–(3.14) and the monotonicity of $(\mathcal{U}, \mathbf{\Pi})$, it follows that (3.11)–(3.12) hold for $\tau \in [T^k, \infty)$, $d = d^k$, $h = h^k + 2\nu d^k$, and $\xi = \xi^k - \nu d^k$. Hence, if we define $\xi(t) = \xi^k - \nu d^k$, $h(t) = h^k + 2\nu d^k$, and $d(t) = d^k$ for all $t \in [T^k, T^{k+1})$, then it follows that for all $(x, t) \in \mathbf{R} \times [T_1, +\infty)$ and $i = 1, \ldots, n$, there holds

$$\mathcal{U}(x - ct + \xi(t)) - d(t) \le u(x,t) \le \mathcal{U}(x - ct + \xi(t) + h(t)) + d(t), \tag{3.26}$$

$$\Pi_i(x - ct + \xi(t) + h(t)) - d(t)k_{0i} \le v_i(x,t) \le \Pi_i(x - ct + \xi(t)) + d(t)k_{0i}. \tag{3.27}$$

Next, for each $t \in [T_1, +\infty)$, there exists a unique nonnegative integer $m = m(t)$ which is the largest integer not greater than $(t - T_1)/t_0^*$. Then we have $t \in [T^m, T^{m+1})$, and from the definitions of $h(t)$ and $d(t)$, it follows that for all $t \geq T_1$, there holds

$$h(t) = h^m + 2\nu d^m \leq (1 + 2\nu d_0^*) \exp\left[\Big(\frac{t - T_1}{t_0^*} - 1\Big) \ln(1 - \kappa_0^*)\right], \tag{3.28}$$

$$d(t) = d^m \leq d_0^* \exp\left[\Big(\frac{t - T_1}{t_0^*} - 1\Big) \ln(1 - \kappa_0^*)\right]. \tag{3.29}$$

Now we will show that $\xi(\infty) := \lim_{t \to +\infty} \xi(t)$ exists. Indeed, for any $\tau \geq t \geq T_1$, there exist nonnegative integers l and p such that $\tau \in [T^l, T^{l+1})$ and $t \in [T^p, T^{p+1})$. Then by the definitions of $\xi(t)$ and (3.25), we can estimate $|\xi(\tau) - \xi(t)|$ as follows:

$$\begin{aligned}
|\xi(\tau) - \xi(t)| &= |(\xi^l - \nu d^l) - (\xi^p - \nu d^p)| \\
&\leq \sum_{i=p}^{l-1} |\xi^{i+1} - \xi^i| + \nu |d^l - d^p| \\
&= \Big(\frac{1 + \nu d_0^*}{\kappa_0^*} + \nu d_0^*\Big) |(1 - \kappa_0^*)^p - (1 - \kappa_0^*)^l)|,
\end{aligned}$$

which tends to zero as $\tau, t \to +\infty$. Hence $\xi(\infty)$ exists. Moreover, we have

$$|\xi(t) - \xi(\infty)| \leq \Big(\frac{1 + \nu d_0^*}{\kappa_0^*} + \nu d_0^*\Big) \exp\left[\Big(\frac{t - T_1}{t_0^*} - 1\Big) \ln(1 - \kappa_0^*)\right] \tag{3.30}$$

for all $t \geq T_1$. Now set

$$\widehat{\kappa} := -\ln(1 - \kappa_0^*)/t_0^*,$$

which may depend on the initial data $(u(\cdot, 0), \mathbf{v}(\cdot, 0))$. Finally, combining (3.26)–(3.30) with the mean-value theorem, we obtain the assertion of this lemma. This completes the proof. $\square$

Final Proof of Theorem 1. With this preparation, we are ready to complete the proof of Theorem 1. Note that it remains to be seen whether the exponent constant $\hat{\kappa}$ can be replaced with a universal constant κ.

First, we define the following two quantities:

$$d^* := \frac{\epsilon^*(1)}{4} \quad \text{and} \quad \kappa^* := \frac{\nu \epsilon^*(1)}{2} < \frac{1}{4},$$

where $\epsilon^*(1)$ is defined in the proof of Lemma 3.5. We then fix a sufficiently large constant $t^* \geq 2$ with the property that

$$e^{-\mu_0 (t^* - 1)} (1 + \epsilon^*(1)/d^*) \leq 1 - \kappa^*.$$

Let ξ^* be defined in Lemma 3.6. Next, by Lemma 3.6, there exists a positive constant T such that there holds

$$\begin{aligned}
&\mathcal{U}(x - cT + \xi^*) - d^* \leq u(x, T) \leq \mathcal{U}(x - cT + \xi^*) + d^*, \\
&\Pi_i(x - cT + \xi^*) - d^* k_{0i} \leq v_i(x, T) \leq \Pi_i(x - cT + \xi^*) + d^* k_{0i}
\end{aligned}$$

for all $x \in \mathbf{R}$ and $i = 1, \ldots, n$. Define

$$\kappa := -\ln(1 - \kappa^*)/t^*,$$

which is independent of the initial data $(u(\cdot,0),\mathbf{v}(\cdot,0))$. Finally, by using a similar argument of Step 3–Step 4 in the proof of Lemma 3.6, we can conclude that

$$\sup_{x\in\mathbf{R}}|u(x,t)-\mathcal{U}(x-ct+\xi^*)|\le Ke^{-\kappa t},\quad \sup_{x\in\mathbf{R}}|v_i(x,t)-\Pi_i(x-ct+\xi^*)|\le Ke^{-\kappa t}$$

hold for all $t>0$ and $i=1,\dots,n$, and for some constant K. This completes the proof. ∎

Acknowledgements. This work was initiated during the visit at the Mathematical Biosciences Institute of The Ohio State University. The author thanks Professor Xinfu Chen for his valuable discussion and careful reading of the manuscript. The author also thanks Professor James Sneyd for his valuable discussion. The author would like to thank the referees for helpful comments on this paper.

REFERENCES

[1] A. Atri, J. Amundson, D. Clapham, and J. Sneyd, *A single pool model for intracellular calcium oscillations and waves in the Xenopus laevis oocyte*, Biophys. J., 65 (1993), pp. 1727–1739.
[2] M. J. Berridge, *Inositol trisphosphate and calcium signalling*, Nature, 361 (1993), pp. 315–325.
[3] I. Bezprozvanny, J. Watras, and B. E. Ehrlich, *Bell-shaped calcium response curves of* $\mathrm{Ins}(1,4,5)\mathrm{P}_3$*- and calcium-gated channels from endoplasmic reticulum of cerebellum*, Nature, 351 (1991), pp. 751–754.
[4] W. B. Busa, J. E. Ferguson, S. K. Joseph, J. R. Williamson, and R. Nuccitelli, *Activation of frog (Xenopus laevis) eggs by inositaltrisphosphate* I: *Characterization of* Ca^{2+} *release from intracellular stores*, J. Cell Biol., 101 (1985), pp. 677–682.
[5] X. F. Chen, *Existence, uniqueness, and asymptotic stability of traveling waves in nonlocal evolution equations*, Adv. Differential Equations, 2 (1997), pp. 125–160.
[6] G. Dupont and A. Goldbeter, *Properties of intracellular* Ca^{2+} *waves generated by a model based on* Ca^{2+}*-induced* Ca^{2+} *release*, Biophys. J., 67 (1994), pp. 2191–2204.
[7] J. Evans and N. Shenk, *Solutions to axon equations*, Biophys. J., 10 (1970), pp. 1090–1101.
[8] M. Falcke, *Buffers and oscillations in intracellular* Ca^{2+} *dynamics*, Biophys. J., 84 (2003), pp. 28–41.
[9] J. E. Ferrel, Jr., and E. M. Machleder, *The biochemical basis of an all-or-none cell fate switch in Xenpous oocytes*, Science, 280 (1998), pp. 895–898.
[10] P. C. Fife and J. B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Ration. Mech. Anal., 65 (1977), pp. 335–361.
[11] R. A. Fontanilla and R. Nuccitelli, *Characterization of the sperm-induced calcium wave in Xenopus eggs using confocal microscopy*, Biophys. J., 75 (1998), pp. 2079–2087.
[12] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[13] S. Girard, A. Luckhoff, J. Lechleiter, J. Sneyd, and D. Clapham, *Two-dimensional model of calcium wves reproduces the patterns observed in Xenopus laevis oocyte*, Biophys. J., 61 (1992), pp. 509–517.
[14] J.-S. Guo and J.-C. Tsai, *The asymptotic behavior of solutions of the buffered bistable system*, J. Math. Biol., 53 (2006), pp. 179–213.
[15] M. S. Jafri and J. Keizer, *On the roles of* Ca^{2+} *diffusion,* Ca^{2+} *buffers and the endoplasmic reticulum in* IP_3*-induced* Ca^{2+} *waves*, Biophys. J., 69 (1995), pp. 2139–2153.
[16] E. J. Kaftan, B. E. Ehrlich, and J. Watras, *Inositol* $1,4,5$*-trisphosphate* (InsP_3) *and calcium interact to increase the dynamic range of* InsP_3 *receptor-dependent calcium signaling*, J. Gen. Physiol., 110 (1997), pp. 529–538.
[17] J. Keener and J. Sneyd, *Mathematical Physiology*, Springer-Verlag, New York, 1998.
[18] Y.-X. Li and J. Rinzel, *Equations for* InsP_3 *receptor-mediated* $[\mathrm{Ca}^{2+}]_i$ *oscillations derived from a detailed kinetic model: A Hodgkin-Huxley like formalism*, J. Theor. Biol., 166 (1994), pp. 461–473.
[19] I. C. B. Marshall and C. W. Taylor, *Regulation of inositol* $1,4,5$*-trisphosphate receptors*, J. Exp. Biol., 184 (1993), pp. 161–182.

[20] M. Naraghi and E. Neher, *Linearized buffered Ca^{2+} diffusion in microdomains and its implications for calculation of [Ca^{2+}] at the mouth of a calcium channel*, J. Neurosci., 17 (1997), pp. 6961–6973.
[21] M. Naraghi, T. H. Muller, and E. Neher, *Two-dimensional determination of the cellular Ca^{2+} binding in bovine chromaffin cells*, Biophys. J., 75 (1998), pp. 1635–1647.
[22] E. Neher, *Usefulness and limitations of linear approximations to the understanding of Ca^{2+} signals*, Cell Calcium, 24 (1998), pp. 345–57.
[23] M. C. Nowycky and M. J. Pinter, *Time courses of calcium and calcium-bound buffers following calcium influx in a model cell*, Biophys. J., 64 (1993), pp. 77–91.
[24] R. Nuccitelli, D. L. Yim, and T. Smart, *The sperm-induced* Ca^{2+} *wave following fertilization of the Xenopus egg requires the production of* $\mathrm{Ins}(1,4,5)\mathrm{P}_3$, Dev. Biol., 158 (1993), pp. 200–212.
[25] J.-M. Roquejoffre, D. Terman, and V. A. Volpert, *Global stability of traveling fronts and convergence towards stacked families of waves in monotone parabolic systems*, SIAM J. Math. Anal., 27 (1996), pp. 1261–1269.
[26] T. A. Rooney and A. P. Thomas, *Intracellular calcium waves generated by $Ins(1,4,5)P_3$ dependent mechanisms*, Cell Calcium, 14 (1993), pp. 674–690.
[27] F. Sala and A. Hernández-Cruz, *Calcium diffusion modeling in a spherical neuron: relevance of buffering properties*, Biophys. J., 57 (1990), pp. 313–324.
[28] B. M. Slepchenko, J. C. Schaff, and Y. S. Choi, *Numerical approach to fast reactions in reaction-diffusion systems: application to buffered calcium waves in bistable model*, J. Comput. Phys., 162 (2000), pp. 186–218.
[29] G. D. Smith, *Analytical steady-state solution to the rapid buffering approximation near an open Ca^{2+} channel*, Biophys. J., 71 (1996), pp. 3064–3072.
[30] G. D. Smith, L. Dai, R. M. Miura, and A. Sherman, *Asymptotic analysis of buffered calcium diffusion near a point source*, SIAM J. Appl. Math., 61 (2001), pp. 1816–1838.
[31] G. D. Smith, J. E. Pearson, and J. Keizer, *Modeling Intracellular Calcium Waves and Sparks*, in Computatiional Cell Biology, C. P. Fall, E. S. Marland, J. M. Wagner, and J. J. Tyson, eds., Springer-Verlag, New York, 2002, pp. 198–229.
[32] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Springer-Verlag, New York, 1994.
[33] J. Sneyd, J. Keizer, and M. J. Sanderson, *Mechanisms of calcium oscillations and waves: a quantitative analysis*, FASEB J., 9 (1995), pp. 1463–1472.
[34] J. Sneyd, P. D. Dale, and A. Duffy, *Traveling waves in buffered systems: applications to calcium waves*, SIAM J. Appl. Math., 58 (1998), pp. 1178–1192.
[35] J.-C. Tsai and J. Sneyd, *Existence and stability of traveling waves in buffered systems*, SIAM J. Appl. Math., 66 (2005), pp. 237–265.
[36] J.-C. Tsai and J. Sneyd, *Are Buffers Boring? Uniqueness and asymptotical stability of traveling wave fronts in the buffered bistable system*, J. Math. Biol., 54 (2007), pp. 513–553.
[37] A. I. Volpert, V. A. Volpert, and V. A. Volpert, *Traveling-wave solutions of parabolic systems*, Translations of Mathematical Monographs 140, AMS, Providence, RI, 1994.
[38] J. Wagner and J. Keizer, *Effects of rapid buffers on* Ca^{2+} *diffusion and* Ca^{2+} *oscillations*, Biophys. J., 67 (1994), pp. 447–456.
[39] J. Wagner, Y.-X. Li, J. Pearson, and J. Keizer, *Simulation of the fertilization* Ca^{2+} *wave in Xenopus laevis eggs*, Biophys. J., 75 (1998), pp. 2088–2097.
[40] D. Xu and X.-Q. Zhao, *Bistable waves in an epidemic model*, J. Dynam. Differential Equations, 16 (2004), pp. 679–707. Erratum, J. Dynam. Differential Equations, 17 (2005), pp. 219–247.
[41] G. De Young and J. Keizer, *A single-pool inositol* $1,4,5$*-trisphosphate-receptor-based model for agonist-stimulated oscillations in* Ca^{2+} *concentration*, Proc. Natl. Acad. Sci. USA, 89 (1992), pp. 9895–9899.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 160–173

EXISTENCE OF SOLUTIONS FOR SUPPLY CHAIN MODELS BASED ON PARTIAL DIFFERENTIAL EQUATIONS*

M. HERTY†, A. KLAR†, AND B. PICCOLI‡

Abstract. We consider a model for supply chains governed by partial differential equations. The mathematical properties of a continuous model are discussed and existence and uniqueness are proven. Moreover, Lipschitz continuous dependence on the initial data is proven. We make use of the front tracking method to construct approximate solutions. The obtained results extend the preliminary work of [S. Göttlich, M. Herty, and A. Klar, *Commun. Math. Sci.*, 3 (2005), pp. 545–559].

Key words. supply chains, networks, front tracking

AMS subject classifications. 90B10, 65Mxx

DOI. 10.1137/060659478

1. Introduction.

Supply chain modeling is characterized by different mathematical approaches: on one hand, there are discrete event simulations based on considerations of individual parts; on the other hand, continuous models like [1, 2, 3] using partial differential equations have been introduced. We consider supply chain modeling based on the latter—the continuous models. Recently those models based on scalar conservation laws have been reformulated in the framework of network models where the dynamics on the arcs is governed by a partial differential equation; see [12]. This approach is inspired by other recent discussions on networks; see, e.g., [4, 8, 13, 14].

We recall the basic supply chain model under consideration: a supply chain network consists of connected suppliers which are going to process parts. Further, each supplier consists of a processor for assembling and construction and a buffer for unprocessed parts, called a queue. We have the following definition.

Definition 1.1 (network definition). *A supply chain network is a finite, connected, directed graph consisting of a finite set of arcs $\mathcal{J}$ and a finite set of vertices $\mathcal{V}$. Each supplier j is modeled by an arc j, which is again parameterized by an interval $[a_j, b_j]$.*

Each processor is characterized by a maximum processing capacity μ_j, its length L_j, and the processing time T_j. The rate L_j/T_j describes the processing velocity and we assume for simplicity that $L_j/T_j = 1$ for all j. To model the evolution of parts inside the processor we introduce the function $\rho_j(x,t)$, i.e., the density of parts in processor j at point x and time t. Now, the dynamics of each processor on an arc j are governed by an advection equation as in [2]:

$$\partial_t \rho_j(x,t) + \partial_x \min\left\{\mu_j, \frac{L_j}{T_j}\rho_j(x,t)\right\} = 0 \quad \forall x \in [a_j, b_j],\ t \in \mathbb{R}^+, \tag{1.1a}$$

$$\rho_j(x,0) = \rho_{j,0}(x) \quad \forall x \in [a_j, b_j]. \tag{1.1b}$$

*Received by the editors May 10, 2006; accepted for publication (in revised form) November 7, 2006; published electronically April 27, 2007.
http://www.siam.org/journals/sima/39-1/65947.html

†Fachbereich Mathematik, Technische Universität Kaiserslautern, Kaiserslautern, Germany (herty@rhrk.uni-kl.de, klar@itwm.fhg.de).
‡Istituto per Applicazioni del Calcolo "M. Picone," Rome, Italy (b.piccoli@iac.cnr.it).

Equation (1.1) can be derived from a discrete event simulation [2] and allows for the following interpretation: The parts are processed with velocity L_j/T_j but with a maximal flux of μ_j. The dynamical behavior of the queues is discussed in detail in the following sections. Roughly speaking, if the inflow is greater than the maximum possible outflow, then the queue increases proportionally to the difference of the two, while it decreases in the opposite case.

First, in section 2, we consider a chain-like network geometry as in Figure 1 for which the discussion below simplifies. Then, in the following section, we turn to the situation of arbitrary networks—in particular, those with vertices having more than two connected arcs.

Our main achievement is the extension of results proposed in [12]. The correct space to be considered is that of couples (ρ_j, q_j): density of parts and queue buffer occupancy. We prove existence and uniqueness of weak solutions for a general network of supply chains and bounded variation (BV) initial data. The densities ρ_j are Lipschitz continuous in time w.r.t. the L^1 metric, while the queue's buffer occupancies q_j are absolutely continuous.

Moreover, we prove Lipschitz continuous dependence on the initial data. This, in turn, permits us to extend the corresponding semigroup trajectories to L^∞ initial data.

The main idea of the proof is to construct approximated solutions by wave front tracking [5] and derive bounds on the total variation by a careful estimate of the interactions at the vertices of the network. The proof of Lipschitz dependence uses the approach as in [6].

2. Consecutive processors. In this section we recall the supply chain network model introduced and investigated in [12] and extend the existence results obtained therein.

First, we consider the case where each vertex is connected to exactly to one incoming arc and one outgoing arc and we assume that the arcs are consecutively labeled, i.e., arc j is connected to arc $j+1$, and that $b_j = a_{j-1}$; see also Figure 1.

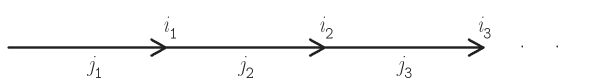


FIG. 1. *Example of a simple network structure.*

2.1. Modeling and previous results. As in the introduction, the supplier j is defined by a queue j and a processor j. Physically, the queue is located in front of each processor, i.e., at $x = a_j$. To avoid technical difficulties, we assume that the first supplier consists of a processor only and the last has infinite length, so $a_1 = -\infty$ and $b_N = +\infty$ for the first and, respectively, the last supplier in the supply chain.

In addition to (1.1), the queue buffer occupancy in front of each processor is modeled as time-dependent function $t \to q_j(t)$. If the capacity of processor $j-1$ and the demand of processor j are not equal, the queue q_j increases or decreases its buffer. Mathematically, this implies that each queue q_j satisfies the following equation:

$$\partial_t q_j(t) = f_{j-1}(\rho_{j-1}(b_{j-1}, t)) - f_j(\rho_j(a_j, t)), \quad j = 2, \ldots, N. \tag{2.1}$$

Last, a reasonable mathematical condition for the boundary values for outgoing arcs

j is given by (see [12])

$$f_j(\rho_j(a_j,t)) = \begin{cases} \min\{f_{j-1}(\rho_{j-1}(b_{j-1},t)),\mu_j\}, & q_j(t)=0, \\ \mu_j, & q_j(t)>0. \end{cases} \tag{2.2}$$

This allows for the following interpretation: If the outgoing buffer is empty, we process as many parts as possible but at most μ_j. If the buffering queue contains parts, then we process at the maximal possible rate, namely, again μ_j. Finally, the supply chain model is a coupled system of partial and ordinary differential equations on a network given by (1.1), (2.1), and (2.2).

We recall some preliminary facts from [12].

Note that due to the very special flux function

$$f_j(x) := \min\{\mu_j; L_j/T_j\rho\}, \tag{2.3}$$

a Riemann problem for (1.1) and $(x,t)\in\mathbb{R}\times\mathbb{R}^+$ admits one of the following two solutions. Let $\rho_{j,0}(x)=\rho_l$ for $x<0$ and $\rho_{j,0}=\rho_r$ for $x\geq 0$. If $\rho_l<\rho_r$, then the solution ρ_j is given by

$$\rho_j(x,t) = \begin{cases} \rho_l, & -\infty < \frac{x}{t} \leq \frac{f_j(\rho_r)-f_j(\rho_l)}{\rho_r-\rho_l}, \\ \rho_r, & \frac{f_j(\rho_r)-f_j(\rho_l)}{\rho_r-\rho_l} < \frac{x}{t} < \infty. \end{cases} \tag{2.4}$$

If, on the contrary, $\rho_l>\rho_r$, then the following happens. If either $\rho_l\leq\mu_j$ or $\rho_r\geq\mu_j$, then the solution is given by (2.4). Otherwise (i.e., if $\rho_r<\mu_j<\rho_l$), we obtain the solution given by

$$\rho(x,t) = \begin{cases} \rho_l, & -\infty < \frac{x}{t} \leq \frac{f_j(\rho_l)-\mu_j}{\rho_l-\mu_j}, \\ \mu_j, & \frac{f_j(\rho_l)-\mu_j}{\rho_l-\mu_j} < \frac{x}{t} \leq \frac{\mu_j-f_j(\rho_r)}{\mu_j-\rho_r}, \\ \rho_r, & \frac{\mu_j-f_j(\rho_r)}{\mu_j-\rho_r} < \frac{x}{t} < \infty. \end{cases} \tag{2.5}$$

Notice that the right-hand side (RHS) of the first two inequalities is always 0 or 1.

We can introduce the following definition.

Definition 2.1 (network solution). *A family of functions $\{\rho_j,q_j\}_{j\in\mathcal{J}}$ is called an* admissible *solution for a network as in Figure* 1 *if, for all j, ρ_j is a weak entropic solution* [16] *to* (1.1)*, q_j is absolutely continuous, and, in the sense of traces for ρ_js,* (2.1) *and* (2.2) *hold for a.e. t.*

For the particular situation of a single vertex $v\in\mathcal{V}$ with incoming arc $j=1$ and outgoing arc $j=2$ and constant initial data $\rho_{j,0}(x)\leq\mu_j$, there exists an admissible solution $\{\rho_1,\rho_2,q_2\}$. The solution has the explicit form

$$\rho_1(x,t)=\rho_{1,0}, \tag{2.6a}$$

$$\rho_2(x,t) = \begin{cases} f_1(\rho_{1,0})<\mu_2 & \begin{cases} \rho_{1,0}, & 0\leq (x-t_0)/t<1=\frac{f_2(\mu_2)-f_2(\rho_{1,0})}{\mu_2-\rho_{1,0}}, \\ \mu_2, & 1\leq (x-t_0)/t \text{ and } x/t<1, \\ \rho_{2,0}, & 1\leq x/t<\infty, \end{cases} \\ f_1(\rho_{1,0})\geq\mu_2 & \begin{cases} \mu_2, & 0\leq x/t<1=\frac{f_2(\mu_2)-f_2(\rho_{2,0})}{\mu_2-\rho_{2,0}}, \\ \rho_{2,0}, & 1\leq x/t<\infty, \end{cases} \end{cases} \tag{2.6b}$$

$$q_2(t)=q_{2,0}+\int_0^t f_1(\rho_{1,0})-f_2(\rho_2(a_2+,\tau))d\tau, \tag{2.6c}$$

wherein $t_0 = q_{2,0}/(\mu_2 - f_1(\rho_{1,0}))$. For a network as in Figure 1, for initial data $\{\rho_{j,0}(x)\}_j$ where each $\rho_{j,0}$ is a step function, and for initial values $q_j(0) = 0$, there exists an admissible solution $\{\rho_j, q_j\}_j$ to the network problem (1.1), (2.1), (2.2); see [12]. The construction of the solution is based on wave or front tracking (see below and in [9, 5, 15]). For applications of this method in the context of network problems we also refer the reader to [14, 8].

2.2. Wave front tracking approximations. To start, we introduce a equidistant grid $(i\delta)_{i0}^{N_x}$ such that $0 \leq (i\delta) \leq \max\{\mu_j : j \in \mathcal{J}\}$ and such that $\forall j \exists i_j : i\delta\mu_j$. Here, it is implicitly assumed that μ_i/μ_j is rational. We approximate the initial data by step functions $\rho_{j,0}^\delta$ taking values in the set $\{i\delta : i0, \ldots, N_x\}$. Then each Riemann problem inside an arc or at a vertex is solved, obtaining various traveling discontinuities. If discontinuities collide, then the collision can be resolved either by solving a Riemann problem inside the arc j (see (2.4),(2.5)) or as a collision with a vertex (see (2.6)). In both cases we obtain new discontinuities propagating until the next collision.

At the same time, an evolution of the queue's buffers q_j is automatically defined when solving the Riemann problems at vertices.

This construction guarantees that the solution on arcs takes values only in the set $\{i\delta : i = 0, \ldots, N_x\}$ and we obtain a wave front tracking approximate solution (denoted by $(\rho^\delta, q^\delta) := \{(\rho_j^\delta, q_j^\delta)\}_j$) consisting of a set of moving discontinuities along the intervals $[a_j, b_j]$ and the queue's buffers' evolutions.

As usual [5, 10], to guarantee the good definition of wave front tracking approximate solutions and, passing to the limit, prove existence of solutions in the sense of Definition 2.1, three basic estimates are in order:

1. the estimate on the number of waves;
2. the estimate on the number of interactions (between waves and of waves with queues); and
3. the estimate on total variation of solutions for ρ_j.

Moreover, in our case, we need to prove some compactness of the sequence q_j^δ in an appropriate space.

It is easy to check that every collision inside an arc decreases the number of waves, while the interactions with a vertex may produce two new waves; cf. (2.6). Also, since the characteristic velocity of waves is always positive and is bounded from above, then the first two estimates are readily obtained; see [12]. Therefore the construction of wave front tracking approximations is well-defined up to any given time T.

2.3. Total variation estimates on densities. Here, we provide total variation estimates on ρ_j^δ (i.e., along wave front tracking approximate solutions). This will imply the existence of an admissible solution for BV initial data $\rho_{j,0}$.

First, we discuss the case of initial data $\rho_{j,0}$ additionally satisfying the following assumption (K).

(K) For every j the initial datum satisfies $\rho_{j,0} \leq \mu_j$.

The above construction guarantees that (K) remains valid for every time along wave front tracking approximate solutions.

Each $\rho_j^\delta(x,t)$ is a piecewise constant function in x and thus will define a number of constant states $\rho_{j,i}^\delta, i = 1, \ldots, N_j$, where we assume that $\rho_j^\delta(a_j, \cdot) = \rho_{j,1}^\delta$, and so forth. We define the total variation of the flux on the network as

$$T.V.(f(\rho^\delta)) = \sum_{j\in\mathcal{J}} T.V.(f_j(\rho_j^\delta(\cdot,t))) = \sum_{j\in\mathcal{J}} \sum_{i1}^{N_j-1} |f_j(\rho_{j,i}^\delta) - f_j(\rho_{j,i+1}^\delta)|. \tag{2.7}$$

Note that, thanks to assumption (K), a bound on $T.V.(f_j(\rho_j^\delta(\cdot,t)))$ provides also a bound on $T.V.(\rho_j^\delta(\cdot,t))$, since $\rho_{j,i}^\delta \leq \mu_j$ for all j,i. Furthermore, $T.V.(f(\rho^\delta))$ does not increase when discontinuities collide *inside* an arc j; see [5]. Next, we discuss the collision of a discontinuity with a vertex.

LEMMA 2.2. *Assume a single vertex with incoming arc $j=1$ and outgoing arc $j=2$. Furthermore, assume constant states $\rho_{j,0}$, $j=1,2$, at the vertex and consider a discontinuity colliding at time t_0. Denote the new solution at the vertex after the collision by $\bar\rho_j$. Assume no more collision of discontinuities happens until t^*. Then, for all $t_0 < t < t^*$,*

$$\sum_{j=1}^{2} T.V.(f_j(\rho_j(\cdot,t))) + |\partial_t q_2(t)| \leq \sum_{j=1}^{2} T.V.(f_j(\rho_j(\cdot,t_0))) + |\partial_t q_2(t_0)|. \tag{2.8}$$

Proof. By construction the colliding discontinuity has to arrive from arc $j=1$, and therefore the total variation of the flux on this arc decreases by $|f_1(\bar\rho_1) - f_1(\rho_{1,0})|$. On the outgoing arc $j=2$ we distinguish two cases. First, assume that $f_2(\rho_{2,0}) = f_1(\rho_{1,0})$. Then, due to (2.1) we have $\partial_t q_2(t_0) = 0$. If $f_1(\bar\rho_1) \leq \mu_2$, then $f_2(\bar\rho_2) = f_1(\bar\rho_1)$ and (2.8) holds. If, on the other hand, $f_1(\bar\rho_1) > \mu_2$, then due to (2.2), $f_2(\bar\rho_2) = \mu_2$ and again (2.8) holds, since for $t > t_0$

$$\begin{aligned} |f_1(\bar\rho_1) - f_1(\rho_{1,0})| &= |\mu_2 - f_1(\rho_{1,0})| + |f_1(\bar\rho_1) - \mu_2| \\ &= |f_2(\bar\rho_2) - f_2(\rho_{2,0})| + |\partial_t q_2(t)|. \end{aligned}$$

In the second case, we assume $f_2(\rho_{2,0}) = \mu_2$. Then,

$$|\partial_t q_2(t_0)| = |f_1(\rho_{1,0}) - \mu_2|$$

and we distinguish two more subcases depending on whether the queue is increasing or decreasing after the collisions. First, assume $f_1(\bar\rho_1) \geq \mu_2$, i.e., the queue q_2 is increasing with

$$|\partial_t q_2(t)| = f_1(\bar\rho_1) - \mu_2$$

and

$$f_2(\bar\rho_2) = f_2(\rho_{2,0}) = \mu_2. \tag{2.9}$$

Inequality (2.8) still holds, since

$$|f_1(\rho_{1,0}) - f_1(\bar\rho_1)| + |\partial_t q_2(t_0)| \geq |\partial_t q_2(t)|$$

for $t > t_0$. Second, assume $f_1(\bar\rho_1) < \mu_2$, i.e., the queue q_2 is decreasing. Let $\bar t$ be such that $q_2(\bar t) = 0$. Then (2.8) holds since for $t < \min\{\bar t, t^*\}$

$$T.V.(f_2(\bar\rho_2(\cdot,t))) = 0$$

and

$$\begin{aligned} |f_1(\bar\rho_1) - f_1(\rho_{1,0})| + |f_1(\rho_{1,0}) - \mu_2| &\geq |\mu_2 - f_1(\bar\rho_1)| \\ &= |\partial_t q_2(t)|. \end{aligned}$$

If $\bar{t} < t^*$ we obtain a new traveling discontinuity on the outgoing arc $j = 2$ for times $t > \bar{t}$ when the queue q_2 becomes empty: $f_2(\bar{\rho}_2(a_2+, t)) = f_1(\bar{\rho}_1)$ and $\partial_t q(t) = 0$ for $t > \bar{t}$. Then, (2.8) still holds, since

$$T.V.(f_2(\rho_2(\cdot, t))) + |\partial_t q_2(t)|$$

is constant for this interaction. This finishes the proof. ∎

Summarizing, we conclude that for all $\delta > 0$ the following holds for all $t > 0$:

$$\sum_{j=1}^{N} T.V.(\rho_j^\delta(\cdot, t)) + \sum_{j=2}^{N} |\partial_t q_j^\delta(t)| \le \sum_{j=1}^{N} T.V.(\rho_{j,0}^\delta(\cdot)) + \sum_{j=2}^{N} |\partial_t q_j^\delta(0)| \tag{2.10a}$$

$$\text{and } \rho_j^\delta(x, t) \le \max_j \mu_j \ \forall j, x. \tag{2.10b}$$

2.4. Total variation estimates on queues buffers. Let us now estimate the total variation of $\partial_t q_j$.

LEMMA 2.3. *Assume we have a single vertex with incoming arc $j = 1$ and outgoing arc $j = 2$ (of infinite length). Furthermore, assume we have constant states $\rho_{j,0}, j = 1, 2$, at the vertex and consider a discontinuity collision at time t_0. Denote the new solution at the vertex after the collision by $\bar{\rho}_j$. Assume no more collisions of discontinuities happen until t^*. Then, for all $t_0 < t < t^*$,*

$$T.V.(\partial_t q_2, [t_0, t]) \le 2\ |f_1(\bar{\rho}_1) - f_1(\rho_{1,0})| + |\partial_t q_2(t_0)|. \tag{2.11}$$

Proof. The interactions are clearly the same examined in Lemma 2.2.

First, assume that $f_2(\rho_{2,0}) = f_1(\rho_{1,0})$. Then, due to (2.1) we have $\partial_t q_2(t_0) = 0$. If $f_1(\bar{\rho}_1) \le \mu_2$, then $f_2(\bar{\rho}_2) = f_1(\bar{\rho}_1)$ and $\partial_t q_2(t) = 0$; thus (2.11) holds because the left-hand side vanishes.

If on the other hand, $f_1(\bar{\rho}_1) > \mu_2$, then

$$|f_1(\bar{\rho}_1) - f_1(\rho_{1,0})| = |f_2(\bar{\rho}_2) - f_2(\rho_{2,0})| + |\partial_t q_2(t)| \ge |\partial_t q_2(t)|;$$

thus (2.11) holds because $\partial_t q_2(t_0) = 0$.

In the second case, we assume $f_2(\rho_{2,0}) = \mu_2$. Then,

$$\partial_t q_2(t_0) = f_1(\rho_{1,0}) - \mu_2, \quad \partial_t q_2(t_0+) = f_1(\bar{\rho}_1) - \mu_2. \tag{2.12}$$

If the queue is increasing after the interaction, then

$$T.V.(\partial_t q_2(t), [t_0, t]) = |f_1(\rho_{1,0}) - \mu_2 - (f_1(\bar{\rho}_1) - \mu_2)| = |f_1(\rho_{1,0}) - f_1(\bar{\rho}_1)|. \tag{2.13}$$

Second, assume $f_1(\bar{\rho}_1) < \mu_2$, i.e., the queue q_2 is decreasing. Let $\bar{t}$ be such that $q_2(\bar{t}) = 0$. For $t < \min\{\bar{t}, t^*\}$, (2.13) still holds; thus we conclude the case $t^* \le \bar{t}$. If, on the contrary, $\bar{t} < t^*$, we obtain a new traveling discontinuity on the outgoing arc $j = 2$ for times $t > \bar{t}$ when the queue q_2 becomes empty: $f_2(\bar{\rho}_2(a_2+, t)) = f_1(\bar{\rho}_1)$ and $\partial_t q_2(t) = 0$ for $t > \bar{t}$. Then,

$$\begin{aligned} T.V.(\partial_t q_2(t), [t_0, t]) &= |\partial_t q_2(t_0) - \partial_t q_2(t_0+)| + |\partial_t q_2(t_0+) - \partial_t q_2(t)| \\ &\le |f_1(\rho_{1,0}) - f_1(\bar{\rho}_1)| + |f_1(\bar{\rho}_1) - \mu_2| \\ &\le 2\ |f_1(\rho_{1,0}) - f_1(\bar{\rho}_1)| + |\partial_t q_2(t_0)|. \end{aligned}$$

∎

We can now reason as follows. Define

$$\eta = \min_j |b_j - a_j|$$

as the minimum length of a supplier and set

$$TV_j^k = T.V.(f_j(\rho_j^\delta(\cdot, k\eta))), \qquad q_j^k = \partial_t q_j^\delta(k\eta).$$

Then by Lemmas 2.2 and 2.3, we get

$$T.V.(\partial_t q_j^\delta, [k\eta, (k+1)\eta]) \le 2\, TV_{j-1}^k + q_j^k,$$

$$q_j^k + TV_{j-1}^k + TV_j^k \le q_j^{k-1} + TV_{j-1}^{k-1} + TV_j^{k-1}.$$

Moreover, defining by $\widetilde{TV}_N^k$ the variation in the flux produced on the last supplier by the queue q_N^δ on the time interval $[k\eta, (k+1)\eta]$, we get

$$TV_1^k \le TV_1^0, \quad q_N^k + TV_{N-1}^k + \widetilde{TV}_N^k \le q_N^{k-1} + TV_{N-1}^{k-1} + \widetilde{TV}_N^{k-1}.$$

Therefore, summing up on j and k we get the following:

$$\sum_{j=2}^{N} T.V.(\partial_t q_j^\delta, [0, K\eta]) = \sum_{j=2}^{N} \sum_{k=0}^{K-1} T.V.(\partial_t q_j^\delta, [k\eta, (k+1)\eta])$$

$$\le \sum_{j=2}^{N} \sum_{k=0}^{K-1} \left(2\, TV_{j-1}^k + q_j^k\right) \le K \sum_{j=2}^{N} \left(2\, TV_{j-1}^0 + q_j^0\right).$$

Restating, we have

$$\sum_{j=2}^{N} T.V.(\partial_t q_j^\delta, [0, K\eta]) \le K \sum_{j=2}^{N} \left(2\ T.V.(\rho_{j-1,0}^\delta(\cdot)) + |\partial_t q_j^\delta(0)|\right)\ \forall t. \tag{2.14}$$

2.5. Existence of a network solution for BV initial data. For existence of solutions, we consider the space the space of data (ρ, q) on the supply chain with the norm

$$\|(\rho, q)\| = \sum_j \|\rho_j\|_{L^1} + \sum_j |q_j|. \tag{2.15}$$

Then, we want to find a solution in the space $Lip([0,T], L^1((a_j, b_j)))$ for the ρ components and in the space $W^{1,1}([0,T])$ for the q components.

Due to the special flux function, we obtain discontinuities traveling with speed v at most equal to 1. Therefore, we have for $t_1 < t_2$ and every j

$$\int_{a_j}^{b_j} |\rho_j^\delta(x, t_1) - \rho_j^\delta(x, t_2)| dx \le T.V.(\rho_j^\delta(\cdot, t_1))|t_1 - t_2| + \int_{t_1}^{t_2} |f(\rho_j^\delta(a_j, t))|\, dt. \tag{2.16}$$

The estimate (2.16) guarantees Lipschitz dependence w.r.t. time in L^1, while (2.10) ensures uniform BV bounds. Therefore, by using standard techniques [5, 17], one can show that for $\delta \to 0$ a subsequence of ρ^δ converges in L^1 provided that $T.V.(\rho_{j,0}(x))$ is bounded. Furthermore, the limit solution ρ^* is a weak entropic solution for (1.1).

Concerning q_j, we observe that $\partial_t q_j$ are of bounded variation. Again by BV compactness, we have that $\partial_t q_j$ converges by subsequences in BV, in particular, almost everywhere and strongly in L^1. Thus q_j converges uniformly. Finally q_j converges by subsequences in $W^{1,1}$.

Remark 2.4. Notice that we can pass to the limit using the uniform Lipschitz continuities of q_j. In fact, by definition, $Lip_t(q_j) \leq \max\{\mu_{j-1}, \mu_j\}$. Thus we can pass to the limit obtaining Lipschitz continuous functions with the same bound on the Lipschitz constant.

Also, we can pass to the limit using estimate (2.10) and the Ascoli–Arzelá theorem, but in that case we cannot guarantee that $\partial_t q_j$ is in BV and that q_j is in $W^{1,1}$.

Consider now the case in which (K) is violated. For every j, the data entering the supplier from a_j satisfies (K). Consider the generalized characteristic $\pi_j(t)$ starting from a_j at time 0 and let τ_j (possibly $+\infty$) be the time in which it reaches b_j. We can divide the supplier into two regions:

$$A_j = \{(t,x) : x \leq \pi_j(t)\}, \qquad B_j = \{(t,x) : x > \pi_j(t)\};$$

see Figure 2.

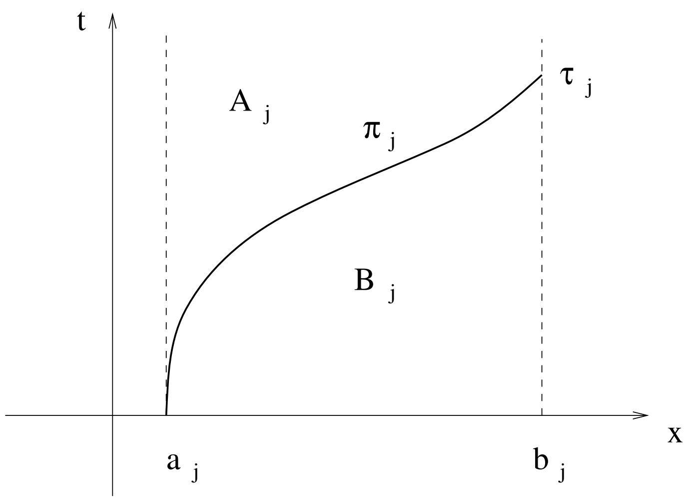


FIG. 2. *Regions A_j and B_j.*

A_j is the region influenced by the incoming flux from a_j, while B_j is the region where ρ depends only on the initial datum $\rho_{j,0}$. Notice that for $t \geq \tau_j$, $B_j \cap \{(t,x) : a_j \leq x \leq b_j\} = \emptyset$. On A_j, (K) holds true; thus the estimate (2.10) also holds. On B_j the solution is the same as the solution to a scalar problem; thus the total variation is decreasing. We thus again reach compactness in BV and the existence of a solution.

Finally, we get the following.

PROPOSITION 2.5. *If $T.V.(\rho_{j,0}(x)) \leq C$ for some $C > 0$, then there exists a solution (ρ, q) on the network such that $(\rho, q) \in Lip([0,T], L^1((a_j,b_j))) \times W^{1,1}([0,T])$, ρ is BV for every time, and $\partial_t q_j$ is in BV.*

2.6. Uniqueness and Lipschitz continuous dependence. We want to prove uniqueness and Lipschitz continuous dependence on the space of data (ρ, q) on the supply chain with the norm (2.15). We use the same approach of [6, 10] and thus consider a Riemannian metric on this space where the tangent vectors are considered only for ρ_j piecewise constant functions.

Let us first focus on the ρ_j's: a "generalized tangent vector" consists of two components (v, ξ), where $v \in L^1$ describes the L^1 infinitesimal displacement, while $\xi \in \mathbb{R}^n$ describes the infinitesimal displacement of discontinuities. A family of piecewise constant functions $\theta \to \rho^\theta$, $\theta \in [0,1]$, with the same number of jumps, say, at the points $x_1^\theta < \cdots < x_M^\theta$, admits a tangent vector in the following functions is well defined (see Figure 3):

$$L^1 \ni v^\theta(x) \doteq \lim_{h \to 0} \frac{\rho^{\theta+h}(x) - \rho^\theta(x)}{h},$$

and also the numbers

$$\xi_\beta^\theta \doteq \lim_{h \to 0} \frac{x_\beta^{\theta+h} - x_\beta^\theta}{h}, \qquad \beta = 1, \ldots, M.$$

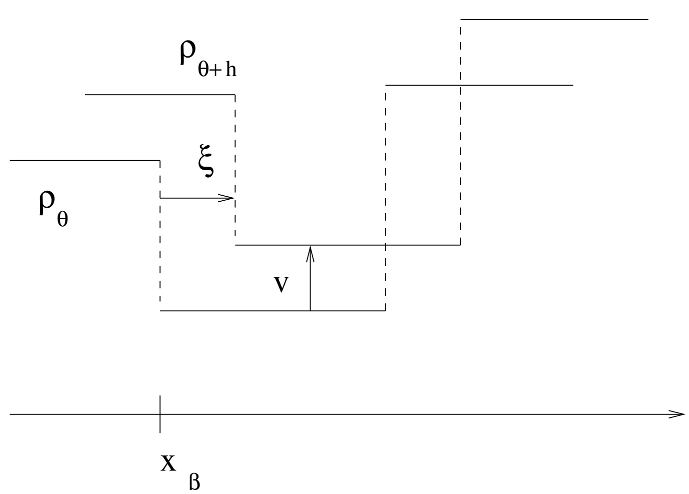


FIG. 3. *Construction of "generalized tangent vectors."*

Notice that the path $\theta \to \rho^\theta$ is not differentiable w.r.t. the usual differential structure of L^1; in fact, if $\xi_\beta^\theta \neq 0$, as $h \to 0$ the ratio $\left[\rho^{\theta+h}(x) - \rho^\theta(x)\right]/h$ does not converge to any limit in L^1.

The L^1 length of the path $\gamma : \theta \to \rho^\theta$ is given by

$$\|\gamma\|_{L^1} = \int_0^1 \left\|v^\theta\right\|_{L^1} d\theta + \sum_{\beta=1}^{M} \int_0^1 \left|\rho^\theta(x_\beta+) - \rho^\theta(x_\beta-)\right| \left|\xi_\beta^\theta\right| d\theta. \tag{2.17}$$

According to (2.17), the L^1 length of a path γ is the integral of the norm of its tangent vector, defined as follows:

$$\|(v, \xi)\| \doteq \|v\|_{L^1} + \sum_{\beta=1}^{M} |\Delta\rho_\beta| \, |\xi_\beta| \,,$$

where $\Delta\rho_\beta = \rho(x_\beta+) - \rho(x_\beta-)$ is the jump across the discontinuity x_β.

Now, given two piecewise constant functions ρ and ρ', call $\Omega(u, u')$ the family of all "differentiable" paths $\gamma : [0, 1] \to \gamma(t)$ with $\gamma(0) = u$, $\gamma(1) = u'$. The Riemannian distance between u and u' is given by

$$d(u, u') \doteq \inf \left\{ \|\gamma\|_{L^1}, \gamma \in \Omega(u, u') \right\}.$$

To define d on all L^1, for given $u, u' \in L^1$ we set

$$\begin{aligned} d(u, u') \doteq \inf \{ &\|\gamma\|_{L^1} + \|u - \tilde{u}\|_{L^1} + \|u' - \tilde{u}'\|_{L^1} : \\ &\tilde{u}, \tilde{u}' \text{ piecewise constant functions, } \gamma \in \Omega(u, u') \}. \end{aligned}$$

It is easy to check that this distance coincides with the distance of L^1.

To estimate the L^1 distance among wave front tracking approximate solutions we proceed as follows. Take ρ, ρ' piecewise constant initial data and let $\gamma_0(\vartheta) = u^\vartheta$ be a regular path joining $\rho = \rho^0$ with $\rho' = \rho^1$. Define $\rho^\vartheta(t, x)$ to be a wave front tracking approximate solution with initial data ρ^ϑ and let $\gamma_t(\vartheta) = \rho^\vartheta(t, \cdot)$. Then for every $t \geq 0$, γ_t is a differentiable path. If we can prove that

$$\|\gamma_t\|_{L^1} \leq \|\gamma_0\|_{L^1} \tag{2.18}$$

for every $t \geq 0$, then

$$\|\rho(t, \cdot) - \rho'(t, \cdot)\|_{L^1} \leq \inf_{\gamma_t} \|\gamma_t\|_{L^1} \leq \inf_{\gamma_0} \|\gamma_0\|_{L^1} = \|\rho(0, \cdot) - \rho'(0, \cdot)\|_{L^1}. \tag{2.19}$$

Now, to obtain (2.18), and hence (2.19), it is enough to prove that, for every tangent vector $(v, \xi)(t)$ to any regular path γ_t, one has

$$\|(v, \xi)(t)\| \leq \|(v, \xi)(0)\|; \tag{2.20}$$

i.e., the norm of a tangent vector does not increase in time. Moreover, if (2.19) is established, then uniqueness and Lipschitz continuous dependence of solutions to Cauchy problems are straightforwardly achieved passing to the limit on the wave front tracking approximate solutions.

Remark 2.6. Since the Riemannian distance d is equivalent to the L^1 metric, the reader could think that the whole framework is not so useful. On the contrary, the different differential structure permits one to rely on tangent vectors, whose norm can be easily controlled. This would not be possible using the tangent vectors of the usual differential structure of L^1, i.e., having only the v component.

Also, while for systems of conservation laws it is possible to find a decreasing functional (see [7]), this is not the case for networks (see [10]), even for a scalar conservation law.

Let us now turn to the supply chains case. It is easy to see that all paths in L^1 connecting piecewise constant functions can be realized using only the ξ component of the tangent vector; see [5, 6]. Therefore, indicating by $x_{\beta_i^j}$ the positions of discontinuities, $j = 1, \ldots, N$, $i = 1, \ldots, M_j$, a tangent vector to a function defined on the network is given by

$$(\xi_{\beta_i^j}, \eta_j),$$

where $\xi_{\beta_i^j}$ is the shift of the discontinuity $x_{\beta_i^j}$, while η_j is the shift of the queue buffer occupancy q_j. The norm of a tangent vector is given by

$$\|(\xi_{\beta_i^j}, \eta_j)\| = \sum_{j,i} |\xi_{\beta_i^j}| |\Delta\rho_{\beta_i^j}| + \sum_j |\eta_j|.$$

Again, to control the distance among solutions it is enough to control the evolution of norms of tangent vectors. Finally, we have the following lemma.

LEMMA 2.7. *The norm of tangent vectors are decreasing along wave front tracking approximations.*

Proof. The norm of tangent vectors changes only at interaction times or if a wave is generated (see [6]); thus we have to consider three cases.

(i) Two waves interact on a supplier.
(ii) A wave interacts with a vertex.
(iii) One queue empties down.

Case (i) is the same as the classical case; see [5, 6].

Consider case (ii) and assume that the interaction happens with vertex j at time t. Let us indicate by $f_j^{\pm}$ the value of the flux at a_j before and after the interaction and, similarly, by $f_{j-1}^{\pm}$ the value of the flux at b_{j-1}. In general we use the symbols $+$ and $-$ to indicate quantities before and after the interaction, respectively.

Assume first that $q_j(t) = 0$; then $f_{j-1}^- = f_j^- < \mu_j$. If $f_{j-1}^+ \leq \mu_j$, then the queue remains empty, a ρ wave is generated on supplier j, and the tangent vector norm remains unchanged. If $f_{j-1}^+ > \mu_j$, then $\xi^+ = \xi^-$, $\Delta\rho^+ = \mu_j - f_j^-$, and $\eta^+ = \eta^- + \xi^-(f_{j-1}^+ - \mu_j)$. Since $\Delta\rho^- = f_{j-1}^+ - f_{j-1}^- = f_{j-1}^+ - f_j^-$, the norm is conserved.

Assume now that $q_j(t) > 0$; then $f_j^- = f_j^+ = \mu_j$. No ρ wave is produced, $\eta^+ = \eta^- + \xi^-\Delta\rho^-$, and again we conclude.

Let us pass to case (iii) and use the same notation of case (ii). Then $f_j^- = \mu_j$ and $f_j^+ = f_{j-1}^- = f_{j-1}^+ < \mu_j$. We get $\Delta\rho^+ = \mu_j - f_{j-1}^-$, $\xi^+ = \eta^-/(\mu_j - f_{j-1}^-)$, and $\eta^+ = 0$; thus we are finished. $\square$

2.7. Existence for L^1 initial data. Since we proved Lipschitz continuous dependence, by an approximation argument, we also get existence for L^1 initial data. More precisely, we get the following theorem.

THEOREM 2.8. *There exists a Lipschitz continuous semigroup S_t defined on the domain $\mathcal{D} = \{(\rho_j, q_j) : \rho_j \in L^\infty, q_j \in \mathbb{R}\}$. Moreover, for every initial datum (ρ_j, q_j) with ρ_j of bounded variation, the semigroup trajectory $t \mapsto S_t(\rho_j, q_j)$ is a network solution.*

We point out that assumption (K) guarantees the existence of a solution on the network, while this is not granted in the general case, as shown by the following example.

Example 2.9. Consider a simple network formed by only one vertex connecting an incoming arc $j = 1$ and an outgoing arc $j = 2$ and initial data

$$\rho_1(0, x) = \mu_1 = \mu_2, \qquad \rho_2(0, x) = \mu_2 + \sin^2\left(\frac{1}{x - a_2}\right), \qquad q_2(0) > 0.$$

Clearly on the outgoing arc $j = 2$ the solution takes values in the flat part of the flux; thus it is constant in time. In particular, $\rho_2(t, x)$ has no trace as $x \to a_2$ for any value of t.

Remark 2.10. Notice that (2.2) still makes sense for Example 2.9 if we interpret the relation to hold for every limit $\lim_n \rho_2(t, x_n)$ with $x_n \to a_2$. On the other side, we can make oscillations in ρ_2 arbitrarily large if we put no constraints on the possible values of ρ_2.

3. General networks. Now, we turn to the case of more general networks as, for example, depicted in Figure 4.

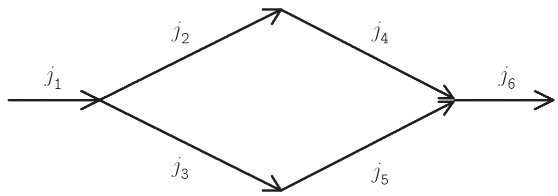


FIG. 4. *Network geometry for a supply chain.*

3.1. Modeling. We consider general vertices $v \in \mathcal{V}$ with m_v incoming and n_v outgoing arcs. The set of arc indices of incoming (outgoing) arcs is denoted by δ_v^- (δ_v^+). If we have more than one outgoing arc, we need to define the distribution of the goods from the incoming arcs. Similar to [8], we model this as follows. We assume that for each single vertex v a matrix $A_v := A(\alpha_{ij})_{i,j} \in \mathbb{R}^{m_v \times n_v}$ is given and that the total flux willing to go to arc $j \in \delta_v^+$ is given by

$$\sum_{i \in \delta_v^-} \alpha_{ij} f_i(\rho_i(b_i-, t)).$$

Therefore, we assume that the matrix A satisfies for all $i \in \delta_v^-, j \in \delta_v^+ : 0 \leq \alpha_{ij} \leq 1$ and $\sum_{j \in \delta_v^+} \alpha_{ij} = 1$. Then, the supply chain network model is given by (1.1) and for each junction v by the equations for the queues (see also [11])

$$\forall j \in \delta_v^+ : \partial_t q_j(t) = \sum_{i \in \delta_v^-} \alpha_{ij} f_i(\rho_i(b_i-, t)) - f_j(\rho_j(a_j+, t)), \tag{3.1}$$

and the boundary values $\forall j \in \delta_v^+$,

$$f_j(\rho_j(a_j+, t)) = \begin{cases} \min\{\sum_{i \in \delta_v^-} \alpha_{ij} f_i(\rho_i(b_i-, t)); \mu_j\}, & q_j(t) = 0, \\ \mu_j, & q_j(t) > 0. \end{cases} \tag{3.2}$$

Note that due to the positive velocity of the occurring waves the boundary conditions are well defined. In particular, and in contrast to [8, 14], no additional maximization problem near the vertex has to be solved. Moreover, due to (3.1) and the assumption on A, we conserve the total flux at each vertex v for all times $t > 0$:

$$\sum_{j \in \delta_v^+} (\partial_t q_j(t) + f_j(\rho_j(a_j+, t))) = \sum_{i \in \delta_v^-} f_i(\rho_i(b_i-, t)).$$

Now, the construction of a solution to the network problem (1.1), (3.1), (3.2) is as before. In particular, the results of [12] extend to problem (1.1), (3.1), (3.2) on the network $(\mathcal{J}, \mathcal{V})$. It is enough to control the number of waves and interactions: Let $\eta = \min_j(b_j - a_j)$ be the minimum length of a supplier. Since all waves move at positive velocity at most equal to 1, two interactions with vertices of the same wave can happen at most every η units of time. If N is the number of suppliers, then there is at most a multiplication by N every η unit of time; thus we control the number of waves and interactions.

Therefore, for given piecewise constant initial data $\rho_{j,0}^\delta$ on a network, a solution (ρ^δ, q^δ) can be defined by the wave tracking method up to any time T. Next, we extend Lemma 2.2 to the more general situation of a vertex v above.

3.2. Existence, uniqueness, and Lipschitz continuous dependence of a weak solution. We can get again BV estimates.

LEMMA 3.1. *Assume we have a single vertex with incoming arcs* $\delta^- = \{1, \dots, m\}$ *and outgoing arcs* $\delta^+ = \{m+1, \dots, m+n\}$*. Furthermore, assume we have constant states* $\rho_{j,0}, j \in \delta^- \cup \delta^+$*, at the vertex and consider a discontinuity collision at time* t_0*. Denote the new solution at the vertex after the collision by* $\bar{\rho}_j$*. Assume that there are no more collision of discontinuities until* t^**. Then, for all* $t_0 < t < t^*$*,*

$$\begin{aligned}\sum_{j\in\delta^-\cup\delta^+} T.V.(f_j(\rho_j(\cdot,t))) + \sum_{j\in\delta^+} |\partial_t q_j(t)| \\ \le \sum_{j\in\delta^-\cup\delta^+} T.V.(f_j(\rho_j(\cdot,t_0))) + \sum_{j\in\delta^+} |\partial_t q_j(t_0)|.\end{aligned} \tag{3.3}$$

Proof. The proof is very similar to the proof of Lemma 2.2. The colliding discontinuity has to arrive on an arc $i \in \delta^-$ and we assume $i = 1$. The total variation on the incoming arc $i = 1$ therefore decreases by

$$|f_1(\bar{\rho}_1) - f_1(\rho_{1,0})| = \sum_{j\in\delta^+} |\alpha_{1j} f_1(\bar{\rho}_1) - \alpha_{1j} f_1(\rho_{1,0})|.$$

Hence, it suffices to prove that for any fixed outgoing arc $j \in \delta^+$ and for all $t > t_0$ the following inequality holds:

$$|\alpha_{1j} f_1(\bar{\rho}_1) - \alpha_{1j} f_1(\rho_{1,0})| + |\partial_t q_j(t_0)| \ge T.V.(f_j(\rho_j(\cdot,t))) + |\partial_t q_j(t)|. \tag{3.4}$$

Fix $j \in \delta^+$. With the other cases being similar we discuss only the (most interesting) case: Assume

$$\sum_{i\in\delta^-} \alpha_{ij} f_i(\rho_{i,0}) > f_j(\rho_{j,0})$$

and

$$\alpha_{1j} f_1(\bar{\rho}_1) + \sum_{i,i\ne1} f_i(\rho_{i,0}) < \mu_j.$$

Then, the queue q_j is decreasing after the collision at time t_0, and we denote again by $\bar{t}$ the time when $q_j(\bar{t})0$. Then for $t < \min\{\bar{t}, t^*\}$ we obtain $T.V.(f_j(\bar{\rho}_j(\cdot,t))) = 0$ and $|\alpha_{1j} f_1(\bar{\rho}_1) - \alpha_{1j} f_1(\rho_{1,0})| + |\partial_t q_j(t_0)| \ge \mu_j - \sum_{i,i\ne1} \alpha_{ij} f_i(\rho_{i,0}) - \alpha_{1j} f_1(\bar{\rho}_1) = |\partial_t q_j(t)|$. If $\bar{t} < t^*$, then a new discontinuity is generated since the queue q_j empties. By (3.2) we have

$$f_j(\bar{\rho}_j(a_j+,t)) = \sum_{i,i\ne1} \alpha_{ij} f_j(\rho_{j,0}) + \alpha_{1j} f_1(\bar{\rho}_1),$$

and therefore

$$|\partial_t q(\bar{t})| T.V.(f_j(\rho_j(\cdot,t)))$$

for $t > \bar{t}$. Hence, (3.4) holds for all $t > t_0$. This finishes the proof. ∎

Therefore, we again obtain the estimate (2.10), where the sum now should run over all arcs and nodes of the network. Moreover, the estimates on $\partial_t q_j$ work in the same way.

The same arguments as above give existence and uniqueness of a weak solution as well as the Lipschitz continuous dependence on the data in the general case for BV initial data. Finally, Theorem 2.8 holds for a general network.

4. Summary. We have proven existence, uniqueness, and Lipschitz continuous dependence of a weak solution to a network model for supply chains. The model consists of a scalar hyperbolic equation governing the dynamics of a supplier and an ordinary differential equation for describing the behavior of the queues. The proof of existence relies on the front tracking approximations and estimates on the total variation.

REFERENCES

[1] D. Armbruster, P. Degond, and C. Ringhofer, *Kinetic and fluid models for supply chains supporting policy attributes*, Transport Theory Statist. Phys., to appear.

[2] D. Armbruster, P. Degond, and C. Ringhofer, *A model for the dynamics of large queuing networks and supply chains*, SIAM J. Appl. Math., 66 (2006), pp. 896–920.

[3] D. Armbruster, D. Marthaler, and C. Ringhofer, *Kinetic and fluid model hierarchies for supply chains*, Multiscale Model. Simul., 2 (2003), pp. 43–61.

[4] M. K. Banda, M. Herty, and A. Klar, *Coupling conditions for gas networks governed by the isothermal Euler equations*, Netw. Heterog. Media, 1 (2006), pp. 295–314.

[5] A. Bressan, *Hyperbolic Systems of Conservation Laws*, Oxford University Press, Oxford, UK, 2000.

[6] A. Bressan, G. Crasta, and B. Piccoli, *Well-posedness of the Cauchy problem for $n \times n$ systems of conservation laws*, Mem. Amer. Math. Soc., 146 (694) (2000).

[7] A. Bressan, T. P. Liu, and T. Yang, *l^1 stability estimates for $n \times n$ conservation laws*, Arch. Ration. Mech. Anal., 149 (1999), pp. 1–22.

[8] G. M. Coclite, M. Garavello, and B. Piccoli, *Traffic flow on a road network*, SIAM J. Math. Anal., 36 (2005), pp. 1862–1886.

[9] C. M. Dafermos, *Polygonal approximations of solutions of the initial value problem for a conservation law*, J. Math. Anal. Appl., 38 (1972), pp. 33–41.

[10] M. Garavello and B. Piccoli, *Traffic Flow on Network*, Appl. Math. Ser. 1, Institute of Mathematical Sciences, Springfield, MO, 2006.

[11] S. Göttlich, M. Herty, and A. Klar, *Modelling and optimization of supply chains on complex networks*, Commun. Math. Sci., 4 (2006), pp. 315–330.

[12] S. Göttlich, M. Herty, and A. Klar, *Network models for supply chains*, Commun. Math. Sci., 3 (2005), pp. 545–559.

[13] M. Herty and M. Rascle, *Coupling conditions for a class of second-order models for traffic flow*, SIAM J. Math. Anal., 38 (2006), pp. 595–616.

[14] H. Holden and N. H. Risebro, *A mathematical model of traffic flow on a network of unidirectional roads*, SIAM J. Math. Anal., 26 (1995), pp. 999–1017.

[15] H. Holden and N. H. Risebro, *Front Tracking for Hyperbolic Conservation Laws*, Springer-Verlag, New York, Berlin, Heidelberg, 2002.

[16] S. N. Kruzkov, *First order quasi linear equations in several independent variables*, Math. USSR Sbornik, 10 (1970), p. 217.

[17] J. Smoller, *Shock Waves and Reaction–Diffusion Equations*, Springer-Verlag, Heidelberg, 1983.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 174–194

BIFURCATION FOR A FREE BOUNDARY PROBLEM MODELING TUMOR GROWTH BY STOKES EQUATION*

AVNER FRIEDMAN† AND BEI HU‡

Abstract. We consider a free boundary problem modeling tumor growth in fluid-like tissue. The model equations include a diffusion equation for the nutrient concentration, and the Stokes equation with a source which represents the proliferation density of the tumor cells. The proliferation rate μ and the cell-to-cell adhesiveness γ which keeps the tumor intact are two parameters which characterize the "aggressiveness" of the tumor. For any positive radius R there exists a unique radially symmetric stationary solution with radius $r = R$. We prove that for a sequence $\mu/\gamma = M_n(R)$ there exist symmetry-breaking bifurcation branches of solutions with free boundary $r = R + \varepsilon Y_{n,0}(\theta) + O(\varepsilon^2)$ (n even ≥ 2) for small $|\varepsilon|$, where $Y_{n,0}$ is the spherical harmonic of mode $(n, 0)$. Furthermore, the smallest $M_n(R)$, say, $M_{n_*}(R)$, is such that $n_* = n_*(R) \to \infty$ as $R \to \infty$. The biological implications of this result are discussed at the end of the paper.

Key words. free boundary problems, stationary solution, stability, instability, bifurcation, symmetry-breaking, tumor growth

AMS subject classifications. Primary, 35R35, 35K55; Secondary, 35Q80, 35C20, 92C37

DOI. 10.1137/060656292

1. Introduction.

Mathematical models of solid tumor growth, which consider the tumor tissue as a density of proliferating cells, have been developed and studied in many papers; see [1, 2, 3, 4, 7, 9, 10, 18, 20, 21, 23, 27, 28, 30, 31] and the references given there. Most of the models discuss the case of radially symmetric tumors.

Tumors grown in vitro have a nearly spherical shape, but tumors in vivo may develop all kinds of protrusions. It is therefore interesting to explore the existence of nonspherical solutions of tumor models.

Let $\Omega(t)$ denote the tumor domain at time t, and p the pressure within the tumor resulting from proliferation of the tumor cells. The density of the cells, c, depends on the concentration of nutrients, σ, and, assuming that this dependence is linear, we simply identify c with σ. We also assume a linear dependence of the proliferation rate S on σ, that is,

$$S = \mu(\sigma - \widetilde{\sigma}),$$

where $\widetilde{\sigma} > 0$ is a threshold concentration and μ is a positive parameter. If the consumption rate of nutrients by tumor cells is proportional to the concentration of the nutrients, then σ satisfies

$$\beta\sigma_t - \Delta\sigma + \sigma = 0 \quad \text{in } \Omega(t), \qquad \sigma = \overline{\sigma} \quad \text{on } \partial\Omega(t). \tag{1.1}$$

Most tumor models assume that the tissue has the structure of a porous medium for which Darcy's law holds. If we denote by p the pressure within the tumor resulting

*Received by the editors April 4, 2006; accepted for publication (in revised form) November 13, 2006; published electronically April 27, 2007. This work was partially supported by National Science Foundation upon agreement 0112050.
http://www.siam.org/journals/sima/39-1/65629.html
†Department of Mathematics, The Ohio State University, 231 West 18th Avenue, Columbus, OH 43210 (afriedman@mbi.osu.edu).
‡Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556 (b1hu@nd.edu).

from the proliferation of tumor cells, then p is related to the velocity $\vec{v}$ by which the cells move, by Darcy's law $\vec{v} = -\nabla p$. Since by conservation of mass

$$\operatorname{div} \vec{v} = S = \mu(\sigma - \widetilde{\sigma}), \tag{1.2}$$

we obtain for the pressure p the equation

$$\Delta p = -\mu(\sigma - \widetilde{\sigma}) \quad \text{in } \Omega(t). \tag{1.3}$$

There are, however, tumors for which the tissue is more naturally modeled as fluid. For example, in early stages of breast cancer the tumor is confined to the duct of a mammary gland, which consists of epithelial cells, a meshwork of proteins, and extracellular fluid. Several recent papers on ductal carcinoma in the breast use the Stokes equation in their mathematical models [15, 16, 17]. If we denote the fluid velocity by $\vec{v} = (v_1, v_2, v_3)$ and the fluid pressure by p, then the constitutive law is $\sigma_{ij} = -p\delta_{ij} + 2\nu(e_{ij} - \frac{1}{3}\overline{\Delta}\,\delta_{ij})$, where σ_{ij} is the stress tensor, $p = -\frac{1}{3}\sigma_{kk}$, $e_{ij} = \frac{1}{2}(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i})$ is the strain tensor, $\overline{\Delta} = e_{kk} = \operatorname{div}\vec{v}$ is the dilation, and ν is the viscosity coefficient. If there are no body forces, then $\frac{\partial \sigma_{ij}}{\partial x_j} = 0$.

We can rewrite this equation as the Stokes equation

$$-\nu\Delta\vec{v} + \nabla p - (\nu/3)\nabla \operatorname{div}\vec{v} = 0 \quad \text{in } \Omega(t), \quad t > 0. \tag{1.4}$$

We now turn to the boundary conditions at the boundary $\Gamma(t)$ of $\Omega(t)$. We assume that the tumor is held together by the forces of cell-to-cell adhesion with constant γ; the role of γ is discussed in [5, 6, 8]. Introducing the stress tensor $Q = \nu(\nabla\vec{v} + (\nabla\vec{v})^T) - (p + \frac{2}{3}\nu \operatorname{div}\vec{v})I$ with components $Q_{ij} = \nu(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}) - \delta_{ij}(p + \frac{2\nu}{3}\operatorname{div}\vec{v})$, we then have $Q\vec{n} = -\gamma\kappa\vec{n}$ on $\Gamma(t)$, where $\vec{n}$ is the outward normal and κ is the mean curvature ($\kappa > 0$ if $\Gamma(t)$ is the surface of a convex body). Noting that $\frac{2\nu}{3}\operatorname{div}\vec{v} = \frac{2\nu}{3}\mu(\overline{\sigma} - \widetilde{\sigma})$ on $\Gamma(t)$, this boundary condition can be written in the form

$$T\vec{n} = \big\{-\gamma\kappa + (2\nu/3)\mu(\overline{\sigma} - \widetilde{\sigma})\big\}\vec{n} \quad \text{on } \Gamma(t), \quad t > 0, \tag{1.5}$$

where $T_{ij} = \nu(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}) - \delta_{ij}p$.

We also assume the kinematic condition $\vec{v}\cdot\vec{n} = V_n$ on $\Gamma(t)$, $t > 0$.

The system (1.2), (1.4), and (1.5) has six-dimensional kernel V_0 consisting of rigid motions $\vec{v}_0 = \vec{a} + \vec{b} \times \vec{x}$. We must therefore add six scalar constraints. These constraints can be written in the form $\int_{\Omega(t)} \vec{v}\, dx = \vec{A}(t)$, $\int_{\Omega(t)} \vec{v} \times \vec{x}\, dx = \vec{B}(t)$, where $\vec{A}(t)$, $\vec{B}(t)$ are prescribed functions. Finally, we prescribe the initial condition

$$\Omega(t)\Big|_{t=0} = \Omega_0, \qquad \sigma\Big|_{t=0} = \sigma_0(x) \quad \text{in } \Omega_0. \tag{1.6}$$

Local existence and uniqueness of solutions for the system (1.4)–(1.6) for a general domain were recently proved by Friedman [19].

In this paper we are interested in discovering a nonspherical stationary solution to the tumor model based on the Stokes equation. For the case of Darcy's law it was proved in Friedman and Reitich [23] that if $\widetilde{\sigma} < \overline{\sigma}$ then there exists a unique stationary spherical solution with radius $R = R_S$ which is determined by $\widetilde{\sigma}/\overline{\sigma}$. It was further proved in Fontelos and Friedman [13] that there exists a family of symmetry-breaking bifurcation branches of stationary solutions with free boundary

$$r = R_S + \varepsilon Y_{n,0} + O(\varepsilon^2) \qquad (n \geq 2) \tag{1.7}$$

with μ, γ such that

$$\frac{\mu}{\gamma} = M_n(R_S) + O(\varepsilon), \quad M_n(R) = \frac{(n-1)n(n+2)}{2R^5 P_0(R)[P_1(R) - P_n(R)]}, \tag{1.8}$$

where $Y_{n,0}$ is the spherical harmonic of order $(n, 0)$, $P_n(R) = I_{n+3/2}(R)/[RI_{n+1/2}(R)]$, and $I_m(r)$ is the modified Bessel function of order m; furthermore, $M_n(R) < M_{n+1}(R)$. For dimension 2, such a result was proved earlier in Friedman and Reitich [24].

The purpose of the present paper is to establish the existence of symmetry-breaking bifurcation branches of stationary solutions in the case where we replace Darcy's law by the Stokes equation and maintain the conservation law (1.2).

By scaling we may assume that $\nu = 1$ and $\overline{\sigma} = 1$. We also take $\vec{A} = 0$, $\vec{B} = 0$ in the constraints. Then the stationary problem becomes

$$-\Delta\sigma + \sigma = 0 \quad \text{in } \Omega, \qquad \sigma = 1 \quad \text{on } \partial\Omega, \tag{1.9}$$

$$-\Delta\vec{v} + \nabla p = (\mu/3)\nabla(\sigma - \widetilde{\sigma}) \quad \text{in } \Omega, \tag{1.10}$$

$$\operatorname{div}\vec{v} = \mu(\sigma - \widetilde{\sigma}) \quad \text{in } \Omega \quad (\widetilde{\sigma} < 1), \tag{1.11}$$

$$T(\vec{v}, p)\vec{n} = \Big(-\gamma\kappa + \frac{2\nu}{3}\mu(1 - \widetilde{\sigma})\Big)\vec{n} \quad \text{on } \partial\Omega, \tag{1.12}$$

$$\vec{v} \cdot \vec{n} = 0 \quad \text{on } \partial\Omega, \tag{1.13}$$

$$\int_\Omega \vec{v}\, dx = 0, \qquad \int_\Omega \vec{v} \times \vec{x}\, dx = 0, \tag{1.14}$$

where $T(\vec{v}, p) = (\nabla\vec{v})^T + \nabla\vec{v} - p\, I$, $I = (\delta_{ij})_{i,j=1}^3$.

In this paper, we shall prove that there exists a family of symmetry-breaking bifurcation branches of stationary solutions of (1.9)–(1.14) with free boundary as (1.7) for n even ≥ 2, where

$$\frac{\mu}{\gamma} = M_n(R_S) + O(\varepsilon),\ M_n(R) = \frac{n(n+2)(2n+1)}{4(n+1)(2n+3)} \frac{1}{R^3 P_0(R)[P_1(R) - P_n(R)]} \tag{1.15}$$

instead of (1.8). In an interesting contrast to the situation for (1.8), the present sequence $M_n(R_S)$ is generally not monotone increasing for all $n \geq 2$. Instead, it is monotone increasing beginning only from some $n = \overline{n}(R_S)$; furthermore, if $M_{n_*(R_S)}(R_S) = \min\{M_n(R_S);\ n = 2, 3, 4, \dots\}$, then $n_*(R_S) \to \infty$ if $R_S \to \infty$. The biological implications of this result will be discussed in the concluding section of this paper. For the reader's convenience, some important formulas on vector spherical harmonics and on Bessel functions are collected in Appendices A and B.

2. Radially symmetric stationary spherical solution. The only radially symmetric solution of (1.9) is

$$\sigma_S(r) = \frac{R_S}{\sinh R_S}\frac{\sinh r}{r} = \frac{R_S^{1/2}}{I_{1/2}(R_S)}\frac{I_{1/2}(r)}{r^{1/2}} \quad \text{in } \{r < R_S\}, \tag{2.1}$$

where R_S is uniquely determined by integrating (1.11) over $\{r < R_S\}$ and using the relation (1.13), $\int_0^{R_S} \mu(\sigma_S - \widetilde{\sigma})r^2 dr = 0$, or (cf. [23])

$$P_0(R_S) \equiv (R_S \coth R_S - 1)/R_S^2 = \widetilde{\sigma}/3. \tag{2.2}$$

We shall show that the pressure and velocity corresponding to σ_S are

$$p_S(r) = \overline{p} + \frac{4\mu}{3}\sigma_S(r), \quad \vec{v}_S = \mu G(r) r \vec{e}_r, \quad G(r) = \frac{R^{1/2}}{I_{1/2}(R)} \frac{I_{3/2}(r)}{r^{3/2}} - P_0(R), \tag{2.3}$$

where $R = R_S$, and the constant $\overline{p}$ in (2.3) is determined by

$$p_S(R) \;=\; \frac{\gamma}{R} + \frac{4\mu}{3}(1 - \widetilde{\sigma}) \;\equiv\; \frac{\gamma}{R} + \frac{4\mu}{3}(1 - 3P_0(R)). \tag{2.4}$$

In fact, from (B.14) with $n = 0$ we get $G(R) = 0$, so that (1.13) is satisfied. Using various properties of the Bessel functions we verify

$$RG'(R) = 1 - 3P_0(R) = \sigma_S(R) - \widetilde{\sigma}, \tag{2.5}$$

$$4G'(r) + rG''(r) = (\sigma_S)_r(r) = \frac{R^{1/2}}{I_{1/2}(R)} \frac{I_{3/2}(r)}{r^{1/2}}, \tag{2.6}$$

$$\Big(\frac{\partial}{\partial x_i}(G(r)x_j) + \frac{\partial}{\partial x_j}(G(r)x_i)\Big)\Big|_{r=R} = 2G'(R)\frac{x_i x_j}{R}\Big|_{r=R}, \tag{2.7}$$

from which we deduce that (1.10)–(1.12) are satisfied. Finally, the constraints in (1.14) are obviously satisfied. From now on we set $R = R_S$.

3. The bifurcation problem. Consider a family of domains with boundaries $\partial\Omega_\varepsilon : r = R_S + \widetilde{R}(\theta, \phi)$, where $\widetilde{R}(\theta, \phi) = \varepsilon S(\theta, \phi)$. Let $(\sigma, \vec{v}, p)$ be the solution of (1.9)–(1.12) with the constraints (1.14). We define a function F by

$$F(\widetilde{R}, \mu) = \vec{v} \cdot \vec{n}. \tag{3.1}$$

Then $(\sigma, \vec{v}, p, R_S + \widetilde{R})$ is a stationary solution if and only if $F(\widetilde{R}, \mu) = 0$.

The function $S(\theta, \phi)$ may be viewed as a function defined on the unit sphere $\Sigma = \{x;\ |x| = 1\}$. We shall later assume that $S(\theta, \phi)$ is in $C^{m+\alpha}(\Sigma)$, that is, S is $C^{m+\alpha}$ as a function defined on Σ; note that this does not mean that $S(\theta, \phi)$ is in $C^{m+\alpha}$ in the variable (θ, ϕ). We shall see in sections 4–6 that F maps the space $C^{m+\alpha}(\Sigma) \times \mathbb{R}$ into the space $C^{m+\alpha-1}(\Sigma)$. In section 4 we shall formally compute the Fréchet derivatives of F. Since F is smooth (as will be shown in section 6), the formal derivation will be rigorously justified.

In order to compute the Fréchet derivatives of F, we need the expansion of $(\sigma, \vec{v}, p)$ of order ε. Note that the normal vector $\vec{n}$ is given by

$$\vec{n} = \vec{e}_r - \varepsilon \frac{S_\theta}{R} \vec{e}_\theta - \varepsilon \frac{S_\phi}{R \sin\theta} \vec{e}_\phi + O(\varepsilon^2) = \vec{e}_r - \frac{\varepsilon}{R} \nabla_\omega S + O(\varepsilon^2). \tag{3.2}$$

We can formally write, for any fixed value of the positive parameter μ,

$$\sigma = \sigma_S + \varepsilon\sigma_1 + O(\varepsilon^2), \quad p = p_S + \varepsilon p_1 + O(\varepsilon^2), \quad \vec{v} = \vec{v}_S + \varepsilon \vec{v}_1 + O(\varepsilon^2), \tag{3.3}$$

$$\partial\Omega_\varepsilon : r = R + \varepsilon S(\theta, \phi). \tag{3.4}$$

Since $\vec{v}_S(R) = 0$, $F(\varepsilon S, \mu) = \varepsilon(\vec{v}_1 + S \frac{\partial}{\partial r} \vec{v}_S)|_{r=R} \cdot \vec{e}_r + O(\varepsilon^2)$. Using the relation $S(\theta, \phi) \frac{\partial}{\partial r} \vec{v}_S|_{r=R} \cdot \vec{e}_r = \mu G'(R) R\, S(\theta, \phi)$ and (2.5), the Fréchet derivative is given by

$$[F_{\widetilde{R}}(0, \mu)]S = \Big(\vec{v}_1 + S \frac{\partial}{\partial r} \vec{v}_S\Big)\Big|_{r=R} \cdot \vec{e}_r = \vec{v}_1\Big|_{r=R} \cdot \vec{e}_r + \mu(1 - 3P_0(R))\, S(\theta, \phi). \tag{3.5}$$

Without the constraint (1.14), the velocity $\vec{v}_1$ is determined only up to an additive vector of the form $\vec{a}+\vec{b}\times\vec{x}$. In section 4 we shall compute this Fréchet derivative, and in section 5 we shall prove that this derivative satisfies the conditions of the bifurcation theorem of Crandall and Rabinowitz for a sequence of parameters M_n, thus allowing us to conclude the existence of symmetry-breaking bifurcations initiating from M_n.

4. Computation of Fréchet derivative. In this section we shall compute the Fréchet derivative $[F_{\widetilde{R}}(0,\mu)]S$. For clarity we shall first perform the computation in a formal way; namely, we assume that the asymptotic formulas (3.3) are valid and that these relations may be differentiated several times; we also drop all the $O(\varepsilon^2)$ terms. All these formal computations will be made rigorous in section 6.

LEMMA 4.1.

$$[F_{\widetilde{R}}(0,\mu)]Y_{l,0} = \Big\{\mu R^2 P_0(R)[P_1(R)-P_l(R)]\Big(1+\frac{l}{(2l^2+4l+3)}\Big) - \frac{\gamma(l+2)(2l+1)l}{4R(2l^2+4l+3)}\Big\}Y_{l,0} \quad \textit{for } l\neq 1, \tag{4.1}$$

$$[F_{\widetilde{R}}(0,\mu)]Y_{1,0}=0 \quad \textit{for } l=1. \tag{4.2}$$

The rigorous proof of this lemma will be given in section 5.

Since we will not impose in this section the constraints of (1.14), $\vec{v}$ and $\vec{v}_1$ are determined only up to additive vectors $\vec{a}+\vec{b}\times\vec{x}$. In the next section we shall impose the constraints (1.14) for $\vec{v}$ and then determine $\vec{v}_1$ in a unique way in the form $\vec{v}_1 = \vec{v}_{1*}+\vec{a}+\vec{b}\times\vec{x}$, where $\vec{v}_{1*}$ is a special solution satisfying (1.9)–(1.13) and $\vec{a},\vec{b}$ are such that (1.14) is satisfied.

4.1. Computation of $\boldsymbol{\sigma_1}$. Formally, $s_1|_{\partial B_R} = \sigma_1|_{\partial\Omega_\varepsilon}+O(\varepsilon)$, so that

$$\varepsilon\sigma_1\Big|_{\partial B_R} = \varepsilon\sigma_1\Big|_{\partial\Omega_\varepsilon}+O(\varepsilon^2) = -\varepsilon(\sigma_S)_r(R)S+O(\varepsilon^2). \tag{4.3}$$

Substituting (3.3) and (4.3) into (1.9), we find that σ_1 satisfies

$$-\Delta\sigma_1+\sigma_1=0 \quad \text{in } B_R, \qquad \sigma_1=-(\sigma_S)_r(R)S(\theta,\phi) \quad \text{on } \partial B_R. \tag{4.4}$$

Hence, we have the following lemma.

LEMMA 4.2. *If $S=Y_{l,m}$, then*

$$\sigma_1 = -(\sigma_S)_r(R)\frac{I_{l+1/2}(r)}{r^{1/2}}\,\frac{R^{1/2}}{I_{l+1/2}(R)}Y_{l,m}(\theta,\phi). \tag{4.5}$$

Remark 4.1. Lemma 4.2 is understood in the sense that it holds for the real and imaginary parts of $Y_{l,m}$ separately. The same interpretation will be used in subsequent computations.

4.2. Computation of $\boldsymbol{\vec{v}_1}$ and $\boldsymbol{p_1}$. Substituting (3.3) into (1.10), (1.11), we get

$$-\Delta\vec{v}_1+\nabla p_1=(\mu/3)\nabla\sigma_1 \quad \text{in } B_R, \tag{4.6}$$

$$\operatorname{div}\vec{v}_1=\mu\sigma_1 \quad \text{in } B_R. \tag{4.7}$$

By taking the divergence in (4.6) and using (4.7), we also have

$$\Delta\big(p_1-(4\mu/3)\sigma_1\big)=0 \quad \text{in } B_R. \tag{4.8}$$

Using (4.4), we can rewrite (4.6) in the form

$$\Delta\big(\vec{v}_1 - \mu\nabla\sigma_1\big) = \nabla\big(p_1 - (4\mu/3)\sigma_1\big) \quad \text{in } B_R. \tag{4.9}$$

In the special case $S = Y_{l,m}$, we look for a solution $(\vec{v}_1, p_1)$ of the form

$$\vec{v}_1 = \vec{a} + \vec{b} \times \vec{x} + \mu\nabla\sigma_1 + v_{l,m}(r)\vec{V}_{l,m} + x_{l,m}(r)\vec{X}_{l,m} + w_{l,m}(r)\vec{W}_{l,m}, \tag{4.10}$$

$$p_1 = \frac{4\mu}{3}\sigma_1 + p_{l,m}(r)Y_{l,m}, \tag{4.11}$$

where $\vec{V}_{l,m}$, $\vec{X}_{l,m}$, and $\vec{W}_{l,m}$ are vector spherical harmonics; see (A.10)–(A.12) for their expressions.

Then by (A.15)

$$\begin{aligned} \nabla\Big(p_1 - \frac{4\mu}{3}\sigma_1\Big) \;=\; & \Big(\frac{l+1}{2l+1}\Big)^{1/2}\Big[-\frac{\partial}{\partial r}p_{l,m} + \frac{l}{r}p_{l,m}\Big]\vec{V}_{l,m} \\ & + \Big(\frac{l}{2l+1}\Big)^{1/2}\Big[\frac{\partial}{\partial r}p_{l,m} + \frac{l+1}{r}p_{l,m}\Big]\vec{W}_{l,m}. \end{aligned} \tag{4.12}$$

Applying (A.15) to $I_{l+1/2}(r)/r^{1/2}$ and using (B.9) and (B.10), we obtain

$$\begin{aligned} \nabla\sigma_1 \;=\; & -(\sigma_S)_r(R)\frac{R^{1/2}}{I_{l+1/2}(R)}\Big\{-\Big(\frac{l+1}{2l+1}\Big)^{1/2}\frac{I_{l+3/2}(r)}{r^{1/2}}\vec{V}_{l,m} \\ & + \Big(\frac{l}{2l+1}\Big)^{1/2}\frac{I_{l-1/2}(r)}{r^{1/2}}\vec{W}_{l,m}\Big\}. \end{aligned} \tag{4.13}$$

Taking the divergence in (4.10) and using (4.4) and (A.16)–(A.18), we find that

$$\begin{aligned} \operatorname{div}\vec{v}_1 \;=\; & \mu\sigma_1 - \Big(\frac{l+1}{2l+1}\Big)^{1/2}\Big[\frac{dv_{l,m}}{dr} + \frac{l+2}{r}v_{l,m}\Big]Y_{l,m} \\ & + \Big(\frac{l}{2l+1}\Big)^{1/2}\Big[\frac{dw_{l,m}}{dr} - \frac{l-1}{r}w_{l,m}\Big]Y_{l,m}. \end{aligned} \tag{4.14}$$

Similarly, taking the Laplacian in (4.10) and using (A.20)–(A.22), we obtain

$$\Delta\big(\vec{v}_1 - \mu\nabla\sigma_1\big) = L_{l+1}(v_{l,m})\vec{V}_{l,m} + L_l(x_{l,m})\vec{X}_{l,m} + L_{l-1}(w_{l,m})\vec{W}_{l,m}. \tag{4.15}$$

Then the system (4.7), (4.6) reduces to four ODEs:

$$-\Big(\frac{l+1}{2l+1}\Big)^{1/2}\Big[\frac{dv_{l,m}}{dr} + \frac{l+2}{r}v_{l,m}\Big] + \Big(\frac{l}{2l+1}\Big)^{1/2}\Big[\frac{dw_{l,m}}{dr} - \frac{l-1}{r}w_{l,m}\Big] = 0, \tag{4.16}$$

$$L_{l+1}(v_{l,m}) = \Big(\frac{l+1}{2l+1}\Big)^{1/2}\Big[-\frac{\partial}{\partial r}p_{l,m} + \frac{l}{r}p_{l,m}\Big], \tag{4.17}$$

$$L_l(x_{l,m}) = 0, \tag{4.18}$$

$$L_{l-1}(w_{l,m}) = \Big(\frac{l}{2l+1}\Big)^{1/2}\Big[\frac{\partial}{\partial r}p_{l,m} + \frac{l+1}{r}p_{l,m}\Big]. \tag{4.19}$$

LEMMA 4.3. *The solutions of the system* (4.16)–(4.19) *have the form*

$$p_{l,m}(r) = 2A_1(2l+3)r^l, \tag{4.20}$$

$$v_{l,m}(r) = \frac{2lA_1}{2l+1}\Big(\frac{2l+1}{l+1}\Big)^{1/2}r^{l+1}, \tag{4.21}$$

$$x_{l,m}(r) = B_1 r^l, \tag{4.22}$$

$$w_{l,m}(r) = C_1 l\Big(\frac{2l+1}{l}\Big)^{1/2}r^{l-1} + \frac{A_1(2l+3)l}{(2l+1)}\Big(\frac{2l+1}{l}\Big)^{1/2}r^{l+1}, \tag{4.23}$$

where A_1, B_1, C_1 are arbitrary constants.

Proof. From (4.8), (4.11), and (A.19), we have $L_l(p_{l,m}) = 0$. This equation has two linearly independent solutions, r^l and r^{-l-1}. Since $p_{l,m}$ is not singular at $r = 0$, we have $p_{l,m}(r) = D_1 r^l$ for some constant D_1. Hence (4.17) is reduced to $L_{l+1}(v_{l,m}) = 0$. This equation has linearly independent solutions r^{l+1} and r^{-l-2}. Since $v_{l,m}(r)$ is not singular at $r = 0$, it must be equal to const $\cdot\, r^{l+1}$, which, for convenience in future computations, we write in the form (4.21).

Similarly, $x_{l,m}(r)$ has the form (4.22). Finally, (4.19) is reduced to

$$L_{l-1}(w_{l,m}) = D_1\Big(\frac{l}{2l+1}\Big)^{1/2}(2l+1)r^{l-1}; \tag{4.24}$$

the general solution of this equation which is not singular at $r = 0$ is of the form

$$w_{l,m}(r) = C_1 l\Big(\frac{2l+1}{l}\Big)^{1/2} r^{l-1} + \frac{D_1}{2}\Big(\frac{l}{2l+1}\Big)^{1/2} r^{l+1}, \tag{4.25}$$

where C_1 is an arbitrary constant. A direct computation shows that (4.16) holds if and only if

$$-\frac{2lA_1}{2l+1}(2l+3) + \Big(\frac{l}{2l+1}\Big)^{1/2} D_1 \Big(\frac{l}{2l+1}\Big)^{1/2} = 0. \tag{4.26}$$

Therefore, $w_{l,m}(r)$ is given by (4.23). ∎

4.3. The boundary condition for $\vec{v}_1$ and p_1. We need three equations to determine the constants A_1, B_1, C_1. They should come from the boundary conditions for $\vec{v}_1$ and p_1. By [23], the curvature κ for the surface $r = R_S + \varepsilon S$ is given by

$$\kappa = \frac{1}{R} - \frac{\varepsilon}{R^2}\Big(S + \frac{1}{2}\Delta_\omega S\Big) + O(\varepsilon^2). \tag{4.27}$$

Thus, by (3.2),

$$\kappa\vec{n} = \frac{1}{R}\vec{e}_r - \frac{\varepsilon}{R^2}\Big\{\nabla_\omega S + \Big(S + \frac{1}{2}\Delta_\omega S\Big)\vec{e}_r\Big\} + O(\varepsilon^2). \tag{4.28}$$

A direct computation shows that

$$\begin{aligned} T(\vec{v},p)\vec{n}\Big|_{\partial\Omega_\varepsilon} &= \Big(T(\vec{v}_S,p_S)\Big|_{\partial\Omega_\varepsilon} + \varepsilon T(\vec{v}_1,p_1)\Big|_{\partial\Omega_\varepsilon}\Big)\Big(\vec{e}_r - \frac{\varepsilon}{R}\nabla_\omega S\Big) + O(\varepsilon^2) \\ &= T(\vec{v}_S,p_S)\Big|_{\partial B_R}\vec{e}_r + \varepsilon\Big\{\Big(T(\vec{v}_1,p_1)\Big|_{\partial B_R} + S\frac{\partial}{\partial r}T(\vec{v}_S,p_S)\Big|_{\partial B_R}\Big)\vec{e}_r \\ &\quad - \frac{1}{R}T(\vec{v}_S,p_S)\Big|_{\partial B_R}\nabla_\omega S\Big\} + O(\varepsilon^2), \end{aligned} \tag{4.29}$$

where $T(\vec{v}_1,p_1) = \nu(\nabla\vec{v}_1 + (\nabla\vec{v}_1)^T) - p_1 I$. Thus the boundary condition for $\vec{v}$ becomes

$$T(\vec{v}_1,p_1)\Big|_{\partial B_R}\vec{e}_r = \vec{\Phi}_1, \tag{4.30}$$

where

$$\begin{aligned} \vec{\Phi}_1 \;&=\; \frac{\gamma}{R^2}\Big\{\nabla_\omega S + \Big(S + \frac{1}{2}\Delta_\omega S\Big)\vec{e}_r\Big\} - S\frac{\partial}{\partial r}T(\vec{v}_S,p_S)\Big|_{\partial B_R}\vec{e}_r \\ &\quad + \Big(\frac{1}{R}T(\vec{v}_S,p_S)\Big|_{\partial B_R} - \frac{2\mu}{3R}(1-\widetilde{\sigma})\Big)\nabla_\omega S. \end{aligned} \tag{4.31}$$

For a (column) vector $\vec{Z}$, ∇Z denotes the matrix with components $(\nabla\vec{Z})_{ij} = \partial_{x_i} Z_j$, so that for any vector $\vec{U}$, we have $\nabla\vec{Z}\cdot\vec{U} = (\partial_{x_1}\vec{Z}^T\cdot\vec{U},\, \partial_{x_2}\vec{Z}^T\cdot\vec{U},\, \partial_{x_3}\vec{Z}^T\cdot\vec{U})^T$. Since $\vec{v}_S = \mu G(r) r\vec{e}_r$ and $r\nabla\vec{e}_r = I - \vec{e}_r\vec{e}_r^{\,T}$, for any vector $\vec{U}$

$$\begin{aligned}\nabla\vec{v}_S\cdot\vec{U} &= \Big[\frac{\partial}{\partial r}\Big(\mu G(r) r\Big)\vec{e}_r\Big]\vec{e}_r^{\,T}\cdot\vec{U} + \mu G(r) r\nabla\vec{e}_r\cdot\vec{U}\\ &= \frac{\partial}{\partial r}\Big(\mu G(r) r\Big)(\vec{U}\cdot\vec{e}_r)\vec{e}_r + \mu G(r)\Big(\vec{U} - (\vec{U}\cdot\vec{e}_r)\vec{e}_r\Big)\\ &= \Big\{-\mu G(r) + \frac{\partial}{\partial r}\Big(\mu G(r) r\Big)\Big\}(\vec{U}\cdot\vec{e}_r)\vec{e}_r + \mu G(r)\vec{U}.\end{aligned} \tag{4.32}$$

Since $\nabla_\omega S\cdot\vec{e}_r = 0$, if we apply (4.32) to $\vec{U} = \nabla_\omega S$, the right-hand side of (4.32) will vanish at $r = R$. The same is true if we replace ∇v_S by $(\nabla v_S)^T$ in (4.32), since ∇v_S is symmetric (cf. (2.7)). Hence

$$\begin{aligned}\Big(T(\vec{v}_S, p_S)\Big|_{\partial B_R} - \frac{2\mu}{3}(1-\widetilde{\sigma})\Big)\nabla_\omega S &= \Big(-p_S(R) - \frac{2\mu}{3}(1-\widetilde{\sigma})\Big)\nabla_\omega S\\ &= -\Big(\frac{\gamma}{R} + 2\mu(1 - 3P_0(R))\Big)\nabla_\omega S,\end{aligned} \tag{4.33}$$

where in the last equality we used (2.2), (2.4). Similarly, taking $\vec{U} = \vec{e}_r$ and differentiating in r, we get

$$\begin{aligned}\frac{\partial}{\partial r}T(\vec{v}_S, p_S)\Big|_{\partial B_R}\vec{e}_r &= \Big[2\frac{\partial^2}{\partial r^2}\Big(\mu G(r) r\Big) - \frac{\partial p_S(r)}{\partial r}\Big]_{r=R}\vec{e}_r\\ &= -2\mu\Big[\frac{2}{R}(1 - 3P_0(R)) - \frac{1}{3}RP_0(R)\Big]\vec{e}_r,\end{aligned}$$

where in the last equality we used (2.5) and (2.6). Substituting this and (4.33) into (4.31), we find that, in the special case $S = Y_{l,m}$,

$$\begin{aligned}\vec{\Phi}_1 \;=\; & -\frac{2\mu}{R}(1 - 3P_0(R))\nabla_\omega Y_{l,m} + \Big\{\frac{\gamma}{R^2}\Big(1 - \frac{l(l+1)}{2}\Big)\\ & + \frac{4\mu}{R}(1 - 3P_0(R)) - \frac{2\mu}{3}RP_0(R)\Big\}Y_{l,m}\vec{e}_r.\end{aligned} \tag{4.34}$$

Using (A.13), (A.14), we then obtain the orthogonality relation

$$\langle\vec{\Phi}_1, \vec{X}_{l,m}\rangle = 0. \tag{4.35}$$

According to Lemma 4.1 of [26], if $T(\vec{v}_1, p_1)\vec{e}_r = \vec{\Phi}_1 = M_1\vec{V}_{l,m} + M_2\vec{X}_{l,m} + M_3\vec{W}_{l,m}$, where $\vec{v}_1$, p_1 are as in (4.10), (4.11) with $\vec{a} = 0$, $\vec{b} = 0$, $\nabla\sigma_1 = 0$, then $x_{l,m}$ satisfies

$$\frac{\partial x_{l,m}}{\partial r} - \frac{1}{r}x_{l,m} = M_2 \quad \text{at } r = R. \tag{4.36}$$

The same is true if σ_1 is as in (4.5) since, by (4.13), the introduction of σ_1 does not change M_2 (i.e., $\langle\nabla\sigma_1, \vec{X}_{l,m}\rangle = 0$); the same is true if the constant vectors are not equal to zero since $T(\vec{v}_1, p_1) = T(\vec{v}_1 - \vec{a} - \vec{b}\times\vec{x}, p_1)$. In the present case $M_2 = 0$ (by (4.34) and (A.13), (A.14)), so that, by (4.22), equation (4.36) reduces to

$$B_1\Big(lR^{l-1} - \frac{R^l}{R}\Big) = 0, \tag{4.37}$$

and, since $l \geq 2$, $B_1 = 0$.

Using Lemma 4.3 with $B_1 = 0$, and (4.5), one can verify that

$$\vec{v}_1 = \vec{a} + \vec{b} \times \vec{x} + H_1(r) Y_{l,m} \vec{e}_r + H_2(r) \nabla_\omega Y_{l,m}, \tag{4.38}$$

where

$$\begin{aligned} H_1(r) \;=\; & \Big\{ -\mu R P_0(R) \frac{R^{1/2}}{I_{l+1/2}(R)} \Big(\frac{I_{l+3/2}(r)}{r^{1/2}} + \frac{l}{r^{3/2}} I_{l+1/2}(r) \Big) \\ & + A_1 l\, r^{l+1} + C_1 l\, r^{l-1} \Big\}, \end{aligned} \tag{4.39}$$

$$H_2(r) \;=\; -\mu R P_0(R) \frac{R^{1/2}}{I_{l+1/2}(R)} \frac{I_{l+1/2}(r)}{r^{3/2}} + A_1 \frac{(l+3)}{(l+1)} r^{l+1} + C_1 r^{l-1}. \tag{4.40}$$

Also, by (4.11), (4.5), and the equality $(\sigma_S)_r(R) = R P_0(R)$,

$$p_1\Big|_{r=R} = \Big(-\frac{4\mu}{3} R P_0(R) + 2A_1(2l+3) R^l \Big) Y_{l,m}. \tag{4.41}$$

Lemma 4.4. *For any function $H(r)$, there hold*

$$\nabla\Big(H(r) Y_{l,m} \vec{e}_r\Big) \cdot \vec{e}_r = \nabla\Big(H(r) Y_{l,m}\Big) = H'(r) Y_{l,m} \vec{e}_r + r^{-1} H(r) \nabla_\omega Y_{l,m}, \tag{4.42}$$

$$\Big(\nabla\Big(H(r) Y_{l,m} \vec{e}_r\Big)\Big)^T \cdot \vec{e}_r = H'(r) Y_{l,m} \vec{e}_r, \tag{4.43}$$

$$\nabla\Big(H(r) \nabla_\omega Y_{l,m}\Big) \cdot \vec{e}_r = -r^{-1} H(r)\, \nabla_\omega Y_{l,m}, \tag{4.44}$$

$$\Big(\nabla\Big(H(r) \nabla_\omega Y_{l,m}\Big)\Big)^T \cdot \vec{e}_r = H'(r) \nabla_\omega Y_{l,m}. \tag{4.45}$$

Proof. As in the proof of Lemma 4.1 of [26], the following relations hold for any vectors $\vec{U}, \vec{Z}$ and scalar Ψ:

$$\nabla(\Psi \vec{Z}) \cdot \vec{e}_r = \Psi \nabla \vec{Z} \cdot \vec{e}_r + (\vec{Z} \cdot \vec{e}_r) \nabla \Psi, \tag{4.46}$$

$$(\nabla \vec{Z})^T \vec{U} = (\nabla \vec{Z}) \vec{U} - \vec{U} \times (\nabla \times \vec{Z}), \tag{4.47}$$

$$(\nabla \vec{Z}) \cdot \vec{U} = \nabla(\vec{Z} \cdot \vec{U}) - \nabla \vec{U} \cdot \vec{Z}, \tag{4.48}$$

$$\nabla \vec{e}_r \cdot \vec{Z} = r^{-1} \Big(\vec{Z} - (\vec{Z} \cdot \vec{e}_r) \vec{e}_r \Big). \tag{4.49}$$

Using (4.46) with $\Psi = H(r) Y_{l,m}$ and $\vec{Z} = \vec{e}_r$, and then also recalling (4.49), we obtain (4.42).

Using (A.4)–(A.5), we can directly compute $\nabla \times (H(r) Y_{l,m} \vec{e}_r)$. If we now apply (4.47) with $\vec{U} = \vec{e}_r$, $Z = H(r) Y_{l,m} \vec{e}_r$ and use (4.42), we obtain the assertion (4.43).

Similarly, $\nabla(H(r) \nabla_\omega Y_{l,m}) \cdot \vec{e}_r = H(r) \nabla(\nabla_\omega Y_{l,m}) \cdot \vec{e}_r$, since $\nabla_\omega Y_{l,m} \cdot \vec{e}_r = 0$. Using (4.48) with $\vec{Z} = \nabla_\omega Y_{l,m}$, $\vec{U} = \vec{e}_r$, we derive $\nabla(\nabla_\omega Y_{l,m}) \cdot \vec{e}_r = -\nabla \vec{e}_r \cdot \nabla_\omega Y_{l,m}$, and using (4.49), the inequality (4.44) follows.

Finally, using (A.4)–(A.7) and (A.10)–(A.14), one can verify directly that, for some function $\widetilde{Q}$,

$$\begin{aligned} \nabla \times (H(r) \nabla_\omega Y_{l,m}) \;=\; & H'(r) \Big(\vec{e}_\phi \partial_\theta Y_{l,m} - \frac{\vec{e}_\theta}{\sin\theta} \partial_\phi Y_{l,m} \Big) + \widetilde{Q} \vec{e}_r \\ & + r^{-1} H(r) \Big(\vec{e}_\phi (\partial_\theta Y_{l,m}) - \frac{\vec{e}_\theta}{\sin\theta} (\partial_\phi Y_{l,m}) \Big). \end{aligned} \tag{4.50}$$

Taking the vector product with $\vec{e}_r$ and using (4.44), (4.47), we obtain (4.45). ∎

By Lemma 4.4,

$$\begin{aligned}&\Big((\nabla\vec{v}_1)+(\nabla\vec{v}_1)^T\Big)\Big|_{r=R}\cdot\vec{e}_r\\&=2H_1'(r)\Big|_{r=R}Y_{l,m}\vec{e}_r+\Big(r^{-1}H_1(r)-r^{-1}H_2(r)+H_2'(r)\Big)\Big|_{r=R}\nabla_\omega Y_{l,m}.\end{aligned}$$

A direct computation shows that

$$2H_1'(r)\Big|_{r=R}=-2\mu RP_0(R)\Big(1-2P_l(R)+\frac{l(l-1)}{R^2}\Big)+2A_1l(l+1)R^l+2C_1l(l-1)R^{l-2},$$

and

$$\begin{aligned}&\Big(r^{-1}H_1(r)-r^{-1}H_2(r)+H_2'(r)\Big)\Big|_{r=R}\\&=-2\mu RP_0(R)\Big(P_l(R)+\frac{l-1}{R^2}\Big)+2A_1\frac{l(l+2)}{(l+1)}R^l+2C_1(l-1)R^{l-2}.\end{aligned}$$

Recalling (4.34), one can verify that the boundary condition (4.30) reduces to the two equations

$$\begin{aligned}&2A_1\frac{l(l+2)}{(l+1)}R^l+2C_1(l-1)R^{l-2}\\&=-\frac{2\mu}{R}(1-3P_0(R))+2\mu RP_0(R)\Big(P_l(R)+\frac{l-1}{R^2}\Big)\end{aligned}\tag{4.51}$$

and

$$\begin{aligned}&2A_1R^l(l^2-l-3)+2C_1R^{l-2}l(l-1)\\&=\frac{\gamma}{R^2}\Big(1-\frac{l(l+1)}{2}\Big)+\frac{4\mu}{R}(1-3P_0(R))\\&\quad-4\mu RP_0(R)P_l(R)+2\mu RP_0(R)\frac{l(l-1)}{R^2}.\end{aligned}\tag{4.52}$$

In order to compute the Fréchet derivative in (3.5), we need only to compute the linear combination $A_1R^{l+1}+C_1R^{l-1}$, since they always appear in this form. To do that, note that the combination $3(l+1)R\times(4.51)+(2l+1)R\times(4.52)$ gives

$$\begin{aligned}&2(l-1)(2l^2+4l+3)\Big(A_1R^{l+1}+C_1R^{l-1}\Big)y\\&=-\gamma\frac{(l-1)(l+2)}{2R}(2l+1)+2\mu(1-3P_0(R))(l-1)\\&\quad-2\mu R^2P_0(R)P_l(R)(l-1)+2\mu P_0(R)(l-1)(2l^2+4l+3).\end{aligned}\tag{4.53}$$

If $l\neq 1$, then

$$\begin{aligned}A_1R^{l+1}+C_1R^{l-1}=\mu P_0(R)+\frac{1}{2(2l^2+4l+3)}\Big\{&-\gamma\frac{(l+2)}{2R}(2l+1)\\&+2\mu(1-3P_0(R))-2\mu R^2P_0(R)P_l(R)\Big\},\end{aligned}$$

so that by (4.39), (4.40),

$$\Big(\vec{v}_1 - \vec{a} - \vec{b}\times\vec{x}\Big)\cdot\vec{e}_r\Big|_{r=R} = \Big\{-\mu R P_0(R)\Big(RP_l(R) + \frac{l}{R}\Big) + A_1 l R^{l+1} + C_1 l R^{l-1}\Big\} Y_{l,m}$$
$$= \Big\{-\mu R^2 P_0(R) P_l(R) + \frac{l}{2(2l^2+4l+3)}\Big(-\gamma\frac{(l+2)}{2R}(2l+1) + 2\mu(1-3P_0(R)) - 2\mu R^2 P_0(R)P_l(R)\Big)\Big\} Y_{l,m}.$$

Recalling that $1 - 3P_0(R) = R^2 P_0(R) P_1(R)$ (cf. (B.16) with $n = 0$), we obtain

$$\begin{aligned}&\Big(\vec{v}_1 - \vec{a} - \vec{b}\times\vec{x}\Big)\cdot\vec{e}_r\Big|_{r=R} + \mu(1-3P_0(R))\, Y_{l,m}\\ &= \Big\{\mu R^2 P_0(R)[P_1(R) - P_l(R)]\Big(1 + \frac{l}{(2l^2+4l+3)}\Big) - \frac{\gamma(l+2)(2l+1)l}{4R(2l^2+4l+3)}\Big\} Y_{l,m}.\end{aligned} \tag{4.54}$$

From (3.5) and the fact that $(\vec{b}\times\vec{x})\cdot\vec{e}_r = 0$ we then get

$$\begin{aligned}[F_{\widetilde{R}}(0,\mu)]Y_{l,m} \;=\; &\Big\{\mu R^2 P_0(R)[P_1(R) - P_l(R)]\Big(1 + \frac{l}{(2l^2+4l+3)}\Big)\\ &- \frac{\gamma(l+2)(2l+1)l}{4R(2l^2+4l+3)}\Big\} Y_{l,m} + \vec{a}\cdot\vec{e}_r.\end{aligned} \tag{4.55}$$

Note that we have not yet implemented the constraints (1.14). This will be done in the next section, where we shall prove that (1.14) for $l \geq 2$ is satisfied with $\vec{a} = 0$ and some vector $\vec{b}$.

Hence, for $l \geq 2$, $\vec{a} = 0$ in (4.55). Therefore, the expression in (4.55) vanishes if and only if $\mu/\gamma = M_l(R)$, where

$$M_l(R) = \frac{k(l)}{R^3 P_0(R)[P_1(R) - P_l(R)]}, \quad k(l) = \frac{l(l+2)(2l+1)}{4(l+1)(2l+3)}, \quad l \geq 2. \tag{4.56}$$

So far we have assumed that the Fréchet derivative exists and we computed it at $S = Y_{l,m}$. Since the Fréchet derivative is a linear operator, the formula (4.55) formally extends to any $S = \sum_l a_l Y_{l,0}(\theta)$.

5. Constraints in (1.14).

Proof of Lemma 4.1. When we computed the Fréchet derivative in (4.54) we assumed that $(\sigma, \vec{v}, p)$ is a solution of (1.9)–(1.13), but we have not yet imposed the constraints in (1.14). If we impose these constraints on $\vec{v}$, then we obtain similar constraints on $\vec{v}_1$:

$$\int_{B_R} \vec{v}_1\, dx = 0, \qquad \int_{B_R} \vec{v}_1 \times \vec{x}\, dx = 0. \tag{5.1}$$

Writing v_1 in the form (4.38), namely, $\vec{v}_1 = \vec{a} + \vec{b}\times\vec{x} + \vec{v}_{1*}$, where

$$\vec{v}_{1*} = H_1(r) Y_{l,m}\vec{e}_r + H_2(r)\nabla_\omega Y_{l,m} \tag{5.2}$$

and H_1, H_2 are given by (4.39)–(4.40), we then need to determine $\vec{a}, \vec{b}$ such that

$$\int_{B_R} (\vec{a} + \vec{b}\times\vec{x}) dx = -\int_{B_R} \vec{v}_{1*}\, dx, \quad \int_{B_R} (\vec{a} + \vec{b}\times\vec{x})\times\vec{x}\, dx = -\int_{B_R} \vec{v}_{1*}\times\vec{x}\, dx, \tag{5.3}$$

by (5.1). By a direct computation using (A.1), (5.3) is reduced to

$$\int_{B_R} \vec{v}_{1*}\, dx = -\frac{4\pi}{3} R^3 \vec{a}, \qquad \int_{B_R} \vec{v}_{1*} \times \vec{x}\, dx = \frac{8\pi}{15} R^5 \vec{b}. \tag{5.4}$$

The case $l \geq 2$. Since $Y_{l,0}$ is orthogonal to $Y_{1,0}$ for $l \neq 1$,

$$\int_0^\pi Y_{l,0}(\theta)\cos\theta\, \vec{e}_3(\sin\theta d\theta) = \int_0^\pi Y_{l,0}(\theta) 2\sqrt{\frac{\pi}{3}} Y_{1,0}(\theta)\, \vec{e}_3(\sin\theta d\theta) = 0, \tag{5.5}$$

and using (A.1), (5.5), we obtain

$$\int_0^{2\pi} \Big(\int_0^\pi Y_{l,0}(\theta)\vec{e}_r \sin\theta d\theta\Big) d\phi = 2\pi \int_0^\pi Y_{l,0}(\theta)\cos\theta\, \vec{e}_3 \sin\theta d\theta = 0; \tag{5.6}$$

similarly, by integration by parts,

$$\int_0^{2\pi} \Big(\int_0^\pi \nabla_\omega Y_{l,0}(\theta)\sin\theta d\theta\Big) d\phi = 4\pi \int_0^\pi Y_{l,0}(\theta)\cos\theta\, \vec{e}_3 \sin\theta d\theta \;=\; 0. \tag{5.7}$$

If we substitute (5.2) into the left-hand side of (5.4) and then use (5.6), (5.7), we get $\int_{B_R} \vec{v}_{1*}\, dx = 0$ if $l \neq 1$; hence

$$\vec{a} = 0 \quad \text{ if } l \neq 1. \tag{5.8}$$

We can solve $\vec{b}$ by (5.4), but since $(\vec{b}\times\vec{x})\cdot\vec{e}_r = r(\vec{b}\times\vec{e}_r)\cdot\vec{e}_r = 0$, the extra term from $\vec{b}\times\vec{x}$ does not contribute anything to the Fréchet derivative (cf. (3.5), (4.54)), and hence Lemma 4.1 holds in this case.

In the derivation of some of the formulas above we have assumed $l \geq 2$. So the cases $l = 0$ and $l = 1$ have to be treated separately.

The case $l = 0$. Since $Y_{0,0}$ is a constant, if we perturb the system by $R + \varepsilon Y_{0,0}$, the system remains radially symmetric, and in fact, as easily verified,

$$\sigma_1 = -Y_{0,0}(\sigma_S)_r(R)\frac{I_{1/2}(r)}{r^{1/2}}\frac{R^{1/2}}{I_{1/2}(R)} = -Y_{0,0}(\sigma_S)_r(R)\sigma_S(r), \tag{5.9}$$

$$\vec{v}_1 = \mu\nabla\sigma_1 = \mu(\sigma_1)_r(r)\vec{e}_r, \qquad p_1 = \overline{p}_1 + \frac{4\mu}{3}\sigma_1, \tag{5.10}$$

where $\overline{p}_1$ is a uniquely determined constant. Thus

$$\vec{v}_1\Big|_{r=R} \cdot \vec{e}_r = \mu(\sigma_1)_r(R) = -\mu Y_{0,0}[(\sigma_S)_r(R)]^2 = -\mu Y_{0,0} R^2 P_0^2(R).$$

By (3.5) and the relation $1 - 3P_0(R) = R^2 P_0(R) P_1(R)$ (by (B.16)), we then obtain

$$[F_{\tilde{R}}(0,\mu)]Y_{0,0} = -\mu Y_{0,0} R^2 P_0(R)[P_0(R) - P_1(R)]. \tag{5.11}$$

Thus (4.1) is valid also for $l = 0$. □

The case $l = 1$. The case $l = 1$ has to be treated in a special way. In this case, if we follow the proof for $l \geq 2$ we find that the constants C_1 and B_1 in (4.22) and (4.23) are not determined by the boundary conditions, so some changes need to be made. However, it is easier to compute the Fréchet derivative in a different and more illuminating way. We shall write the spherically symmetric solution $(\sigma_S, \vec{v}_S, p_S)$

on $B_{R_S}(0) = \{|x| < R_S\}$ in the form $(\sigma_S(x), \vec{v}_S(x), p_S(x))$. Let $\vec{x}_0 = \frac{3}{\sqrt{4\pi}}\vec{e}_3$. Then the translated functions $(\sigma_S(x - \varepsilon x_0),\ \vec{v}_S(x - \varepsilon x_0),\ p_S(x - \varepsilon x_0))$ form a solution on $B_{R_S}(\varepsilon x_0) = \{|x - \varepsilon x_0| < R_S\}$. We can rewrite the boundary $|x - \varepsilon x_0| = R_S$ as

$$r = |x| = R_S + \varepsilon r \vec{e}_r \frac{3}{\sqrt{4\pi}}\vec{e}_3 + O(\varepsilon^2) \ = \ R_S + \varepsilon Y_{1,0}(\theta) + O(\varepsilon^2).$$

This means that when the perturbation is given by $\varepsilon Y_{1,0}$, we have the explicit translated solution up to an error of order $O(\varepsilon^2)$. Thus $v_S(x - \varepsilon x_0) = O(\varepsilon^2)$ on $r = R_S + \varepsilon Y_{1,0}(\theta)$ and, by (3.1), $F(\varepsilon Y_{10}, \mu) = O(\varepsilon^2)$. It follows that

$$[F_{\widetilde{R}}(0, \mu)]Y_{1,0} = 0 \quad \text{ for any } \mu > 0. \qquad \square \tag{5.12}$$

6. Rigorous justification. For $S \in C^{k+\alpha}(\Sigma)$, $k \geq 3$, we set $\Omega_\varepsilon = \{r < R + \varepsilon S\}$, and define σ_1 by (4.4) and $(p_1, \vec{v}_1)$ by (4.6)–(4.7), (4.30)–(4.31), and (5.1).

Note that $(\sigma, p, \vec{v})$ is defined only on Ω_ε, while $(\sigma_S, p_S, \vec{v}_S)$ is defined on whole $\mathbb{R}^3$ and $(\sigma_1, p_1, \vec{v}_1)$ is defined on B_R. We need to transform all these functions to the same domain Ω_ε and shall do it by the Hanzawa transformation, which is a diffeomorphism defined by $(r, \theta, \phi) = H_\varepsilon(r', \theta', \phi') \equiv (r' + \chi(R - r')\varepsilon S(\theta', \phi'), \theta', \phi')$, where $\chi \in C^\infty$, $\chi(z) = 0$ if $|z| \geq 3\delta_0/4$, and $\chi(z) = 1$ if $|z| < \delta_0/4$. Observe that H_ε maps B_R onto Ω_ε while keeping the ball $\{r < R - (3\delta_0/4)\}$ fixed. The inverse Hanzawa transformation H_ε^{-1} maps Ω_ε onto B_R. Note that $H_\varepsilon = I + \varepsilon N$, where N is a C^∞ operator.

It was rigorously proved in [22, Lemma 3.2] that

$$\|\sigma - (\sigma_S + \varepsilon\sigma_1(H_\varepsilon^{-1}\cdot))\|_{C^{k+\alpha}(\Omega_\varepsilon)} \leq C|\varepsilon|^2 \|S\|_{C^{k+\alpha}(\Omega_\varepsilon)} \tag{6.1}$$

for $k = 3$; the proof extends to all $k \geq 2$.

From (6.1), it follows that the formal expansion for σ is rigorous. We now proceed to rigorously justify the expansion used in section 5 for p and $\vec{v}$ in the same manner as was done in [22], using estimates derived for the (inhomogeneous) Stokes equation instead of the parabolic equation of [22].

By (4.6)–(4.7), (4.30)–(4.31), and (5.1), we can apply the estimates for the Stokes equation [32, Proposition 2] to obtain

$$\|\vec{v}_1\|_{C^{k-1+\alpha}(B_R)} + \|p_1\|_{C^{k-2+\alpha}(B_R)} \leq C\|\vec{\Phi}_1\|_{C^{k-2+\alpha}(\Sigma)} \leq C\|S\|_{C^{k+\alpha}(\Sigma)}. \tag{6.2}$$

Using (4.6), (4.7) and (6.1), (6.2), we can rewrite (1.10)–(1.11) in the form

$$-\Delta\Big(\vec{v} - \vec{v}_S - \varepsilon\vec{v}_1(H_\varepsilon^{-1}x)\Big) + \nabla\Big(p - p_S - \varepsilon p_1(H_\varepsilon^{-1}x)\Big) = \varepsilon^2 \widetilde{f} \quad \text{ in } \Omega_\varepsilon, \tag{6.3}$$

$$\operatorname{div}\Big(\vec{v} - \vec{v}_S - \varepsilon\vec{v}_1(H_\varepsilon^{-1}x)\Big) = \varepsilon^2 \widetilde{g} \quad \text{ in } \Omega_\varepsilon, \tag{6.4}$$

where

$$\|\widetilde{f}\|_{C^{k-3+\alpha}(\overline{\Omega}_\varepsilon)} + \|\widetilde{g}\|_{C^{k-2+\alpha}(\overline{\Omega}_\varepsilon)} \leq C\|S\|_{C^{k+\alpha}(\Sigma)}. \tag{6.5}$$

In order to derive estimates for the expansion of $(\vec{v}, p)$, we also need to estimate the boundary conditions. By the proof of [25, Theorem 8.1],

$$\big\|\kappa - R^{-1} + \varepsilon R^{-2}\big(S + \Delta_\omega S/2\big)\big\|_{C^{k-2+\alpha}(\Sigma)} \leq C|\varepsilon|^2 \|S\|_{C^{k+\alpha}(\Sigma)}. \tag{6.6}$$

We also have, from (3.2),

$$\big\|\vec{n} - \big(\vec{e}_r - \varepsilon R^{-1}\nabla_\omega S\big)\big\|_{C^{k+\alpha}(\Sigma)} \leq C|\varepsilon|^2 \|S\|_{C^{k+1+\alpha}(\Sigma)}. \tag{6.7}$$

By essentially repeating all the formal computations of section 4 while keeping track of the error terms (using (6.6) and (6.7)), we find that

$$\|T[\vec{v} - \vec{v}_S - \varepsilon \vec{v}_1(H_\varepsilon^{-1}x), p - p_S - \varepsilon p_1(H_\varepsilon^{-1}x)] \cdot \vec{n}\|_{C^{k-2+\alpha}(\Sigma)} \le C|\varepsilon|^2 \|S\|_{C^{k+\alpha}(\Sigma)}.$$

The constraints in (1.14) imposed on $\vec{v}_1$ ensure that

$$\left|\int_{\Omega_\varepsilon} \Big(\vec{v} - \vec{v}_S - \varepsilon \vec{v}_1(H_\varepsilon^{-1}x)\Big) dx\right| + \left|\int_{\Omega_\varepsilon} \Big(\vec{v} - \vec{v}_S - \varepsilon \vec{v}_1(H_\varepsilon^{-1}x)\Big) \times \vec{x}\, dx\right| \le C|\varepsilon|^2. \tag{6.8}$$

We can then apply again the estimates from [32] for the Stokes equation to obtain

$$\|\vec{v} - \vec{v}_S - \varepsilon \vec{v}_1(H_\varepsilon^{-1}\cdot)\|_{C^{k-1+\alpha}(\overline{\Omega}_\varepsilon)} \le C|\varepsilon|^2 \|S\|_{C^{k+\alpha}(\Sigma)}, \tag{6.9}$$

$$\|p - p_S - \varepsilon p_1(H_\varepsilon^{-1}\cdot)\|_{C^{k-2+\alpha}(\overline{\Omega}_\varepsilon)} \le C|\varepsilon|^2 \|S\|_{C^{k+\alpha}(\Sigma)}. \tag{6.10}$$

These estimates combined with (6.1) ensure that the expansion in the computations of the Fréchet derivative are rigorous, that is,

$$\|F(\widetilde{R}, \mu) - F(0, \mu) - \varepsilon F_{\widetilde{R}}(0, \mu) S\|_{C^{k-1+\alpha}} \le \text{const} \cdot |\varepsilon|\, \|\widetilde{R}\|_{C^{k+\alpha}(\Sigma)}, \quad \widetilde{R} = \varepsilon S,$$

where $S \to F_{\widetilde{R}}(0, \mu) S$ is the linear operator defined by (4.1), (4.2).

7. Monotonicity of $M_n(R)$. In the model with Darcy's law (1.2), it was proved in [13] that $M_n(R) < M_{n+1}(R)$ for all $n \ge 2$ and all $R > 0$.

For the present model (1.9)–(1.14) we have a different result:

$$M_n(R) \text{ is monotone increasing beginning only from some } n = \overline{n}(R); \tag{7.1}$$

furthermore, if $M_{n_*(R)}(R) = \min\{M_n(R);\, n = 2, 3, 4, \dots\}$, then

$$n_*(R) \to \infty \quad \text{if } R \to \infty, \tag{7.2}$$

whereas

$$\text{for } R \text{ small, } n_*(R) = 3, \text{ and } M_3(R) < M_2(R) < M_4(R). \tag{7.3}$$

It is clear from (4.56) that $k(n) = n/4 + O(n^{-1})$. Using this in (4.56), we obtain the assertion (7.1). We next establish the assertion (7.2).

THEOREM 7.1. (i) *For every $n > m \ge 2$, there exists an $R_*(n, m)$ such that*

$$M_n(R) < M_m(R) \quad \textit{for } R > R_*(n, m); \tag{7.4}$$

(ii) (7.2) *holds.*

Proof. By (B.21), $P_n(R) = \frac{1}{R} - \frac{n+1}{R^2} + O(\frac{n^2}{R^3})$, so that

$$M_n(R) = \frac{k(n)}{(n-1)RP_0(R)}\Big[1 + O\Big(\frac{n^2}{R}\Big)\Big]. \tag{7.5}$$

Clearly $\frac{k(n)}{n-1} = \frac{1}{4}(1 + \frac{2n^2+4n+3}{(n^2-1)(2n+3)}) = \frac{1}{4}(1 + \frac{2}{2n+3} + \frac{4n+5}{(n^2-1)(2n+3)})$ and the sequence $\frac{2}{2n+3}$ is monotone decreasing. Since also, for $x > 2$, the function $f(x) = \frac{4x+5}{(x^2-1)(2x+3)}$ satisfies $f'(x) < 0$, the assertion (7.4) follows.

To prove (7.2) note that if $n_*(R_j)$ remains bounded for a sequence of $R_j \to \infty$, then (7.5) is valid uniformly for $R = R_j$. By taking a subsequence, we may assume that $n_*(R_j)$ is a number n_* independent of j. By the previous analysis, if $n > n_*$ and R_j is large enough, then $M_n(R_j) < M_{n_*}(R_j)$, which is a contradiction. ∎

We next state (7.3) more precisely.

THEOREM 7.2. *If* $0 < R < R^\#$, *where* $R^\# \approx 2.601086$, *then* $M_n(R)$ *is monotone increasing for* $n \geq 4$, *and*

$$M_3(R) < M_2(R) < M_4(R) \quad \textit{for } 0 < R < R^\#. \tag{7.6}$$

Proof. The inequality (7.6) can be verified numerically with $R^\#$ uniquely determined by the equation $M_2(R^\#) = M_4(R^\#)$. The monotonicity of $M_n(R)$ for $0 < R < R^\#$ and $n \geq 4$ can be rigorously established using (B.16), (B.20), and (B.22); the details are omitted. ∎

8. Bifurcation. We shall apply the following Crandall–Rabinowitz theorem.

THEOREM 8.1 (see [11, Theorem 1.7]). *Let* X, Y *be real Banach spaces and* $F(x,\mu)$ *a* C^p *map,* $p \geq 3$, *of a neighborhood* $(0,\mu_0)$ *in* $X \times \mathbb{R}$ *into* Y. *Suppose*

(i) $F(0,\mu) = 0$ *for all* μ *in a neighborhood of* μ_0.
(ii) $\ker F_x(0,\mu_0)$ *is one dimensional space, spanned by* x_0.
(iii) $\operatorname{Im} F_x(0,\mu_0) = Y_1$ *has codimension* 1.
(iv) $F_{\mu x}(0,\mu_0)x_0 \notin Y_1$.

Then $(0,\mu_0)$ *is a bifurcation point of the equation* $F(x,\mu) = 0$ *in the following sense: In a neighborhood of* $(0,\mu_0)$ *the set of solutions of* $F(x,\mu) = 0$ *consists of two* C^{p-2} *smooth curves* Γ_1 *and* Γ_2 *which intersect only at the point* $(0,\mu_0)$; Γ_1 *is the curve* $(0,\mu)$ *and* Γ_2 *can be parameterized as follows:*

$$\Gamma_2 : (x(\varepsilon), \mu(\varepsilon)),\ |\varepsilon| \textit{ small}, \qquad (x(0),\mu(0)) = (0,\mu_0),\ x'(0) = x_0.$$

As in [13] we introduce the Banach spaces: for $k \geq 3$,

$$\begin{aligned}
X^{k+\alpha} &= \left\{\widetilde{R} \in C^{k+\alpha}(\Sigma),\ \widetilde{R} \text{ is } \pi\text{-periodic in } \theta,\ 2\pi\text{-periodic in } \phi\right\},\\
X_1^{k+\alpha} &= \text{closure of the linear space spanned by } \{Y_{j,0}(\theta),\ j = 0,1,2,\dots\} \text{ in } X^{k+\alpha},\\
X_2^{k+\alpha} &= \text{closure of the linear space spanned by } \{Y_{j,0}(\theta),\ j = 0,2,4,\dots\} \text{ in } X^{k+\alpha}.
\end{aligned}$$

Notice that (A.8) implies that $Y_{j,0}(\pi-\theta) = Y_{j,0}(\theta)$ if and only if j is even. Thus $X_2^{k+\alpha}$ coincides with the subspace of the $C^{k+\alpha}(\Sigma)$-closure of the smooth functions consisting of those functions u that are independent of ϕ and satisfies $u(\theta) = u(\pi-\theta)$. This property ensures that F defined by (3.1) satisfies $F : X_2^{k+\alpha} \to X_2^{k-1+\alpha}$.

We shall take $X = X_2^{k+\alpha}$ and $Y = X_2^{k-1+\alpha}$ and define F by (3.1) for any $\widetilde{R} \in X_2^{k+\alpha}$. It is clear that

$$\ker[F_{\widetilde{R}}(0,\mu)] = \{0\} \quad \text{if } \mu/\gamma \neq M_2, M_4, M_6, \dots. \tag{8.1}$$

If, for some $R > 0$ and some even integer $n \geq 2$,

$$M_n(R) \neq M_{2m}(R) \quad \text{for all } 2m \neq n, \tag{8.2}$$

then $\ker[F_{\widetilde{R}}(0,\mu)] = \operatorname{span}\{Y_{n,0}\}$. Since $[F_{\widetilde{R}}(0,M_n)]Y_{2m,0} = f(2m,n)Y_{2m,0}$, where $f(2m,n) \neq 0$ for $2m \neq n$, we have

$$\operatorname{Im}[F_{\widetilde{R}}(0,M_n)] \oplus \operatorname{span}\{Y_{n,0}\} = Y,$$

so that the codimension of $Y_1 = \mathrm{Im}[F_{\widetilde{R}}(0, M_n)]$ is 1. From Lemma 4.1 we also have

$$[F_{\mu\widetilde{R}}(0, M_n)]Y_{n,0} = R^2 P_0(R)[P_1(R) - P_n(R)]\Big(1 + \frac{n}{2n^2+4n+3}\Big)Y_{n,0} \notin Y_1.$$

Thus all the assumptions of Theorem 8.1 are satisfied and we obtain the following theorem.

THEOREM 8.2. *For even $n \geq 2$, if $R = R_S$ is such that* (8.2) *is satisfied, then the point* $(0, M_n)$ *is a bifurcation point for the problem* (1.9)–(1.14)*, and the corresponding branch of solutions has free boundaries of the form $r = R + \varepsilon Y_{l,0}(\theta) + O(\varepsilon^2)$.*

Remark 8.1. From Theorems 7.1 and 7.2 it follows that for any n, m larger than 3, there exists at least one R such that $M_n(R) = M_m(R)$. Therefore, the assumption (8.2) for $R = R_S$ cannot be dropped. On the other hand, for any given $n \geq 2$, (8.2) is satisfied for all R with the exception of a discrete sequence.

Remark 8.2. The choice of $X_2^{k+\alpha}$ is necessary in order to be able to use the Crandall–Rabinowitz theorem (this choice is also necessary in the proof of [14, Theorem 2] in order to apply the Crandall–Rabinowitz theorem). Indeed, if we replace $X_2^{k+\alpha}$ by $X_1^{k+\alpha}$, then, for $n \geq 2$, the subspace $\ker[F_{\widetilde{R}}(0, \mu)] = \mathrm{span}\{Y_{n,0}, Y_{1,0}\}$ is of dimension 2, so that the condition (ii) in Theorem 8.1 is not satisfied. Note that $Y_{1,0}$ is the kernel element which corresponds to translation.

Remark 8.3. Theorem 8.2 establishes bifurcation only for even integer $n \geq 2$. One can probably also establish bifurcation for any odd integer $n \geq 3$ by the method of expansion into the power series of ε as done in [13]. However, we should be able to establish bifurcation for any odd $n \geq 3$ also by applying the Crandall–Rabinowitz theorem in a more delicate manner, working with the space

$$M^{k+\alpha} = \text{closure of the linear space spanned by } \{Y_{j,0}(\theta),\ j = 0, 2, 3, 4, \dots\} \text{ in } X^{k+\alpha}$$

for $k \geq 3$. Here we face the problem that F does not map $M^{k+\alpha}$ into $M^{k-1+\alpha}$, but we propose to shift the center of the system in order to eliminate the mode of $Y_{1,0}$ and so that the modified F will map $M^{k+\alpha}$ into $M^{k-1+\alpha}$. Since the image of the shift is of higher order than $\varepsilon Y_{1,0}$, we should be able to use a fixed point theorem in order to make a shift which will eliminate the term with mode $(1, 0)$. The mapping $\widetilde{F}$ composed of F followed by this shift of the origin will map $M^{k+\alpha}$ onto $M^{k-1+\alpha}$, and the Crandall–Rabinowitz theorem could then be applied to $\widetilde{F}$. Indeed, a similar argument was carried out in [21].

9. Conclusion: Biological interpretation. Although the tumor model analyzed in this paper is quite simple, we may nevertheless draw some interesting biological conclusions from the mathematical results. Tumors grown in culture are typically spherical. However, tumors in vivo can have a variety of shapes. In particular, invasion of tumors into their surrounding stroma is associated with growth of protrusions, or "fingers." In our model, these protrusions are expressed by the shape $r = R_S + \varepsilon Y_{n,0}(\theta) + O(\varepsilon^2)$ of the free boundary; the number of protrusions is proportional to n.

The aggressiveness of a tumor is measured by two parameters, μ and γ. The parameter μ is the proliferation rate; the larger the μ is the more aggressive the tumor is. The parameter γ is the cell-to-cell adhesiveness; it plays an important role in keeping the tumor cohesive (see [5, 6, 8]). A smaller value of γ enables the tumor to develop fingers more easily and thus be more prone to invasion. In our model the two parameters appear as a quotient μ/γ. As this parameter increases, the tumor will lose its spherical shape, develop fingers, and become invasive.

The ability of a tumor to invade the surrounding tissue depends also on the material properties of its surroundings. If the tissue is a porous medium, then, according to [13], the smallest value of μ/γ which generates protrusions is $M_2(R_S)$, at which time the tumor will have just three protrusions, no matter how large the radius R_S is. In contrast, in fluid-like tissue as in the present paper, the smallest value of μ/γ which generates protrusions is $M_{n_*}(R_S)$, where $n_* \to \infty$ as $R_S \to \infty$. Thus, when a large spherical tumor develops protrusions, it does so right away with a large number of protrusions, namely, with a number proportional to $n_*(R_S)$. This makes the tumor invasion more hazardous, since it increases the probability that one or several of the many invasive protrusions will reach a blood vessel and lead to metastasis.

Appendix A. Vector spherical harmonics. We use the notation $\vec{e}_r$, $\vec{e}_\theta$, $\vec{e}_\phi$ for the unit normal vectors in the r, θ, ϕ directions, respectively; here $0 \le r < \infty$, $0 \le \theta \le \pi$, $0 \le \phi \le 2\pi$. Then, written in the Cartesian coordinates in $\mathbb{R}^3$,

$$\vec{e}_r = \sin\theta\cos\phi\,\vec{e}_1 + \sin\theta\sin\phi\,\vec{e}_2 + \cos\theta\,\vec{e}_3, \tag{A.1}$$

$$\vec{e}_\theta = \cos\theta\cos\phi\,\vec{e}_1 + \cos\theta\sin\phi\,\vec{e}_2 - \sin\theta\,\vec{e}_3, \tag{A.2}$$

$$\vec{e}_\phi = -\sin\phi\,\vec{e}_1 + \cos\phi\,\vec{e}_2, \tag{A.3}$$

where $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ is the standard basis in $\mathbb{R}^3$ in Cartesian coordinates, i.e., $\vec{e}_1 = (1,0,0)^T$, $\vec{e}_2 = (0,1,0)^T$, $\vec{e}_3 = (0,0,1)^T$. The gradient is given by

$$\nabla_x = \vec{e}_r\,\partial_r + \vec{e}_\theta\,r^{-1}\,\partial_\theta + \vec{e}_\phi\,(r\sin\theta)^{-1}\,\partial_\phi \equiv \vec{e}_r\,\partial_r + r^{-1}\nabla_\omega,$$

and

$$\vec{e}_r \times \vec{e}_\theta = \vec{e}_\phi, \quad \vec{e}_\phi \times \vec{e}_r = \vec{e}_\theta, \quad \vec{e}_\theta \times \vec{e}_\phi = \vec{e}_r, \tag{A.4}$$

$$\partial_r\vec{e}_r = 0, \quad \partial_r\vec{e}_\theta = 0, \quad \partial_r\vec{e}_\phi = 0, \tag{A.5}$$

$$\partial_\theta\vec{e}_r = \vec{e}_\theta, \quad \partial_\theta\vec{e}_\theta = -\vec{e}_r, \quad \partial_\theta\vec{e}_\phi = 0, \tag{A.6}$$

$$\partial_\phi\vec{e}_r = \sin\theta\,\vec{e}_\phi, \quad \partial_\phi\vec{e}_\theta = \cos\theta\,\vec{e}_\phi, \quad \partial_\phi\vec{e}_\phi = -\sin\theta\,\vec{e}_r - \cos\theta\,\vec{e}_\theta. \tag{A.7}$$

The spherical harmonic $Y_{l,m}(\theta,\phi)$ is defined by

$$Y_{l,m}(\theta,\phi) = (-1)^m\sqrt{\frac{(2l+1)(l-m)!}{2(l+m)!}}\,P_l^m(\cos\theta)\frac{e^{im\phi}}{\sqrt{2\pi}} \quad (m = -l,\dots,l), \tag{A.8}$$

where $P_l^m(z) = \frac{1}{2^l l!}(1-z^2)^{m/2}\frac{d^{l+m}}{dz^{l+m}}(z^2-1)^l$. The family of functions $\{Y_{l,m}\}$ forms a complete orthonormal basis for $L^2(\Sigma)$, where Σ is the unit sphere, and

$$\Delta_\omega Y_{l,m} = -l(l+1)Y_{l,m}, \tag{A.9}$$

where $\Delta_\omega = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\frac{\partial}{\partial\theta}) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}$ is the Laplace operator on Σ. The vector spherical harmonics are defined by

$$\begin{aligned}\vec{V}_{l,m} = \vec{e}_r\Big\{-\Big(\frac{l+1}{2l+1}\Big)^{1/2}Y_{l,m}\Big\} + \vec{e}_\theta\Big\{\frac{1}{[(l+1)(2l+1)]^{1/2}}\frac{\partial Y_{l,m}}{\partial\theta}\Big\}\\ + \vec{e}_\phi\Big\{\frac{imY_{l,m}}{[(l+1)(2l+1)]^{1/2}\sin\theta}\Big\},\end{aligned} \tag{A.10}$$

$$\vec{X}_{l,m} = \vec{e}_\theta\Big\{\frac{-mY_{l,m}}{[l(l+1)]^{1/2}\sin\theta}\Big\} + \vec{e}_\phi\Big\{\frac{-i}{l(l+1)]^{1/2}}\frac{\partial Y_{l,m}}{\partial\theta}\Big\}, \tag{A.11}$$

$$\vec{W}_{l,m} = \vec{e}_r\Big\{\Big(\frac{l}{2l+1}\Big)^{1/2} Y_{l,m}\Big\} + \vec{e}_\theta\Big\{\frac{1}{[l(2l+1)]^{1/2}}\frac{\partial Y_{l,m}}{\partial\theta}\Big\} + \vec{e}_\phi\Big\{\frac{imY_{l,m}}{[l(2l+1)]^{1/2}\sin\theta}\Big\}. \tag{A.12}$$

The family of vector spherical harmonics $\{\vec{V}_{l,m}, \vec{X}_{l,m}, \vec{W}_{l,m}\}$ forms a complete orthonormal basis for $(L^2(\Sigma))^3$. We shall need the well-known formulas (see [29] or [26])

$$\nabla Y_{l,m} = \frac{1}{r}\nabla_\omega Y_{l,m} \;=\; \frac{l}{r}\Big(\frac{l+1}{2l+1}\Big)^{1/2}\vec{V}_{l,m} + \frac{l+1}{r}\Big(\frac{l}{2l+1}\Big)^{1/2}\vec{W}_{l,m}, \tag{A.13}$$

$$\vec{e}_r Y_{l,m} = -\Big(\frac{l+1}{2l+1}\Big)^{1/2}\vec{V}_{l,m} + \Big(\frac{l}{2l+1}\Big)^{1/2}\vec{W}_{l,m}; \tag{A.14}$$

for any function $H(r)$,

$$\nabla[H(r)Y_{l,m}] = \Big(\frac{l+1}{2l+1}\Big)^{1/2}\Big[-\frac{dH}{dr} + \frac{l}{r}H\Big]\vec{V}_{l,m} + \Big(\frac{l}{2l+1}\Big)^{1/2}\Big[\frac{dH}{dr} + \frac{l+1}{r}H\Big]\vec{W}_{l,m}, \tag{A.15}$$

$$\operatorname{div}[H(r)\vec{V}_{l,m}] = -\Big(\frac{l+1}{2l+1}\Big)^{1/2}\Big[\frac{dH}{dr} + \frac{l+2}{r}H\Big]Y_{l,m}, \tag{A.16}$$

$$\operatorname{div}[H(r)\vec{X}_{l,m}] = 0, \tag{A.17}$$

$$\operatorname{div}[H(r)\vec{W}_{l,m}] = \Big(\frac{l}{2l+1}\Big)^{1/2}\Big[\frac{dH}{dr} - \frac{l-1}{r}H\Big]Y_{l,m}, \tag{A.18}$$

and

$$\Delta[H(r)Y_{l,m}] = L_l(H)Y_{l,m}, \tag{A.19}$$

$$\Delta[H(r)\vec{V}_{l,m}] = L_{l+1}(H)\vec{V}_{l,m}, \tag{A.20}$$

$$\Delta[H(r)\vec{X}_{l,m}] = L_l(H)\vec{X}_{l,m}, \tag{A.21}$$

$$\Delta[H(r)\vec{W}_{l,m}] = L_{l-1}(H)\vec{W}_{l,m}, \tag{A.22}$$

where

$$L_l = \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} - \frac{l(l+1)}{r^2}. \tag{A.23}$$

Appendix B. Bessel functions. We introduce the Bessel functions (see [12])

$$I_l(\xi) = \Big(\frac{\xi}{2}\Big)^l \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(l+k+1)}\Big(\frac{\xi}{2}\Big)^{2k} \tag{B.1}$$

and recall some well-known properties:

$$I_{1/2}(\xi) = \sqrt{2/(\pi\xi)}\;\sinh\xi, \tag{B.2}$$

$$I_{3/2}(\xi) = \sqrt{2/(\pi\xi)}\;\big(-\xi^{-1}\sinh\xi + \cosh\xi\big), \tag{B.3}$$

$$I_{5/2}(\xi) = \sqrt{2/(\pi\xi)}\;\big\{(3\xi^{-2}+1)\sinh\xi - 3\xi^{-1}\cosh\xi\big\}, \tag{B.4}$$

$$I_l''(\xi) + \xi^{-1}I_l'(\xi) - \big(1 + l^2\xi^{-2}\big)I_l(\xi) = 0, \tag{B.5}$$

$$I_l'(\xi) + l\xi^{-1}I_l(\xi) = I_{l-1}(\xi),\ \ l \ge 1, \tag{B.6}$$

$$I_l'(\xi) - l\xi^{-1}I_l(\xi) = I_{l+1}(\xi),\ \ l \ge 0, \tag{B.7}$$

$$I_{l-1}(\xi) - I_{l+1}(\xi) = 2l\xi^{-1}I_l(\xi),\ \ l \ge 1. \tag{B.8}$$

From (B.6) and (B.7) we obtain

$$\text{(B.9)} \qquad \Big(\frac{d}{dr} - \frac{l}{r}\Big)\frac{I_{l+1/2}(r)}{r^{1/2}} = \frac{I_{l+3/2}(r)}{r^{1/2}},$$

$$\text{(B.10)} \qquad \Big(\frac{d}{dr} + \frac{l+1}{r}\Big)\frac{I_{l+1/2}(r)}{r^{1/2}} = \frac{I_{l-1/2}(r)}{r^{1/2}}.$$

One can also easily verify that, for the operator L_l given in (A.23),

$$\text{(B.11)} \qquad L_l\big(I_{l+1/2}(r)/r^{1/2}\big) = I_{l+1/2}(r)/r^{1/2}.$$

From (B.7) we derive, for any complex number s,

$$\text{(B.12)} \qquad \frac{d}{dr}\Big[\frac{I_{l+1/2}(r\sqrt{s+1}\,)}{r^{1/2}}\Big] = \frac{l}{r^{3/2}}I_{l+1/2}(r\sqrt{s+1}\,) + \frac{\sqrt{s+1}}{r^{1/2}}I_{l+3/2}(r\sqrt{s+1}\,).$$

By differentiating (B.12) in r and using (B.6) and (B.7), we obtain

$$\text{(B.13)} \qquad \begin{aligned} \frac{d^2}{dr^2}\Big[\frac{I_{l+1/2}(r\sqrt{s+1}\,)}{r^{1/2}}\Big] \quad = \quad & \Big(\frac{s+1}{r^{1/2}} + \frac{l(l-1)}{r^{5/2}}\Big)I_{l+1/2}(r\sqrt{s+1}\,) \\ & - \frac{2\sqrt{s+1}}{r^{3/2}}I_{l+3/2}(r\sqrt{s+1}\,). \end{aligned}$$

We next introduce the functions

$$\text{(B.14)} \qquad P_n(\xi) = \frac{I_{n+3/2}(\xi)}{\xi I_{n+1/2}(\xi)}, \qquad n = 0, 1, 2, 3, \ldots.$$

From the preceding properties of the Bessel functions we can derive the following relations (see [20]). For ξ complex,

$$\text{(B.15)} \qquad P_0(\xi) = \xi^{-1}\coth\xi - \xi^{-2}, \qquad P_1(\xi) = (\xi\coth\xi - 1)^{-1} - 3\xi^{-2},$$

$$\text{(B.16)} \qquad P_n(\xi) = 1/\{\xi^2 P_{n+1}(\xi) + (2n+3)\}, \qquad P_n(0) = 1/(2n+3),$$

$$\text{(B.17)} \qquad (d/d\xi)P_n(\xi) = \xi^{-1} - (2n+3)\xi^{-1}P_n(\xi) - \xi P_n^2(\xi),$$

$$\text{(B.18)} \qquad |P_n(\xi)| \le \sqrt{2}P_n(|\xi|) \le \sqrt{2}/(2n+3) \quad \text{for } |\arg(\xi)| \le \pi/4,$$

and, for $r > 0$,

$$\text{(B.19)} \qquad P_n(r) > P_{n+1}(r), \ n \ge 0,$$

$$\text{(B.20)} \qquad \frac{2n+5}{r^2 + (2n+3)(2n+5)} < P_n(r) < \frac{r^2 + (2n+5)(2n+7)}{2(2n+5)r^2 + (2n+3)(2n+5)(2n+7)},$$

$$\text{(B.21)} \qquad \frac{1}{r} - \frac{n+1}{r^2} < P_n(r) < \frac{1}{r} - \frac{n+1}{r^2} + \frac{n^2+n+1}{2r^3},$$

$$\text{(B.22)} \qquad \frac{d}{dr}P_n(r) < 0.$$

Acknowledgment. We thank the referees for helpful suggestions.

REFERENCES

[1] J. A. Adam, *General aspect of modeling tumor growth and immune response*, in A Survey of Models for Tumor-Immune System Dynamics, J. A. Adam and N. Bellomo, eds., Birkhäuser, Boston, 1996, pp. 14–87.

[2] J. A. Adam and S. A. Maggelakis, *Diffusion regulated growth characteristics of a spherical prevascular carcinoma*, Bull. Math. Biol., 52 (1990), pp. 549–582.
[3] B. Bazaliy and A. Friedman, *Global existence and asymptotic stability for an elliptic-parabolic free boundary problem: An application to a model of tumor growth*, Indiana Univ. Math. J., 52 (2003), pp. 1265–1304.
[4] N. Britton and M. A. J. Chaplain, *A qualitative analysis of some models of tissue growth*, Math. Biosci., 113 (1993), pp. 77–89.
[5] H. M. Byrne, *The importance of intercellular adhesion in the development of carcinomas*, IMA J. Math. Appl. Med. Biol., 14 (1997), pp. 305–323.
[6] H. M. Byrne, *A weakly nonlinear analysis of a model of avascular solid tumor growth*, J. Math. Biol., 39 (1999), pp. 59–89.
[7] H. M. Byrne and M. A. J. Chaplain, *Growth of nonnecrotic tumors in the presence and absence of inhibitors*, Math. Biosci., 130 (1995), pp. 151–181.
[8] H. M. Byrne and M. A. J. Chaplain, *Modelling the role of cell-cell adhesion in the growth and development of carcinomas*, Math. Comput. Modelling, 12 (1996), pp. 1–17.
[9] H. M. Byrne and M. A. J. Chaplain, *Growth of nonnecrotic tumors in the presence and absence of inhibitors*, Math. Biosci., 135 (1996), pp. 187–216.
[10] M. A. J. Chaplain, *The development of a spatial pattern in a model for cancer growth*, in Experimental and Theoretical Advances in Biological Pattern Formation, H. G. Othmer, P. K. Maini, and J. D. Murray, eds., Plenum Press, New York, 1993, pp. 45–60.
[11] M. G. Crandall and P. H. Rabinowitz, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.
[12] A. Erdelyi, W. Magnus, F. Oberhettinger, and F. Tricomi, *Higher Transcendental Functions*, Vol. 2, McGraw–Hill, New York, 1953.
[13] M. Fontelos and A. Friedman, *Symmetry-breaking bifurcations of free boundary problems in three dimensions*, Asymptot. Anal., 35 (2003) pp. 187–206.
[14] M. Fontelos and A. Friedman, *Symmetry-breaking bifurcations of charged drops*, Arch. Ration. Mech. Anal., 172 (2004), pp. 267–294.
[15] S. J. H. Franks, H. M. Byrne, J. C. E. Underwood, and C. E. Lewis, *Biological inferences from a mathematical model of comedo ductal carcinoma in situ of the breast*, J. Theoret. Biol., 232 (2005), pp. 523–543.
[16] S. J. H. Franks, H. M. Byrne, J. R. King, J. C. E. Underwood, and C. E. Lewis, *Modelling the early growth of ductal carcinoma in situ of the breast*, J. Math. Biol., 47 (2003), pp. 424–452.
[17] S. J. H. Franks, H. M. Byrne, J. R. King, J. C. E. Underwood, and C. E. Lewis, *Modelling the growth of ductal carcinoma in situ*, Mathematical Medicine & Biology, 20 (2003), pp. 277–308.
[18] S. J. H. Franks and J. R. King, *Interaction between a uniformly proliferating tumor and its surroundings: Uniform material properties*, Mathematical Medicine & Biology, 20 (2003), pp. 47–89.
[19] A. Friedman, *A free boundary problem for a coupled system of elliptic, hyperbolic, and Stokes equations modeling tumor growth*, Interfaces Free Bound., 8 (2006), pp. 247–261.
[20] A. Friedman and B. Hu, *Bifurcation from stability to instability for a free boundary problem arising in a tumor model*, Arch. Ration. Mech. Anal., 180 (2006), pp. 292–330.
[21] A. Friedman and B. Hu, *Asymptotic stability for a free boundary problem arising in a tumor model*, J. Differential Equations, 227 (2006), pp. 598–639.
[22] A. Friedman and B. Hu, *Stability and instability of Liapounov–Schmidt and Hopf bifurcation for a free boundary problem arising in a tumor model*, Trans. Amer. Math. Soc., to appear.
[23] A. Friedman and F. Reitich, *Analysis of a mathematical model for growth of tumor*, J. Math. Biol., 38 (1999), pp. 262–284.
[24] A. Friedman and F. Reitich, *Symmetry-breaking bifurcation of analytic solutions to free boundary problems: An application to a model of tumor growth*, Trans. Amer. Math. Soc., 353 (2000), pp. 1587–1634.
[25] A. Friedman and F. Reitich, *Nonlinear stability of a quasi-static Stefan problem with surface tension: A continuation approach*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 30 (2001), pp. 341–403.
[26] A. Friedman and F. Reitich, *Quasi-static motion of capillary drop* II: *The three dimensional case*, J. Differential Equations, 186 (2002), pp. 509–557.
[27] H. P. Greenspan, *Models for the growth of a solid tumor by diffusion*, Stud. Appl. Math., 52 (1972), pp. 317–340.
[28] H. P. Greenspan, *On the growth of cell culture and solid tumors*, J. Theoret. Biol., 56 (1976), pp. 229–242.

[29] E. L. Hill, *The theory of vector spherical harmonics*, Amer. J. Phys., 222 (1954), pp. 211–214.
[30] S. A. Maggelakis and J. A. Adam, *Mathematical model for prevascular growth of a spherical carcinoma*, Math. Comput. Modelling, 13 (1990), pp. 23–38.
[31] D. L. S. McEwain and L. E. Morris, *Apoptosis as a volume loss mechanism in mathematical models of solid tumor growth*, Math. Biosci., 39 (1978), pp. 147–157.
[32] V. A. Solonnikov, *On quasistationary approximation in the problem of a capillary drop*, in Topics in Nonlinear Analysis, Progr. Nonlinear Differential Equations Appl. 35, J. Escher and G. Simonett, eds., Birkhäuser, Basel, 1999, pp. 643–671.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 195–209

SINGULAR PERTURBATION AS A SELECTION CRITERION FOR YOUNG-MEASURE SOLUTIONS*

M. LILLI†, T. J. HEALEY‡, AND H. KIELHÖFER†

Abstract. We prove existence of Young-measure solutions of an Euler–Lagrange equation arising from a one-dimensional nonconvex variational problem in nonlinear elasticity. In particular, we consider a physically reasonable stored-energy density W such that $W(x, \mu)$ goes to infinity for $\mu \searrow 0$ and $\mu \to \infty$. The selection criterion for the Young measure is a singular perturbation in form of an interfacial energy with capillarity coefficient ε. We first establish uniform a priori bounds on all solutions of the Euler–Lagrange equation, before passing to the limit for $\varepsilon \searrow 0$. Moreover, the singular perturbation allows us to characterize the support of the Young measure.

Key words. Young-measure solution, nonlinear elasticity, singular limits

AMS subject classifications. 34B15, 74N15, 74G55

DOI. 10.1137/060656978

1. Introduction.

This paper is motivated to a large extent by results presented in [11]. There we consider a two-phase model for an elastic solid in the presence of live-body forcing and interfacial or higher-gradient effects, the latter characterized by a small capillarity coefficient $\varepsilon > 0$. The existence of families of equilibria is routine, and the main goal in [11] is the analysis of those solutions in the limit $\varepsilon \searrow 0$. Assuming that the loading is bounded and everywhere nonnegative, we obtain uniform a priori bounds on solutions via maximum-principle arguments. An important by-product of that analysis is a certain monotonicity property for all solutions. In the end we obtain global weak solutions that satisfy the Maxwell condition.

In this work we pursue precisely the same question without presuming that the live-body forcing is bounded or satisfies a sign condition, the latter obviating any apparent maximum principle. Instead we impose very general, physically reasonable growth conditions on both the stored-energy function (double-well) and the body force function. Also in contrast to [11], we impose the pointwise unilateral constraint ensuring that all deformations of the bar are injective (with the concomitant infinite growth of the stored energy).

The outline of this work is as follows: After presenting our formulation in section 2, we obtain uniform a priori bounds on solutions of the regularized ($\varepsilon > 0$) problem in section 3. There we pay careful attention to the injectivity constraint and show that the strains are uniformly (pointwise) positive. The existence of solutions is fairly routine, which we summarize in section 4. In section 5 we examine the limit $\varepsilon \searrow 0$. We demonstrate strong convergence to a continuous displacement field and weak* convergence to a stress field and a strain field, the latter two of which are each

*Received by the editors April 11, 2006; accepted for publication (in revised form) December 6, 2006; published electronically May 10, 2007.
http://www.siam.org/journals/sima/39-1/65697.html

†Institut für Mathematik, Universität Augsburg, 86135 Augsburg, Germany (lilli@math.uni-augsburg.de, kielhofer@math.uni-augsburg.de) The first author thanks the Graduiertenkolleg “Nichtlineare Probleme in Analysis, Geometrie und Physik” at the University of Augsburg for financial support.

‡Department of Theoretical and Applied Mathematics, Cornell University, Ithaca, NY 14853 (tjh10@cornell.edu). This author was supported in part by NSF grant DMS-0406161, which is gratefully acknowledged.

characterized by the family of Young measures generated by the sequence of strains. The limiting strain is generally incompatible with the displacement field. Remarkably the limiting stress has a C^1 representation (within its equivalence class) that satisfies the force-balance equilibrium equation classically. Finally, we show that the family of Young measures reduces to a family of "Dirac" masses in stress field regimes where the stress is an injective function of strain (for "large enough" compressive and tensile stresses). In particular, within those regimes our limiting fields reduce to classical solutions.

L. C. Young first employed (what is now called) the Young measure in the study of problems from the calculus of variations for which a classical minimizer does not exist [23]. Since that time Young measures have become an important tool for characterizing highly oscillatory sequences of functions. Tartar was the first to introduce them in the context of measure-valued solutions of conservation laws (see [20]), and their use in the study of evolutionary PDEs is now well known, e.g., [7], [8], [16]. Existence of Young-measure solutions was proven for PDEs (see, e.g., [9], [19], [21]) and the Young measure serves also as an important tool to prove existence of classical solutions for quasilinear PDEs (see, e.g., [12], [14]). The original ideas of Young have also found great application in the "sharp interface" energetic theory of martensitic phase transformations, e.g., [3], [4]. Our motivation here is similar. However, we focus instead on the behavior of sequences of equilibrium solutions (of a regularized theory), albeit in the context of a simple model problem. In particular, our limiting stress field satisfies equilibrium (pointwise), which is typically not the case for minimizing sequences.

2. Formulation. We consider a one-dimensional elastic bar placed in a soft loading device. We follow the reasoning of [13]; see also [5] and [10].

Let $[0, 1]$ be the reference configuration in the undeformed state and let

$$u(x) \text{ be the placement of the bar;}$$

i.e., $u(x)$ is the position of the material point occupying position x in the undeformed state; $\mu \equiv u'$ is the stretch ratio. The bar is presumed elastic and inhomogeneous, which is reflected by an explicit dependence of the *stored-energy function* $W \in C^2([0, 1] \times \mathbb{R}^+)$ upon the spatial variable x in the interval $[0, 1]$. We assume the function $W(x, \cdot)$ to be a double-well potential. We also require

$$\lim_{\mu \searrow 0} W(x, \mu) = \infty, \quad \lim_{\mu \to \infty} W(x, \mu) = \infty. \tag{2.1}$$

The first condition reflects the fact that an infinite amount of energy is required to compress any finite segment of the bar to zero length, while the second accounts for the fact that an infinite amount of energy is needed to stretch any finite segment of the bar to infinite length.

The bar is presumed "stable" in the undeformed state; i.e., for every $x \in [0, 1]$,

$$W(x, 1) = W_\mu(x, 1) = 0 \text{ and } W_{\mu\mu}(x, 1) > 0. \tag{2.2}$$

The *stress* is defined by $\sigma(x, \mu) := W_\mu(x, \mu)$. Moreover, we impose the following assumptions on W: For arbitrary $x \in [0, 1]$ we have

$$W(x, \mu) \geq W(x, 1) = 0 \text{ for every } \mu \in \mathbb{R}^+. \tag{2.3}$$

Furthermore, there exist for every $x \in [0,1]$ numbers $c_1(x), c_2(x) \in \mathbb{R}$ with $c_j \in C^1(0,1)$ for $j = 1, 2$ such that

$$W_{\mu\mu}(x,\mu) \begin{cases} < 0 \text{ for } \mu \in (c_1(x), c_2(x)), \\ > 0 \text{ otherwise.} \end{cases} \tag{2.4}$$

Let $B : \mathbb{R}_0^+ \times [0,1] \to \mathbb{R}$ denote a loading potential delivering a live-body force

$$-\frac{\partial B}{\partial u}(u,x) =: b(u,x),$$

where

$$b \in C(\mathbb{R}_0^+ \times [0,1]).$$

We require the following growth conditions for σ and b:

(i) We assume for all $x \in [0,1]$

$$\lim_{\mu\to\infty} \frac{W_\mu(x,\mu)\mu}{\mu^{p+1}} \geq K > 0, \tag{2.5}$$

where K is independent of x, $p \in \mathbb{N}$, and

(ii)

$$|b(u,x)| \leq c_3 |u|^r + c_4 \tag{2.6}$$

for some $0 < r < p$, c_3, $c_4 \geq 0$. Moreover, both constants can be chosen independently of x, u.

The left end of the bar is fixed, viz., $u(0) = 0$. The free end of the bar is subjected to a concentrated force τ. The total *potential energy* of the bar is given by

$$\begin{aligned} J(u) &:= \int_0^1 [W(x,u') + B(u,x)]\, dx - \tau u(1), \\ u(0) &= 0. \end{aligned} \tag{2.7}$$

The corresponding Euler–Lagrange equation is

$$\begin{aligned} &\frac{d}{dx}[\sigma(x,u')] + b(u,x) = 0, \\ &u(0) = 0, \quad \text{and} \quad \sigma(1,u'(1)) = \tau. \end{aligned} \tag{2.8}$$

Note that (2.7) has no global minimizer, in general, due to the fact that W is non-convex in u'. In particular, existence of a global minimizer of problem (2.7) cannot be guaranteed by the direct methods of the calculus of variations (see [6] for details). For the same reason, the boundary value problem (2.8) is singular, obviating any systematic solution strategy.

Instead we introduce a "relaxed" variational problem by adding an additional strain gradient term, intended to model interfacial energy:

$$\begin{aligned} J_\varepsilon(u) &:= \int_0^1 \left(\frac{\varepsilon}{2}u''^2 + W(x,u') + B(u,x)\right) dx - \tau u(1), \\ u(0) &= 0, \end{aligned} \tag{2.9}$$

where $\varepsilon > 0$ is a small parameter. The Euler–Lagrange equation of equilibrium is the fourth-order equation

$$-\varepsilon u^{(4)} + \frac{d}{dx}\left[\sigma(x,u')\right] + b(u,x) = 0 \tag{2.10}$$

with boundary conditions

$$u(0) = u''(0) = u''(1) = 0, \qquad \varepsilon u'''(1) = \sigma(1,u'(1)) - \tau. \tag{2.11}$$

Integration of (2.10) yields the system

$$\begin{aligned} &u' = z, \\ &-\varepsilon z'' + \sigma(x,z) = \int_x^1 b(u(\eta),\eta)d\eta + \tau, \\ &u(0) = z'(0) = z'(1) = 0. \end{aligned} \tag{2.12}$$

In contrast to (2.7) and (2.8), (2.9) and (2.12) are amenable to fairly standard existence methods for each $\varepsilon > 0$, which we quote in section 4.

3. A priori bounds. In this section we prove some a priori bounds for solutions of (2.12) (provided they exist) which hold independently of ε. Our first result states that the derivative of all solutions of (2.12) is uniformly bounded away from the singularity of W, namely, 0, which generalizes in a certain sense a result of Antman and Brezis; see [1].

THEOREM 1. *Let W and b satisfy the growth conditions* (2.1), (2.5), *and* (2.6). *Let (u_ε) be a solution of* (2.12) *and set $z_\varepsilon := u'_\varepsilon$. Then there exists some $\delta > 0$ such that for every $\varepsilon > 0$ we have*

$$z_\varepsilon(x) > \delta$$

for all $x \in [0,1]$.

Proof. We prove by contradiction and therefore we assume sequences $(\varepsilon_n)_n$, $(x_n)_n \in [0,1]$, and $(z_{\varepsilon_n})_n := (z_n)_n$ with

$$\varepsilon_n \searrow 0, \quad z_n(x_n) \searrow 0.$$

We separate the proof into several steps.

Step 1. We claim $||z_n||_{L^{p+1}(0,1)} \to \infty$. Without loss of generality we assume that x_n is a location of the global minimum of z_n and hence $z''_n(x_n) \geq 0$ for every $n \in \mathbb{N}$. Equation (2.12) and the property (2.1) yield

$$\int_{x_n}^1 b(u_n(\eta),\eta)d\eta + \tau \leq \sigma(x_n, z_n(x_n)) \to -\infty. \tag{3.1}$$

Assumption (2.6) with $c_3 > 0$ then yields

$$||u_n||_{L^r(0,1)} \to \infty.$$

(If $c_3 = 0$ in (2.6), then (3.1) is a contradiction and Theorem 1 is proved.) By the fundamental theorem of calculus and with Hölder's inequality, $r < p$, we obtain the assertion.

Step 2. We define

$$A_n := \{x \in [0,1] \mid z_n(x) > 1\} := \bigcup_{j=1}^{k(n)} (x_1^j(n), x_2^j(n)).$$

Observe that z_n is monotonically increasing at $x_1^j(n)$ and monotonically decreasing at $x_2^j(n)$ for every $j = 1, \dots, k(n)$ and $z_n'(0) = z_n'(1) = 0$, which yields

$$z_n'(x_1^j(n)) \geq 0, \quad z_n'(x_2^j(n)) \leq 0 \tag{3.2}$$

for all j. Multiplying (2.12) by z_n and integrating over A_n give

$$\int_{A_n} \varepsilon_n (z_n')^2 dx - \sum_j \varepsilon_n z_n z_n' |_{x_1^j}^{x_2^j} = \int_{A_n} \left(-\sigma(x, z_n) + \int_x^1 b(u_n(\eta), \eta) d\eta + \tau \right) z_n dx. \tag{3.3}$$

Note that the left-hand side of (3.3) is nonnegative by (3.2), and hence

$$\int_{A_n} \sigma(x, z_n) z_n dx \leq \int_{A_n} \left[\int_x^1 \left(b(u_n(\eta), \eta) d\eta + \tau \right) z_n \right] dx \tag{3.4}$$

for every $n \in \mathbb{N}$.

Step 3. From Step 1 we deduce

$$||z_n||_{L^{p+1}(A_n)} \to \infty. \tag{3.5}$$

Multiplying (3.4) by $\frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}}$ we obtain for every $n \in \mathbb{N}$

$$\begin{aligned} &\frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \int_{A_n} \sigma(x, z_n) z_n \, dx \\ &\qquad \leq \frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \int_{A_n} \left(\int_x^1 b(u_n(\eta), \eta) d\eta + \tau \right) z_n dx. \end{aligned} \tag{3.6}$$

We claim the following.

(a) *The right side of* (3.6) *converges to* 0 *for* $n \to \infty$. By assumption (2.6) and Hölder's inequality we obtain

$$\begin{aligned} &\left| \int_{A_n} \left(\int_x^1 b(u_n(\eta), \eta) d\eta z_n \right) dx \right| \leq \int_{A_n} \int_x^1 |b(u_n(\eta), \eta)| \, d\eta \, |z_n| \, dx \\ &\leq \int_0^1 |b(u_n(\eta), \eta)| \, d\eta \int_{A_n} |z_n| \, dx \leq \int_0^1 (c_3 |u_n|^r + c_4) \, d\eta \, ||z_n||_{L^1(A_n)} . \end{aligned}$$

Due to the continuous imbedding $L^{p+1}(0,1) \hookrightarrow L^r(0,1)$, the inequality $||u_n||_{L^r(0,1)} \leq ||z_n||_{L^r(0,1)}$, and $z_{n|(0,1)\setminus A_n} < 1$, the following holds:

$$\begin{aligned} &\int_0^1 (c_3 |u_n|^r + c_4) \, d\eta \, ||z_n||_{L^1(A_n)} \leq \left(c_3 \, ||z_n||^r_{L^{p+1}(0,1)} + c_4 \right) ||z_n||_{L^{p+1}(A_n)} \\ &\leq \left(c_3 + c_3 \, ||z_n||^r_{L^{p+1}(A_n)} + c_4 \right) ||z_n||_{L^{p+1}(A_n)} \\ &\leq c_3 \, ||z_n||^{r+1}_{L^{p+1}(A_n)} + (c_3 + c_4) \, ||z_n||_{L^{p+1}(A_n)} . \end{aligned}$$

In view of $r < p$ and (3.5), we see that assertion (a) is proved.

(b) *We have*

$$\lim_{n\to\infty} \frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \int_{A_n} \sigma(x, z_n) z_n \, dx \geq K > 0 \tag{3.7}$$

with K given in (2.5). For any $\varrho > 0$, (2.5) implies the existence of $t_o = t_o(\varrho)$ such that

$$\sigma(x, t)t - K\, |t|^{p+1} > -\frac{\varrho}{2}\, |t|^{p+1} \tag{3.8}$$

for every $t > t_o$. Due to (3.8) we get for an arbitrary $n \in \mathbb{N}$ and for all $x \in [0, 1]$ with $z_n(x) > t_o$ the estimate

$$\frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \Big(\sigma(x, z_n(x))z_n(x) - K\, |z_n(x)|^{p+1}\Big) > -\frac{\varrho}{2} \frac{|z_n(x)|^{p+1}}{||z_n||^{p+1}_{L^{p+1}(A_n)}}. \tag{3.9}$$

Furthermore, we have by the continuity of $\sigma(x, \cdot)$ for every $x \in [0, 1]$ with $z_n(x) \in [1, t_o]$

$$\frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \Big|\sigma(x, z_n(x))z_n(x) - K\, |z_n(x)|^{p+1}\Big| \leq \frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} C.$$

Note that the constant C depends only on ϱ. Hence we deduce the existence of $n_1 = n_1(\varrho)$ such that for every $n > n_1$ we have

$$\frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \Big(\sigma(x, z_n(x))z_n(x) - K\, |z_n(x)|^{p+1}\Big) > -\frac{\varrho}{2} \tag{3.10}$$

for every $x \in [0, 1]$ with $z_n(x) \in [1, t_o]$. We define

$$B_n := \{x \in [0, 1] \mid z_n(x) > t_o\}.$$

Then we obtain by virtue of (3.9) and (3.10)

$$\begin{aligned}
&\frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \int_{A_n} (\sigma(x, z_n)z_n \, dx - K \\
&= \frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \left(\int_{A_n} (\sigma(x, z_n)z_n - K\, |z_n(x)|^{p+1}) \, dx\right) \\
&= \frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \left[\left(\int_{A_n\cap B_n} + \int_{A_n\cap B_n^c}\right) \Big(\sigma(x, z_n)z_n - K\, |z_n|^{p+1}\Big) \, dx\right] \\
&\geq -\frac{\varrho}{2} \frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} ||z_n||^{p+1}_{L^{p+1}(A_n)} - \frac{\varrho}{2} = -\varrho.
\end{aligned} \tag{3.11}$$

Because $\varrho > 0$ was chosen arbitrarily, (3.7) holds.

Now by virtue of (3.6) we get

$$\begin{aligned}
0 < K &\leq \lim_{n\to\infty} \left(\frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \int_{A_n} \sigma(x, z_n)z_n \, dx\right) \\
&\leq \lim_{n\to\infty} \left(\frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \int_{A_n} \left(\int_x^1 b(u_n(\eta), \eta)d\eta + \tau\right) z_n \, dx\right) = 0,
\end{aligned} \tag{3.12}$$

which is obviously a contradiction, and the proof is complete. ∎

In what follows we will always assume the growth conditions (2.5), (2.6). Next we consider an *arbitrary* sequence of solutions of (2.12) $(\varepsilon_n, u_n)_{n\in\mathbb{N}}$, $\varepsilon_n \searrow 0$ and we define $z_n := u_n'$. We now show that the sequence $(z_n)_{n\in\mathbb{N}}$ is uniformly bounded (in some appropriate function space).

LEMMA 2. *Let σ and b satisfy the above-mentioned growth conditions and let $(u_n)_{n\in\mathbb{N}}$ be a sequence of solutions of (2.12) for corresponding ε_n as $\varepsilon_n \searrow 0$. Then there exists a constant $C > 0$ independent of $n \in \mathbb{N}$ such that*

$$||z_n||_{L^{p+1}} \leq C. \tag{3.13}$$

Proof. Let $(u_n)_{n\in\mathbb{N}}$ be a solution of (2.12) and $z_n = u_n'$. We assume

$$||z_n||_{L^{p+1}(0,1)} \to \infty. \tag{3.14}$$

As in the proof of Theorem 1, we obtain

$$\begin{aligned} &\frac{1}{||z_n||^{p+1}_{L^{p+1}(0,1)}} \int_0^1 \sigma(x, z_n) z_n \, dx \\ &\qquad \leq \frac{1}{||z_n||^{p+1}_{L^{p+1}(0,1)}} \int_0^1 \left(\int_x^1 b(u_n(\eta), \eta) d\eta + \tau \right) z_n dx. \end{aligned} \tag{3.15}$$

In particular, as in Step 3 of the proof of Theorem 1 we have on the one hand that the right side of (3.15) converges to 0 for $n \to \infty$ and, on the other hand,

$$\lim_{n\to\infty} \frac{1}{||z_n||^{p+1}_{L^{p+1}(0,1)}} \int_0^1 \sigma(x, z_n) z_n \, dx \geq K > 0$$

with K given in (2.5), which is a contradiction. $\square$

We can actually prove a stronger result, which seems to be limited to the one-dimensional case.

LEMMA 3. *Let $(z_n)_{n\in\mathbb{N}}$ be a sequence as in Lemma 2. Then we obtain the following: The sequence $(z_n)_{n\in\mathbb{N}}$ is uniformly bounded in $C^0([0,1])$.*

Proof. Let $(z_n)_{n\in\mathbb{N}}$ be a sequence as in Lemma 2 and assume for any constant $K > 0$ the existence of a certain $x \in [0,1]$ and $n \in \mathbb{N}$ such that $z_n(x) > K$. In particular, we obtain $x_m \in [0,1]$ and n_m such that $z_{n_m}(x_m) > m$ for every $m \in \mathbb{N}$.

By the continuity of z_{n_m} we know that

$$x \mapsto z_{n_m}(x) \text{ has a global maximum in } [0,1],$$

and without loss of generality, we assume this maximum is at $x = x_m$. By properties (2.5) we deduce

$$\lim_{m\to\infty} \sigma(x_m, z_{n_m}(x_m)) = \infty. \tag{3.16}$$

Moreover, we have for every $m \in \mathbb{N}$

$$-\varepsilon_{n_m} z''_{n_m}(x_m) + \sigma(x_m, z_{n_m}(x_m)) = \int_{x_m}^1 b(u_{n_m}(\eta), \eta) d\eta + \tau. \tag{3.17}$$

The growth conditions (2.6) together with (3.13) imply for every $m \in \mathbb{N}$

$$\left| \int_{x_m}^1 b(u_{n_m}(\eta), \eta) d\eta \right| \leq c_3 \, ||u_{n_m}||^r_\infty + c_4 \leq c_3 \, ||u_{n_m}||^r_{1,p+1} + c_4 \leq c_5.$$

Therefore, we have by (3.16)

$$\varepsilon_{n_m} z''_{n_m}(x_m) \geq \sigma(x_m, z_{n_m}(x_m)) - c_5 \to \infty$$

for $m \to \infty$, and, in particular, we get $z''_{n_m}(x_m) > 0$ for m sufficiently large. On the other hand, we have $z'_{n_m}(x_m) = 0$ for every $m \in \mathbb{N}$ (note that this is also true if $x_m = 0$ or $x_m = 1$ by our boundary conditions). Hence x_m cannot be a maximum, which contradicts our assumption. ∎

LEMMA 4. *Let $(z_n)_{n\in\mathbb{N}}$ be a sequence as in Lemma 2. Then there exists a subsequence of $(z_n)_{n\in\mathbb{N}}$, not relabeled, such that $\varepsilon_n z''_n \rightharpoonup^* 0$ in $L^\infty(0,1)$.*

Proof. Let $(z_n)_{n\in\mathbb{N}}$ be a sequence with the above-mentioned property. To prove the assertion, we multiply (2.12) by z_n, integrate the equation over the unit interval, and integrate by parts to find

$$\int_0^1 \varepsilon_n (z'_n)^2 dx = \int_0^1 \left(\int_x^1 b(u_n(\eta), \eta) d\eta + \tau - \sigma(x, z_n) \right) z_n \, dx,$$

which proves immediately that

$$(\sqrt{\varepsilon_n} z'_n)_{n\in\mathbb{N}} \text{ is bounded in } L^2(0,1). \tag{3.18}$$

Let $\varphi \in C_0^\infty(0,1)$; then integration by parts and (3.18) yield

$$\left| \int_0^1 \varepsilon_n z''_n \varphi dx \right| \leq \sqrt{\varepsilon_n} \int_0^1 |\sqrt{\varepsilon_n} z'_n \varphi'| \, dx \leq \sqrt{\varepsilon_n} \, ||\sqrt{\varepsilon_n} z'_n||_{L^2} \, ||\varphi'||_{L^2} \to 0$$

for $n \to \infty$. Because φ was chosen arbitrarily, we have $\varepsilon_n z''_n \rightharpoonup 0$ in $L^1(0,1)$. Since $(\varepsilon_n z''_n)_n$ is also uniformly bounded in $L^\infty(0,1)$ by (2.12) and Lemma 3, the assertion follows. ∎

4. Critical points of the singular perturbed Euler–Lagrange equation. We have now established various properties of solutions of (2.12) presuming the latter exist. There are at least two well-known approaches to existence, each of which we now briefly discuss. First we may examine the energy functional (2.9), which is readily seen to be weakly lower semicontinuous and coercive over $W^{2,2}(0,1)$. Hence the existence of a minimizer is ensured. That the minimizer also satisfies the weak form of the Euler–Lagrange equation requires a rather delicate argument [1], associated with the pointwise constraint $\mu = u' > 0$, accompanied by the growth condition (2.1). A standard bootstrap argument then shows that the minimizer is actually a classical solution. However, in the context of our problem, the methods of [1] do not provide a uniform bound as in Theorem 1.

Another approach to existence follows along the lines of [11]. We first embed the loading "b" within a one-parameter family and then obtain global branches of solutions. The only difference here is our pointwise restriction $u' > 0$, which does not alter the arguments in [11] in view of Theorem 1.

5. Generalized solutions. In this section we consider a sequence $(z_n)_{n\in\mathbb{N}}$, where $z_n = u'_n$ and u_n is a solution of (2.12) for corresponding ε_n, with $\varepsilon_n \searrow 0$. By virtue of Lemma 3, the sequence $(z_n)_{n\in\mathbb{N}}$ is uniformly bounded in $L^\infty(0,1)$. As such it has a weak* limit, say, z. The uniform boundedness also yields the existence of a family of probability measures, called Young measures,

$$\nu : [0,1] \to \text{Prob}(\mathbb{R}),$$

such that

$$z(x) = \langle \nu_x, Id \rangle \equiv \int_{\mathbb{R}} \rho \, d\nu_x(\rho) \tag{5.1}$$

for a.e. $x \in [0,1]$ (see [2] for details). Recall from Theorem 1 that $z_n(x) > \delta > 0$ for all $x \in [0,1]$ and every n, from which we conclude that

$$z(x) \geq \delta \quad \text{a.e.}$$

In view of Lemma 2 we see that the sequence $(u_n)_{n\in\mathbb{N}}$ is uniformly bounded in $W^{1,p+1}([0,1])$. Thus, by compact imbedding $(u_n)_{n\in\mathbb{N}}$ has a subsequence that converges uniformly to some $u \in C^0([0,1])$. We now prove the following theorem.

THEOREM 5. *Let ν be the Young measure generated by a sequence $(z_n)_{n\in\mathbb{N}}$, where z_n is a solution of* (2.12) *for corresponding ε_n. Then ν satisfies the equation*

$$\langle \nu_x, \sigma(x,\cdot)\rangle = \int_{\mathbb{R}^+} \sigma(x,\rho) d\nu_x(\rho) = \int_x^1 b(u(\eta),\eta)d\eta + \tau \tag{5.2}$$

for a.e. $x \in [0,1]$, where $u \in W^{1,\infty}(0,1) \cap \{v \mid v(0) = 0\}$ is the limit of $(u_n)_{n\in\mathbb{N}}$—in particular, the function

$$x \mapsto \langle \nu_x, \sigma(x,\cdot)\rangle \in C^1(0,1). \tag{5.3}$$

Proof. We consider a sequence $(u_n, z_n)_{n\in\mathbb{N}}$ of solutions of (2.12) for corresponding ε_n with $\varepsilon_n \searrow 0$. Accordingly, for every $n \in \mathbb{N}$ we have

$$-\varepsilon_n z_n'' + \sigma(x, z_n) = \int_x^1 b(u_n(\eta),\eta)d\eta + \tau.$$

By Theorem 1, Lemma 3, and the smoothness of $\sigma(\cdot)$, we see that the sequence of stresses $(\sigma(\cdot, z_n))_{n\in\mathbb{N}}$ is also uniformly bounded in $L^\infty(0,1)$. Accordingly, it too possesses a weak limit, which, by the fundamental theorem for Young measures (cf. [2], [17]), has the following characterization:

$$\sigma(x, z_n) \rightharpoonup^* \langle \nu_x, \sigma(x,\cdot)\rangle \quad \text{in } L^\infty(0,1).$$

Also,

$$\int_x^1 b(u_n(\eta),\eta)d\eta \to \int_x^1 b(u(\eta),\eta)d\eta$$

uniformly. Taking into account Lemma 4, we arrive at (5.2). Finally, the continuity of both $b(\cdot)$ and u shows there is a C^1 member within the equivalence class $\langle \nu_x, \sigma(x,\cdot)\rangle$. $\square$

In summary, our selection procedure delivers a displacement field

$$u_n \to u \quad \text{in } C^0([0,1]),$$

a strain field

$$u_n' \rightharpoonup^* z = \langle \nu_x, Id\rangle \quad \text{in } L^\infty(0,1),$$

and a stress field

$$\sigma(x, z_n) \rightharpoonup^* \tilde{\sigma} := \langle \nu_x, \sigma(x, \cdot) \rangle \quad \text{in } L^\infty(0,1),$$

where ν_x is the family of Young measures generated by the sequence $(z_n)_n$. Equation (5.2) expresses force balance at $x \in [0,1]$. Observe by (5.3) that (5.2) may even be differentiated:

$$\frac{d}{dx} \langle \nu_x, \sigma(x, \cdot) \rangle = b(u(x), x) \quad \text{in } (0,1).$$

We call the pair (u, ν_x) a *Young-measure solution* of (2.12). Note that we recover classical elasticity in the special case when the Young measure reduces to a Dirac mass, viz., $\nu_x = \delta_{z(x)}$ a.e. In particular, (5.2) then expresses the usual stress-strain relation $\tilde{\sigma}(x) = \sigma(x, z(x))$.

Remark 6. Under more specific assumptions on the sign change of b one can prove that the Young-measure solution is a Dirac mass a.e. (see [15]).

Next we show that the Young measures ν_x reduce to a Dirac mass in the "regions" where the stress function is monotonic. Following the approach in [22], the key ingredient here is the following inequality.

LEMMA 7. *Let $\varphi \in C^1([0,1] \times \mathbb{R}_0^+, \mathbb{R})$ be an arbitrary function with $\varphi_z(x,z) \geq 0$ for every $x \in [0,1]$. Then we deduce the inequality*

$$\int_0^1 \langle \nu_x, \sigma(x,\rho)\varphi(x,\rho) \rangle \, dx \leq \int_0^1 \langle \nu_x, \sigma(x,\rho) \rangle \langle \nu_x, \varphi(x,\rho) \rangle \, dx, \tag{5.4}$$

where ν is the Young measure generated by a sequence $(z_n)_{n \in \mathbb{N}}$ of solutions of (2.12) *for corresponding ε_n, $\varepsilon_n \searrow 0$.*

Proof. Multiplying (2.12) by φ and integrating over the unit interval yield

$$\int_0^1 (-\varepsilon z'' \varphi(x,z) + \sigma(x,z)\varphi(x,z)) \, dx = \int_0^1 \left(\int_x^1 b(u(\eta), \eta) d\eta + \tau \right) \varphi(x,z) dx.$$

Integration by parts, taking the boundary conditions and the assumption on φ into account, yields

$$\begin{aligned} &\int_0^1 \sigma(x,z)\varphi(x,z) dx \\ &= \int_0^1 \left(-\varepsilon z' \varphi_x(x,z) - \varepsilon z'^2 \varphi_z(x,z) + \left(\int_x^1 b(u(\eta),\eta) d\eta + \tau \right) \varphi(x,z) \right) dx \\ &\leq \int_0^1 \left(-\varepsilon z' \varphi_x(x,z) + \left(\int_x^1 b(u(\eta),\eta) d\eta + \tau \right) \varphi(x,z) \right) dx. \end{aligned} \tag{5.5}$$

Consider a sequence $(\varepsilon_n, z_n)_{n \in \mathbb{N}}$ with $\varepsilon_n \searrow 0$ and z_n a solution of (2.12) for corresponding ε_n. Let ν be the Young measure generated by $(z_n)_{n \in \mathbb{N}}$. It is easy to see that

$$\lim_{n \to \infty} \int_0^1 -\varepsilon_n z_n' \varphi_x(x, z_n(x)) dx = 0,$$

and (5.5) implies

$$\int_0^1 \langle \nu_x(\rho), \sigma(x,\rho)\varphi(x,\rho) \rangle \, dx \leq \int_0^1 \left[\left(\int_x^1 b(u(\eta),\eta) d\eta + \tau \right) \langle \nu_x(\rho), \varphi(x,\rho) \rangle \right] dx,$$

which together with Theorem 5 immediately implies (5.4). $\square$

Similar to the treatment in [22], we now consider an estimate which is more convenient for our forthcoming analysis. Therefore, we consider the product measure $\tilde{\nu} := \nu \otimes \nu$ over $\mathbb{R} \times \mathbb{R}$.

LEMMA 8. *Let φ be a test function as in Lemma 7 and let ν be a Young measure generated by a sequence $(z_n)_{n\in\mathbb{N}}$. Then*

$$\int_0^1 \left[\int_{\mathbb{R}^2} \big(\varphi(x,\rho) - \varphi(x,\tau)\big)\big(\sigma(x,\rho) - \sigma(x,\tau)\big) d\big(\nu_x(\rho) \otimes \nu_x(\tau)\big) \right] dx \leq 0. \tag{5.6}$$

Proof. Let $\tilde{\nu} := \nu \otimes \nu$ and let ν be the Young measure generated by a sequence of solutions of (2.12). Furthermore, let $(\rho, \tau) \in \mathbb{R}^2$ and φ be an arbitrary test function as the one considered in Lemma 7. Then we obtain by Fubini's theorem

$$\begin{aligned} &\langle \nu_x, \sigma \rangle \langle \nu_x, \varphi \rangle \\ &= \frac{1}{2} \left[\int_{\mathbb{R}} \sigma(x,\rho) d\nu_x(\rho) \int_{\mathbb{R}} \varphi(x,\tau) d\nu_x(\tau) + \int_{\mathbb{R}} \sigma(x,\tau) d\nu_x(\tau) \int_{\mathbb{R}} \varphi(x,\rho) d\nu_x(\rho) \right] \\ &= \frac{1}{2} \left[\int_{\mathbb{R}^2} (\sigma(x,\rho)\varphi(x,\tau) + \sigma(x,\tau)\varphi(x,\rho))\, d(\tilde{\nu}_x(\rho,\tau)) \right]. \end{aligned} \tag{5.7}$$

On the other hand, due to the fact that ν_x is a probability measure for a.e. $x \in [0,1]$, we have

$$\int_{\mathbb{R}} \sigma(x,\rho)\varphi(x,\rho) d\nu_x(\rho) = \frac{1}{2} \left[\int_{\mathbb{R}^2} (\sigma(x,\rho)\varphi(x,\rho) + \sigma(x,\tau)\varphi(x,\tau))\, d(\tilde{\nu}_x(\rho,\tau)) \right]. \tag{5.8}$$

By (5.7) and (5.8) we conclude that

$$\begin{aligned} &\langle \nu_x, \sigma\varphi \rangle - \langle \nu_x, \sigma \rangle \langle \nu_x, \varphi \rangle \\ &\qquad = \frac{1}{2} \left[\int_{\mathbb{R}^2} (\sigma(x,\rho) - \sigma(x,\tau))\, (\varphi(x,\rho) - \varphi(x,\tau))\, d(\tilde{\nu}_x(\rho,\tau)) \right], \end{aligned}$$

and Lemma 7 gives the result. $\square$

We introduce certain functions

$$\begin{aligned} r_a(x) &:= \min \left\{ r \in \mathbb{R}^+ \mid \sigma(x,r) = \sigma(x, c_2(x)) \right\}, \\ r_b(x) &:= \max \left\{ r \in \mathbb{R}^+ \mid \sigma(x,r) = \sigma(x, c_1(x)) \right\}, \end{aligned}$$

where $c_1(x), c_2(x)$ are defined as in (2.4). Furthermore,

$$\sigma \in C^1([0,1] \times \mathbb{R}^+, \mathbb{R}),$$

which implies that $r_a, r_b \in C^1([0,1])$. Note that $\sigma(x,\rho)$ is injective for $\rho \in (-\infty, r_a(x)) \cup (r_b(x), \infty)$. We define

$$T(x) := \mathbb{R} \backslash [\sigma(x, c_2(x)), \sigma(x, c_1(x))],$$

and, furthermore, for some $u \in C^0(0,1) \cap \{v \mid v(0) = 0\}$, let

$$A(u) := \left\{ x \in [0,1] \mid \int_x^1 b(u(\eta),\eta) d\eta + \tau \in T(x) \right\}. \tag{5.9}$$

Note that $A(u)$ is an open subset of $[0,1]$ by the continuity of $x \mapsto \int_x^1 b(u(\eta),\eta) d\eta$ and by the assumptions on W. Now we can prove the following.

THEOREM 9. *Let* $(z_n)_n$ *be a sequence of solutions of* (2.12) *for corresponding* ε_n, $\varepsilon_n \searrow 0$, *let* z *be the weak limit of the sequence, and let* $u(x) = \int_0^x z(s)ds$. *Furthermore, let* ν *be the Young measure generated by this sequence. Then we obtain for a.e.* $x \in A(u)$

$$\nu_x = \delta_{z(x)}. \tag{5.10}$$

In particular, there exists a subsequence, not relabeled, of $(z_n)_{n\in\mathbb{N}}$ *such that* $z_n(x) \to z(x)$ *for a.e.* $x \in A(u)$.

Remark 10. Note that the last assertion is clear by (5.10) (see [17, p. 118]).

Proof. Let $(z_n)_{n\in\mathbb{N}}$ be a sequence of solutions of (2.12) for corresponding ε_n with $\varepsilon_n \searrow 0$ and let ν be the Young measure generated by this sequence. Let u be the limit of the corresponding sequence $(u_n)_{n\in\mathbb{N}}$.

To deduce the result we define a test function $\varphi \in C^1([0,1] \times \mathbb{R}_0^+, \mathbb{R})$ with the following properties:

(i) $\varphi(x,\cdot)_{|\mathbb{R}_0^+\setminus[r_a(x),r_b(x)]}$ is strictly monotonically increasing for every $x \in [0,1]$,

(ii) $\varphi(x,\cdot)_{|[r_a(x),r_b(x)]} \equiv$ const for every $x \in [0,1]$.

Moreover, we define

$$f(x,\rho,\tau) := (\sigma(x,\rho) - \sigma(x,\tau))\,(\varphi(x,\rho) - \varphi(x,\tau))\,,$$
$$f : [0,1] \times \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R},$$

and by the properties of φ we obtain for every $x \in [0,1]$

$$f(x,\rho,\tau) \begin{cases} > 0 & \text{if } (\varrho,\tau) \notin [r_a(x), r_b(x)]^2 \text{ and } \rho \neq \tau, \\ = 0 & \text{elsewhere.} \end{cases} \tag{5.11}$$

By virtue of (5.6) and (5.11), we deduce

$$\int_{\mathbb{R}^2} f(x,\rho,\tau)\, d(\nu_x(\rho) \otimes \nu_x(\tau)) = 0 \text{ for a.e. } x \in [0,1].$$

Hence, we have

$$\operatorname{supp}(\nu_x \otimes \nu_x) \subseteq \left\{(\rho,\tau) \in \mathbb{R}^2 \mid f(x,\rho,\tau) = 0\right\} \text{ for a.e. } x \in [0,1]. \tag{5.12}$$

We now show that if $\operatorname{supp}\nu_x$ consists of at least two points, then $x \notin A(u)$. Assume $\operatorname{supp}\nu_x \supset \{c(x), d(x)\}$; then

$$\{c(x), d(x)\} \times \{c(x), d(x)\} \in \operatorname{supp}(\nu_x \otimes \nu_x).$$

Because of (5.11) and (5.12) we deduce $\{c(x), d(x)\} \in [r_a(x), r_b(x)]$. Thus we obtain

$$\operatorname{supp}\nu_x \subseteq [r_a(x), r_b(x)], \tag{5.13}$$

which, in view of Theorem 5, implies

$$\int_x^1 b(u(\eta),\eta)d\eta + \tau = \langle \nu_x, \sigma\rangle \in [\sigma(x, c_2(x)), \sigma(x, c_1(x))].$$

We conclude immediately that $x \notin A(u)$. This yields the desired result. ∎

In the same way one can prove the following corollary just using $z(x) = \langle \nu_x, Id \rangle$ for a.e. $x \in [0,1]$, where z is the weak limit of $(z_n)_{n\in\mathbb{N}}$ and ν is the Young measure generated by $(z_n)_{n\in\mathbb{N}}$.

COROLLARY 11. *Let $(z_n)_{n\in\mathbb{N}}$ and ν be as in Theorem 9, and let z be the weak limit of $(z_n)_{n\in\mathbb{N}}$. Then for a.e. $x \in [0,1]$ with $z(x) \in \mathbb{R}^+ \setminus [r_a(x), r_b(x)]$ we obtain $\nu_x = \delta_{z(x)}$. Moreover, we have the following: If $\operatorname{supp} \nu_x \subseteq (-\infty, r_a(x)) \cup (r_b(x), \infty)$ is valid, then we also get $\nu_x = \delta_{z(x)}$.*

Another immediate consequence of (5.13) is the following corollary.

COROLLARY 12. *Let the assumptions be as in Corollary 11 and define $u(x) := \int_0^x z(s)ds$. Then for a.e. $x \in [0,1] \setminus A(u)$ we have*

$$\operatorname{supp} \nu_x \subseteq [r_a(x), r_b(x)].$$

Proof. We assume some $x \in [0,1] \setminus A(u)$ such that

$$c \notin [r_a(x), r_b(x)] \text{ and } \{c\} \in \operatorname{supp} \nu_x. \tag{5.14}$$

Without loss of generality we assume $c < r_a(x)$, and hence

$$\sigma(x,c) < \int_x^1 b(u(\eta),\eta)d\eta + \tau = \langle \nu_x, \sigma \rangle \in [\sigma(x, c_2(x)), \sigma(x, c_1(x))].$$

In particular, $\operatorname{supp} \nu_x$ consists of at least two points $\{c, d\}$ with $d > c$. By the same procedure as in Theorem 9, we obtain (see (5.13))

$$\{c, d\} \subseteq [r_a(x), r_b(x)],$$

contradicting (5.14). □

Remark 13. By Theorem 9 and Corollary 12 we conclude the following: Let $(z_n)_{n\in\mathbb{N}}$ be a sequence of solutions of (2.12) for corresponding ε_n, $\varepsilon_n \searrow 0$. By virtue of Theorem 1 there exists $\delta > 0$ with $z_n(x) > \delta > 0$ for every $n \in \mathbb{N}$ and $x \in [0,1]$. Let ν be the Young measure generated by $(z_n)_{n\in\mathbb{N}}$ and let $\{c\} \in \operatorname{supp} \nu_x$. Then $c \geq \delta$.

6. Concluding remarks. (i) The Young measure $\nu = (\nu_x)_{x\in\Omega}$ generated by a sequence $(v_n)_{n\in\mathbb{N}}$, where $v_n : \Omega \mapsto \mathbb{R}^N$, Ω denotes some open set, has the following interpretation: The measure ν_{x_0} can be thought of as the limiting probability of values of $(v_n)_{n\in\mathbb{N}}$ in a small neighborhood of $x_0 \in \Omega$. In order to be mathematically more precise we have the following: First we define

$$\mu_{x_0,\delta,v_n}(C) := \frac{|\{x \in B_\delta(x_0) \mid v_n(x) \in C\}|}{|B_\delta(x_0)|},$$

where $B_\delta(x_0)$ is an open ball with center x_0 and radius δ, C denotes some measurable set in $\mathbb{R}^N$, and $|\cdot|$ denotes the Lebesgue measure. Then we have for every $f \in C_0(\mathbb{R}^N)$

$$\lim_{\delta \searrow 0} \lim_{n\to\infty} \langle \mu_{x_0,\delta,v_n}, f \rangle = \langle \nu_{x_0}, f \rangle .$$

Hence we obtain the following by Theorem 9: Let $u(x) := \int_0^x z(s)ds$, where z is the weak limit of the generating sequence $(z_n)_{n\in\mathbb{N}}$. If $\int_x^1 b(u(\eta),\eta)d\eta + \tau \in A(u)$, where $A(u)$ is defined by (5.9), then the limit of the generating sequence $(z_n)_{n\in\mathbb{N}}$ is determined, because the resulting Young measure ν is a Dirac measure for a.e. $x \in A(u)$. In particular, we have by the Vitalis theorem

$$\lim_{n\to\infty} z_{n|A(u)} = z_{|A(u)} \quad \text{in } L^1(A).$$

On the other hand, if $x \notin A(u)$, then ν_x is not necessarily a Dirac measure, indicating the possibility of strongly oscillatory behavior of the sequence $(z_n)_{n\in\mathbb{N}}$ in this region. However, by virtue of Corollary 12, the values attained by $(z_n)_{n\in\mathbb{N}}$ in $[0,1] \setminus A(u)$ are in the interval $[r_a(x) - \theta, r_b(x) + \theta]$ for arbitrary $\theta > 0$ and for $n = n(\theta)$ sufficiently large. Thus the amplitude of the possible oscillation is restricted by these two values.

(ii) Note that

$$\begin{aligned} &\int_0^1 (\langle \nu_x, W\rangle + B(u,x))\,dx - \tau u(1),\\ &u(0) = 0, \quad u'(x) = \langle \nu_x, Id\rangle\,,\\ &\int_0^1 \int_{\mathbb{R}} ||s||^{p+1}\,d\nu_x(s)dx < \infty \end{aligned} \tag{6.1}$$

possesses a global minimizer in the space of Young measures (see [18]). Since we do not know if $\int_0^1 \varepsilon(z_\varepsilon')^2\,dx \searrow 0$ as $\varepsilon \searrow 0$, it is unclear if the Young-measure solution ν obtained here is the global minimizer of (6.1)—even in the case when the family of critical points corresponds to the global minimizers of (2.9).

(iii) Problem (2.10) with Dirichlet boundary conditions seems to be more difficult than problem (2.10), (2.11). Indeed the latter is statically determinate, a by-product of which is the reformulation (2.12). In particular, it is unclear if Theorem 1 and Lemma 2 hold for the Dirichlet problem.

(iv) In higher-dimensional problems, we consider a nonconvex $W : \Omega \times \mathbb{R}^{N\times M} \to \mathbb{R}$ ($\Omega \subseteq \mathbb{R}^N$ is some bounded smooth domain) with the usual polynomial growth conditions and some subcritical $B : \Omega \times \mathbb{R}^M \to \mathbb{R}$; then by similar techniques to those presented here, we can prove the existence of Young-measure solutions of the corresponding nonelliptic Euler–Lagrange system of the variational problem

$$\begin{aligned} &\int_\Omega W(x, Du(x)) + B(u(x), x)\,dx,\\ &u_{|\Gamma} = 0, \end{aligned}$$

where $\Gamma \subseteq \partial\Omega$ and the $(n-1)$-dimensional Hausdorff measure of Γ is positive. In addition, if we require the analogue of (2.1), namely, $W = W(x, F, \det F)$, where $F \in \mathbb{R}^{N\times N}$ with

$$\lim_{\det F \searrow 0} W(x, F, \det F) = \infty$$

for a.e. $x \in \Omega$, then proving a version of Theorem 1 is an open problem.

REFERENCES

[1] S.S. Antman and H. Brezis, *The existence of orientation–preserving deformations in nonlinear elasticity*, in Nonlinear Analysis and Mechanics: Heriot–Watt Symposium, Volume 2, Pitman, Boston, MA, 1978, pp. 1–29.

[2] J.M. Ball, *A version of the fundamental theorem for Young measures*, in Partial Differential Equations and Continuum Models of Phase Transitions, M. Rascle, D. Serre, and M. Slemrod, eds., Lecture Notes in Phys. 359, Springer-Verlag, New York, 1989, pp. 207–215.

[3] J.M. Ball and R.D. James, *Fine phase mixtures as minimizers of energy*, Arch. Ration. Mech. Anal., 100 (1987), pp. 13–52.

[4] K. Bhattacharya, *Microstructure of Martensite*, Oxford University Press, Oxford, UK, 2003.

[5] J. Carr, M.E. Gurtin, and M. Slemrod, *One–dimensional structured phase transitions under prescribed loads*, J. Elasticity, 15 (1985), pp. 133–142.

[6] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, New York, 1989.
[7] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, New York, 2000.
[8] R.J. DiPerna, *Measure–valued solutions to conservation laws*, Arch. Ration. Mech. Anal., 88 (1985), pp. 223–270.
[9] G. Dolzmann, N. Hungerbühler, and S. Müller, *Non–linear elliptic systems with measure–valued right hand side*, Math. Z., 226 (1997), pp. 545–574.
[10] J.L. Ericksen, *Equilibrium of bars*, J. Elasticity, 5 (1975), pp. 191–201.
[11] T.J. Healey and H. Kielhöfer, *Global continuation via higher-gradient regularization and singular limits in forced one-dimensional phase transitions*, SIAM J. Math. Anal., 31 (2000), pp. 1307–1331.
[12] N. Hungerbühler, *Quasilinear elliptic systems in divergence form with weak monotonicity*, New York J. Math., 5 (1999), pp. 83–90.
[13] R.D. James, *Co–existent phases in the one-dimensional static theory of elastic bars*, Arch. Ration. Mech. Anal., 72 (1979/80), pp. 99–140.
[14] S. Krömer and M. Lilli., *Branches of positive solutions of quasilinear elliptic equations on non–smooth domains*, Nonlinear Anal., 64 (2006), pp. 2183–2202.
[15] M. Lilli, *The Effect of a Singular Perturbation to a 1-d Nonconvex Variational Problem*, Augsburger Schriften zur Mathematik, Physik und Informatik, Logos–Verlag, Berlin, 2005.
[16] J. Málek, J. Neăs, M. Rokyta, and M. Ružič ka, *Weak and Measure Valued Solutions to Evolutionary PDE*, Chapman–Hall, London, 1996.
[17] S. Müller, *Variations models for microstructure and phase transition*, in Calculus of Variations and Geometric Evolution Problems (Cetraro, 1996), Lecture Notes in Math. 1713, Springer-Verlag, New York, 1999, pp. 85–210.
[18] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser, Basel, 1997.
[19] M.O. Rieger, *Young-measure solutions for an elasticity equation with diffusion*, in EQUADIFF 99, International Conference on Differential Equations, World Scientific, River Edge, NJ, 2000, pp. 457–459.
[20] L. Tartar, *The compensated compactness method applied to systems of conservation laws*, in Systems of Nonlinear Partial Differential Equations, J.M. Ball, ed., NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 111, Reidel, Dordrecht, The Netherlands, 1983, pp. 263–285.
[21] F. Theil, *Young-measure solutions for a viscoelastic damped wave equation with nonmonotone stress–strain relation*, Arch. Ration. Mech. Anal., 144 (1998), pp. 47–78.
[22] I. Vecchi, *A note on entropy compactification for scalar conservation laws*, Nonlinear Anal., 15 (1990), pp. 693–695.
[23] L.C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, 2nd ed., Chelsea Publishing Company, New York, 1980.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 210–235

BIFURCATION ANALYSIS OF AN ELLIPTIC FREE BOUNDARY PROBLEM MODELLING THE GROWTH OF AVASCULAR TUMORS*

SHANGBIN CUI† AND JOACHIM ESCHER‡

Abstract. We study bifurcations from radially symmetric solutions of a free boundary problem modelling the dormant state of nonnecrotic avascular tumors. This problem consists of two semilinear elliptic equations with a Dirichlet and a Neumann boundary condition, respectively, and a third boundary condition coupling surface tension effects on the free interface to the internal pressure. By reducing the full problem to an abstract bifurcation equation in terms of the free boundary only and by characterizing the linearization as a Fourier multiplication operator, we carry out a precise analysis of local bifurcations of this problem.

Key words. free boundary problem, elliptic equations, bifurcation, tumor growth, surface tension

AMS subject classifications. 35K35, 35Q80, 35R05

DOI. 10.1137/060657509

1. Introduction.

In this paper we consider the following free boundary problem:

$$
\left\{
\begin{array}{rcll}
\Delta\sigma & = & f(\sigma) & \text{in} \quad \Omega, \\
\Delta p & = & -g(\sigma) & \text{in} \quad \Omega, \\
\sigma & = & \bar{\sigma} & \text{on} \quad \partial\Omega, \\
\frac{\partial p}{\partial\nu} & = & 0 & \text{on} \quad \partial\Omega, \\
p & = & \gamma\kappa & \text{on} \quad \partial\Omega,
\end{array}
\right.
\tag{1.1}
$$

where σ, p are unknown functions defined on an a priori unknown domain Ω. Furthermore, f and g are given functions defined on $[0,\infty)$, satisfying suitable conditions $(A1)$–$(A3)$ on the following page, $\bar{\sigma}$ is a positive constant, $\partial/\partial\nu$ denotes the derivative in the direction of outward normal ν of the free boundary $\partial\Omega$, γ is a positive parameter, and κ is the mean curvature of the free boundary $\partial\Omega$. The manner of determining the sign of κ is such that $\kappa \geq 0$ at points where $\partial\Omega$ is (locally) convex.

The above problem is a mathematical model describing the dormant state of a nonnecrotic avascular tumor. In this model Ω is the domain occupied by the tumor, σ denotes the concentration of nutrient within the tumor, p is the tumor's tissue pressure, and f and g are the nutrient consumption rate function and the tumor-cell proliferation rate function, respectively. It is assumed that all tumor cells are alive and dividable and that their density is constant. This explains why f and g are independent of the cell density. Typical examples of f and g are as follows:

$$
f(\sigma) = \sigma, \quad g(\sigma) = \mu(\sigma - \tilde{\sigma}),
\tag{1.2}
$$

*Received by the editors April 18, 2006; accepted for publication (in revised form) November 27, 2006; published electronically May 22, 2007.

http://www.siam.org/journals/sima/39-1/65750.html

†Institute of Mathematics, Sun Yat-Sen University, Guangzhou, Guangdong 510275, People's Republic of China (cuisb3@yahoo.com.cn).

‡Institute for Applied Mathematics, Leibniz University Hanover, Welfengarten 1, 30167 Hanover, Germany (escher@ifam.uni-hannover.de). This author was supported by the National Natural Science Foundation of China under grant 10471157.

where μ and $\tilde{\sigma}$ are positive constants (see Byrne and Chaplain [2]) and

$$g(\sigma) = \mu(\sigma) - \nu(\sigma) \equiv \frac{A\sigma^{m_1}}{\sigma_c^{m_1} + \sigma^{m_1}} - B\left(1 - \frac{\delta\sigma^{m_2}}{\sigma_d^{m_2} + \sigma^{m_2}}\right),$$
$$f(\sigma) = \beta\mu(\sigma) + h(\sigma), \tag{1.3}$$

where A, B, β, δ, m_1, m_2, σ_c, and σ_d are positive constants, and $h(u)$ is a nonnegative increasing function (see Ward and King [12]). In this paper we shall assume that f, g are general functions satisfying the following condition (A):

(A_1) $\quad f \in C^\infty[0,\infty)$ and $g \in C^\infty[0,\infty)$,

(A_2) $\quad f'(\sigma) > 0$ for $\sigma \geq 0$ and $f(0) = 0$,

(A_3) $\quad g'(\sigma) > 0$ for $\sigma \geq 0$ and $g(\tilde{\sigma}) = 0$ for some $\tilde{\sigma} > 0$,

(A_4) $\quad \tilde{\sigma} < \bar{\sigma}$.

Note that the positive value $\tilde{\sigma}$ in (A_3) plays the role of a threshold in the sense that $g(\sigma) < 0$ for $0 \leq \sigma < \tilde{\sigma}$ and $g(\sigma) > 0$ for $\sigma > \tilde{\sigma}$. Moreover, using the maximum principle and Green's formula, one can easily prove that (1.1) does not have any solution if $\tilde{\sigma} \geq \bar{\sigma}$. Thus it is reasonable to assume that (A_4) holds true.

A basic result is that under the above assumptions (A), the problem (1.1) has a radially symmetric solution, which is unique up to translation and rotation (see section 2). Moreover, it can be shown that this solution is globally asymptotically stable under radially symmetric perturbations, namely, if we consider the corresponding evolutionary problem and regard the solution of (1.1) as a stationary solution of the corresponding evolutionary problem, then all radially symmetric transient solutions converge to this radially symmetric stationary solution as time goes to infinity. A natural and very interesting question is whether the problem (1.1) admits nonradial solutions. This question has the following obvious implication: If (1.1) has a nonradial solution, then we cannot expect that the radially symmetric solution is globally asymptotically stable under nonradially symmetric perturbations.

In the particular case that f and g have the simple form given in (1.2), this question has been answered by Friedman and Reitich [8] in the two-dimensional case, and by Fontelos and Friedman [7] in the three-dimensional case. They considered (1.1) as a bifurcation problem with γ being the bifurcation parameter and used the power series method suggested by the work of Greenspan [11] and Byrne [1] to prove that there exists a sequence of positive numbers γ_k, such that in a neighborhood of each γ_k there exists a bifurcation branch $(\sigma_\varepsilon, p_\varepsilon, \Omega_\varepsilon, \gamma_\varepsilon)$, with ε being a small parameter (bifurcation parameter) of nonradial solutions which bifurcates from the radially symmetric solution at γ_k. This power series method has also been used by Friedman and Reitich [9] to consider the symmetry-breaking bifurcation of some other related free boundary problems. Recently, Borisovich and Friedman [3] used a different approach to reconsider all of those free boundary problems. By this new approach, these problems were reduced to bifurcation problems treatable by the standard bifurcation theory developed, e.g., by Crandall and Rabinowitz [4], and Krasnoselskii.

In this paper we study the problem (1.1) for general nonlinear functions f and g satisfying the condition (A) and for star shaped domains Ω. As in [3, 4, 5, 6, 7, 8, 9], we shall also regard (1.1) as a bifurcation problem with γ being the bifurcation parameter. However, we shall not use the power series method, because in the general situation f, g are not necessarily analytic functions, and therefore this approach is certainly not applicable. Instead, we shall solve the first four equations in terms of an unknown function η describing the free boundary and substitute the solution into the last equation, reducing the problem (1.1) in that way into a bifurcation problem of the following form:

$$A(\eta) + c = \gamma B(\eta), \tag{1.4}$$

where η, c are the unknowns and A, B are nonlinear operators mapping a hypersurface $\mathcal{M}$ of a suitable Banach space X into another Banach space Y. In our application it turns out that X is compactly embedded into Y and that (A, B) is smooth. The key step of our approach is to show that for each eigenvalue γ_k of the linearized problem of (1.4), there exist corresponding closed subspaces X_k of X and Y_k of Y, such that X_k is compactly embedded into Y_k, $A : X_k \cap \mathcal{M} \to X_k$ and $B : X_k \cap \mathcal{M} \to Y_k$, and when restricted to X_k, the linearized problem possesses γ_k as a simple eigenvalue so that the well-known Crandall–Rabinowitz theorem is applicable.

The main result of this paper is the following theorem.

THEOREM 1.1. *Assume that condition* (A) *holds true. Then there exist a null sequence of numbers* $\{\gamma_k\}_{k\geq 2}$ *and an integer* $k^* \geq 2$ *such that in a neighborhood of each* γ_k *with* $k \geq k^*$*, the problem* (1.1) *has a branch of nonradial solutions bifurcating from the radially symmetric solution. If* $\gamma \neq \gamma_k$ *for all* $k \geq 2$*, then* γ *is not a bifurcation point.*

It is worthwhile to note that our proof of Theorem 1.1 yields an asymptotic expansion of the bifurcation branch with respect to the bifurcation parameter; cf. section 5. This expansion particularly shows that the domains corresponding to the above bifurcation solutions are always star shaped perturbations of round spheres.

We also remark that the proof of Theorem 1.1 shows that all mutually distinct γ_k are bifurcation points and that k^* is the smallest number such that all γ_k with $k \geq k^*$ are distinct. It is known that in the linear case (1.2) this threshold value k^* is minimal, i.e., $k^* = 2$ and all γ_k are bifurcation points; cf. [7]. For general nonlinearities f and g this seems to be in unknown. However, it can be shown that if f and g are real analytic functions, then, excluding at most a discrete set of boundary concentrations $\overline{\sigma}$, all γ_k are distinct and thus bifurcation points. We are indebted to one of the referees for bringing this point to our attention.

The structure of this paper is as follows. In section 2 we study the existence and uniqueness of radially symmetric solutions to problem (1.1). The linearization of (1.1) at radially symmetric solutions is determined in section 3, where we also calculate the eigenvalues of the linearized problem. In section 4 we reduce the problem (1.1) into the abstract form (1.4), and we show that if $\gamma \neq \gamma_k$, then no bifurcation occurs at γ. In the last section we prove the local bifurcation result.

2. Radially symmetric solutions. In this section we study the existence and uniqueness of radially symmetric solutions to the system (1.1). Obviously, this problem is equivalent to the study of the following free boundary value problem:

$$\left\{ \begin{array}{rcll} \sigma_0''(r) + \frac{n-1}{r}\sigma_0'(r) & = & f(\sigma_0(r)), & 0 < r < R, \\ p_0''(r) + \frac{n-1}{r}p_0'(r) & = & -g(\sigma_0(r)), & 0 < r < R, \\ \sigma_0, p_0 \in C^1[0,R] & \cap & C^2(0,R), & \\ \sigma_0(R) & = & \bar{\sigma}, & \\ p_0'(R) & = & 0, & \\ p_0(R) & = & \frac{\gamma}{R}. & \end{array} \right. \tag{2.1}$$

That is, if (σ_0, p_0, R) is a solution of (2.1) then, letting $\Omega_0 = \{x \in \mathbb{R}^n : |x| < R\}$, the triple $(\sigma_0(|x|), p_0(|x|), \Omega_0)$ is a radially symmetric solution of (1.1) and vice versa.

Given any $c > 0$, it can be shown by ODE techniques that the first equation in (2.1) has a unique local solution $\sigma_0 \in C^\infty[0, \delta]$ such that $\sigma_0(0) = c$. We extend this local solution to a solution which is not further extendable, and let $[0, R^*)$ be the maximal interval of existence, where $0 < R^* \leq \infty$. We claim that

$$\sigma_0'(r) > 0 \quad \text{for } \; 0 < r < R^*. \tag{2.2}$$

Indeed, it is not difficult to see that $\sigma_0'(0) = 0$. Since $\sigma_0(0) = c > 0$, we infer from the first equation in (2.1) that $\sigma_0''(0) = f(c)/n > 0$. Hence there is a $\delta > 0$ such that $\sigma_0'(r) > 0$ for $0 < r < \delta$. If $\sigma_0'(r) > 0$ does not hold for some $0 < r < R^*$, then for the first positive zero r_0 of $\sigma_0'(r)$ we have $\sigma_0''(r_0) \leq 0$. On the other hand, the first equation in (2.1) shows that $\sigma_0''(r_0) = f(\sigma_0(r_0)) > f(c) > 0$, which is a contradiction. Hence (2.2) holds.

From the just-established relation (2.2), it is clear that if $R^* < \infty$, then

$$\lim_{r \to R^*} \sigma_0(r) = \infty. \tag{2.3}$$

We claim that even if $R^* = \infty$, then (2.3) still holds true. Indeed, by (2.2) we know that $\sigma_0(r) \geq \sigma_0(0) = c$ for all $0 < r < R^* = \infty$. Further, the first equation of (2.1) shows that

$$\sigma_0'(r) = \frac{1}{r^{n-1}} \int_0^r f(\sigma_0(\rho))\rho^{n-1} d\rho \geq \frac{1}{r^{n-1}} \int_0^r f(c)\rho^{n-1} d\rho = \frac{1}{n} f(c) r, \tag{2.4}$$

which implies $\lim_{r \to \infty} \sigma_0(r) = \infty$.

Given $c \in (0, \bar{\sigma})$, it follows from (2.2) and (2.3) that there exists a unique $R_c > 0$ such that

$$\sigma_0(R_c) = \bar{\sigma}.$$

Since $\sigma_0(r)$ is clearly increasing in c and smooth we conclude that R_c is a smooth and decreasing function of c on $0 < c < \bar{\sigma}$. Moreover, it is also clear that

$$\lim_{c \to 0^+} R_c = \infty, \qquad \lim_{c \to \bar{\sigma}^-} R_c = 0.$$

Hence, given $R > 0$, there exists a unique $0 < c < \bar{\sigma}$ such that $R_c = R$. Thus, we have shown that for any given $R > 0$, the problem

$$\sigma_0''(r) + \frac{n-1}{r} \sigma_0'(r) = f(\sigma_0(r)), \qquad \sigma_0(R) = \bar{\sigma}, \tag{2.5}$$

has a unique solution, which we denote by $\sigma_0(r, R)$.

In the following, we shall derive a further representation of $\sigma_0(r, R)$. To this end we consider the following parameter-dependent problem:

$$\begin{cases} \dfrac{\partial^2 U(r,\lambda)}{\partial r^2} + \dfrac{n-1}{r} \dfrac{\partial U(r,\lambda)}{\partial r} = \lambda f(U(r,\lambda)), \quad 0 < r < 1, \\ \dfrac{\partial U}{\partial r}\Big|_{r=0} = 0, \quad U\Big|_{r=1} = \bar{\sigma}, \end{cases} \tag{2.6}$$

where λ is a positive parameter. By a similar argument, as used previously, we see that for every $\lambda > 0$ this problem has a unique solution $U(r, \lambda)$. Clearly, $U(r, \lambda)$ is smooth in both variables r and λ. We have

$$\frac{\partial U(r,\lambda)}{\partial r} > 0, \quad \frac{\partial U(r,\lambda)}{\partial \lambda} < 0 \quad \text{for } 0 < r < 1, \ \ \lambda > 0, \tag{2.7}$$

and

$$\lim_{\lambda \to 0^+} U(r,\lambda) = \bar{\sigma}, \quad \lim_{\lambda \to \infty} U(r,\lambda) = 0 \quad \text{for } 0 \le r < 1. \tag{2.8}$$

Indeed, the first inequality in (2.7) is obvious. The second inequality follows from the maximum principle applied to the equation satisfied by $\partial U(r, \lambda)/\partial\lambda$. Concerning (2.8), the first relation is also obvious, because if $\lambda = 0$ in (2.6), then we get the trivial solution $U = \bar{\sigma}$. To prove the second relation, we apply Lebesgue's dominated convergence theorem to the integral representation of $U(r, \lambda)$ in terms of $f(U(r, \lambda))$ to get $\lim_{\lambda\to\infty} U(r, \lambda) = 0$; cf. also [5], where a similar argument is used. By uniqueness of the solution of the problem (2.5), one can easily verify that

$$\sigma_0(r,R) = U\Big(\frac{r}{R}, R^2\Big) \quad \text{for } 0 \le r \le R. \tag{2.9}$$

Next let us consider the second equation in (2.1), where $\sigma_0(r) = \sigma_0(r, R)$. Clearly this equation has a unique local solution $\bar{p}_0(r)$ satisfying the following initial conditions:

$$\bar{p}_0(0) = 0, \quad \bar{p}_0'(0) = 0;$$

cf. [6]. Since the ODE for $\bar{p}$ is linear and regular on $r > 0$, its solution $\bar{p}_0(r)$ can be extended to the whole interval $[0, R^*)$. Moreover, we have

$$\bar{p}_0'(R) = -\frac{1}{R^{n-1}} \int_0^R g(\sigma_0(r,R)) r^{n-1} dr = -R \int_0^1 g(U(\rho,R^2))\rho^{n-1} d\rho.$$

Consider next the auxiliary function

$$\psi(R) = \int_0^1 g(U(\rho,R^2))\rho^{n-1} d\rho, \quad R > 0.$$

Since $g' > 0$ and $\partial U(r, \lambda)/\partial\lambda < 0$, it is clear that ψ is strictly decreasing. Recalling (2.8) and (A_3), (A_4), we have

$$\lim_{R\to 0^+} \psi(R) = \frac{1}{n} g(\bar{\sigma}) > 0, \qquad \lim_{R\to\infty} \psi(R) = \frac{1}{n} g(0) < 0.$$

Hence, there exists a unique $R > 0$ such that $\psi(R) = 0$, showing that $\bar{p}_0'(R) = 0$. Summarizing, we have constructed a unique solution to the first five equations of (2.1) under the additional condition $\bar{p}_0(0) = 0$. The solution of the full system (2.1) is now obtained by replacing $\bar{p}_0(r)$ with

$$p(r) = \frac{\gamma}{R} + \bar{p}_0(r) - \bar{p}_0(R),$$

which is clearly the unique solution of this problem.

For later purposes we shall derive the following property of $p(r)$:

$$p'(r) > 0 \quad \text{for } 0 < r < R, \qquad p'(r) < 0 \quad \text{for } R < r < R^*. \tag{2.10}$$

The proof is simple: We have

$$\begin{aligned}\frac{d}{dr}\Big(\frac{p'(r)}{r}\Big) &= -\frac{d}{dr}\Big(\frac{1}{r^n}\int_0^r g(\sigma_0(\rho))\rho^{n-1}d\rho\Big) = -\frac{d}{dr}\int_0^1 g\Big(U\Big(\frac{r\rho}{R},R^2\Big)\Big)\rho^{n-1}d\rho\\ &= -\frac{1}{R}\int_0^1 g'\Big(U\Big(\frac{r\rho}{R},R^2\Big)\Big)\frac{\partial U}{\partial r}\Big(\frac{r\rho}{R},R^2\Big)\rho^n d\rho < 0.\end{aligned}$$

Both the relation and the fact that $p'(R)=0$ obviously implies (2.10).

We summarize the above considerations in the following theorem.

THEOREM 2.1. *Under the assumption* (*A*)*, the problem* (1.1) *has a unique radially symmetric solution* $(\sigma_0(r),p_0(r),\Omega_0)$*, where* $\Omega_0=\{x\in\mathbb{R}^n:\ r=|x|<R\}$*. Moreover, the functions* $\sigma_0(r)$ *and* $p_0(r)$ *can be extended to a maximal interval* $[0,R^*)$*, where* $0<R^*\le\infty$*, such that* $\sigma_0(r)$, $p_0(r)\in C^\infty[0,R^*)$*. In addition, the properties* (2.2)*,* (2.3)*, and* (2.10) *hold true.*

As pointed out in section 1, the radially symmetric solution is globally asymptotically stable under radially symmetric perturbations. This means that if we consider the corresponding time-dependent problem, then all radially symmetric transient solutions converge to the solution of (2.1) obtained above. A proof of this fact in the three-dimensional case is given in [5]; the proof for higher dimensional problems is similar.

3. Linearization. In this section we determine the linearization of the problem (1.1) at the unique radially symmetric solution constructed in the previous section. We also provide conditions which imply that the linearized problem has nontrivial solutions.

Consider perturbations of the radially symmetric solution $(\sigma_0(r),p_0(r),\Omega_0)$ of the form

$$\sigma(x)=\sigma_0(r)+\varepsilon\varphi(r,\omega),\quad p(x)=p_0(r)+\varepsilon\psi(r,\omega),\quad \Omega=\{x\in\mathbb{R}^n:\ r<R+\varepsilon\eta(\omega)\},$$

where $r=|x|$ and $\omega=x/|x|$. Moreover, ε is a small parameter and φ, ψ, η are new unknown functions. Let Δ_ω be the Laplace–Beltrami operator on the sphere $\mathbb{S}^{n-1}$. Using the corresponding equations of (2.1), one can easily verify that the linearization of the first two equations of (1.1) is given by

$$\frac{\partial^2\varphi}{\partial r^2}+\frac{n-1}{r}\frac{\partial\varphi}{\partial r}+\frac{1}{r^2}\Delta_\omega\varphi=f'(\sigma_0(r))\varphi, \tag{3.1}$$

$$\frac{\partial^2\psi}{\partial r^2}+\frac{n-1}{r}\frac{\partial\psi}{\partial r}+\frac{1}{r^2}\Delta_\omega\psi=-g'(\sigma_0(r))\varphi. \tag{3.2}$$

Besides, it is also immediate to see that the linearization of the first boundary condition in (1.1) is given by

$$\varphi(R,\omega)+\sigma_0'(R)\eta(\omega)=0. \tag{3.3}$$

To find the linearization of the second boundary condition of (1.1), we first compute the outward normal of $\partial\Omega$:

$$\nu=\frac{\omega-\varepsilon\nabla_\omega\eta(\omega)}{|\omega-\varepsilon\nabla_\omega\eta(\omega)|}=\omega-\varepsilon\nabla_\omega\eta(\omega)+o(\varepsilon),$$

where ∇_ω denotes the tangent mapping on the sphere $\mathbb{S}^{n-1}$, or the orthogonal projection of the gradient ∇ to the tangent space $T_\omega(\mathbb{S}^{n-1})$, when regarding a function in $\omega \in \mathbb{S}^{n-1}$ as a function in $x \in \mathbb{R}^n$ with $\omega = x/|x|$, so that $\omega \cdot \nabla_\omega = 0$. Thus

$$\begin{aligned}\frac{\partial p}{\partial \nu} &= \nu \cdot \nabla p|_{r=R+\varepsilon\eta} = [\omega - \varepsilon \nabla_\omega \eta(\omega) + o(\varepsilon)] \cdot \Big(\frac{\partial p}{\partial r}\omega + \nabla_\omega p\Big)\Big|_{r=R+\varepsilon\eta} \\ &= \frac{\partial p}{\partial r}(R,\omega) + \varepsilon\Big(\frac{\partial^2 p}{\partial r^2}(R,\omega)\eta(\omega) - \nabla_\omega \eta(\omega)\cdot\nabla_\omega p(R,\omega)\Big) + o(\varepsilon) \\ &= p_0'(R) + \varepsilon\Big(\frac{\partial \psi}{\partial r}(R,\omega) + p_0''(R)\eta(\omega)\Big) + o(\varepsilon).\end{aligned}$$

Since $p_0'(R) = 0$ and $p_0''(R) = -g(\bar\sigma)$, we see that the desired linearization is given by

$$\frac{\partial \psi}{\partial r}(R,\omega) - g(\bar\sigma)\eta(\omega) = 0. \tag{3.4}$$

Finally, we have the following formula for the mean curvature of the hypersurface given by $r = R + \varepsilon\eta(\omega)$ (see [9]):

$$\kappa = \frac{1}{R} - \frac{\varepsilon}{R^2}\left[\eta(\omega) + \frac{1}{n-1}\Delta_\omega \eta(\omega)\right] + o(\varepsilon).$$

Therefore the linearization of the last equation in (1.1) reads as

$$\psi(R,\omega) + \frac{\gamma}{R^2}\left[\eta(\omega) + \frac{1}{n-1}\Delta_\omega\eta(\omega)\right] = 0. \tag{3.5}$$

Summarizing, we have the following lemma.

LEMMA 3.1. *The linearization of the problem* (1.1) *at the radially symmetric solution* $(\sigma_0(|x|), p_0(|x|), \Omega_0)$ *is given by the problem* (3.1)–(3.5).

We now investigate the question of whether there exists $\gamma > 0$ such that the problem (3.1)–(3.5) has nontrivial solutions. For this purpose we first note that standard results for second order elliptic partial differential equations imply that all solutions φ, ψ, η are smooth, namely, $\varphi, \psi \in C^\infty(\overline{B}_R) \subseteq C^\infty([0,R], C^\infty(\mathbb{S}^{n-1}))$, and $\eta \in C^\infty(\mathbb{S}^{n-1})$. Thus these functions can be expanded in the following way:

$$\begin{cases} \varphi(r,\omega) = \displaystyle\sum_{k=1}^{\infty}\sum_{l=1}^{d_k} u_{kl}(r) Y_{kl}(\omega), \\ \psi(r,\omega) = \displaystyle\sum_{k=1}^{\infty}\sum_{l=1}^{d_k} v_{kl}(r) Y_{kl}(\omega), \\ \eta(\omega) = \displaystyle\sum_{k=1}^{\infty}\sum_{l=1}^{d_k} c_{kl} Y_{kl}(\omega), \end{cases} \tag{3.6}$$

where we denote by $Y_{kl}(\omega)$ $(l = 1, 2, \ldots, d_k)$ the normalized orthogonal basis of all spherical harmonics of degree k. Note that

$$d_1 = n, \quad d_k = \binom{n+k-1}{k} - \binom{n+k-3}{k-2} \ (k \geq 2).$$

The coefficients $u_{kl}(r)$, $v_{kl}(r)$, c_{kl} satisfy the following estimates: Given positive integers N, k, and $1 \leq l \leq d_k$, we have

$$|u_{kl}(r)| \leq C_N(1+k)^{-N}, \quad |v_{kl}(r)| \leq C_N(1+k)^{-N}, \quad |c_{kl}| \leq C_N(1+k)^{-N}.$$

Substituting (3.6) into (3.1)–(3.5), using the relation

$$\Delta_\omega Y_{kl}(\omega) = -\lambda_k Y_{kl}(\omega), \quad \lambda_k = k^2 + (n-2)k \ \ (k = 1, 2, \dots),$$

and comparing coefficients of every $Y_{kl}(\omega)$, we get

$$u''_{kl}(r) + \frac{n-1}{r} u'_{kl}(r) - \frac{\lambda_k}{r^2} u_{kl}(r) = f'(\sigma_0(r)) u_{kl}(r), \tag{3.7}$$

$$v''_{kl}(r) + \frac{n-1}{r} v'_{kl}(r) - \frac{\lambda_k}{r^2} v_{kl}(r) = -g'(\sigma_0(r)) u_{kl}(r), \tag{3.8}$$

$$u_{kl}(R) + \sigma'_0(R) c_{kl} = 0, \tag{3.9}$$

$$v'_{kl}(R) - g(\bar\sigma) c_{kl} = 0, \tag{3.10}$$

$$v_{kl}(R) + \frac{\gamma}{R^2}\Big(1 - \frac{\lambda_k}{n-1}\Big) c_{kl} = 0. \tag{3.11}$$

Solutions of (1.1) are innately nonunique: Any translation or rotation of a solution is again a solution. This nonuniqueness of solutions to (1.1) causes, correspondingly, a nonuniqueness of solutions to (3.1)–(3.5). To eliminate this nonuniqueness we assume that

$$c_{1l} = 0, \quad u_{1l}(r) = 0, \quad v_{1l}(r) = 0, \quad l = 1, 2, \dots, n. \tag{3.12}$$

Observe that, given $k \ge 2$, we have

$$(n-1) - \lambda_k = (n-1) - k^2 - (n-2)k \le -n - 1 < 0.$$

Using this fact it is not difficult to see that any solution of (3.7) is given by

$$u_{kl}(r) = \alpha_{kl} r^{\mu_k} \bar u_k(r), \tag{3.13}$$

where α_{kl} is an arbitrary constant,

$$\mu_k = \frac{1}{2}\Big[\sqrt{(n-2)^2 + 4\lambda_k} - (n-2)\Big] = k \ge 2, \tag{3.14}$$

and $\bar u_k(r)$ is the unique solution of the problem

$$\begin{cases} \bar u''_k(r) + \dfrac{a_k}{r} \bar u'_k(r) = f'(\sigma_0(r)) \bar u_k(r), \\ \bar u_k(0) = 1, \quad \bar u'_k(0) = 0, \end{cases} \tag{3.15}$$

where

$$a_k = 2\mu_k + n - 1 = 2k + n - 1 \ge n + 3. \tag{3.16}$$

Note that (3.15) is a linear equation, which is regular on $r > 0$. Hence the maximal existence interval of $\bar u_k(r)$, and, consequently, also that of $u_{kl}(r)$, is the same as that for $\sigma_0(r)$, namely, $[0, R^*)$. Besides, since $\sigma_0 \in C^\infty[0, R^*)$ and $a_k > 0$, we have

$\bar{u}_k \in C^\infty[0, R^*)$; cf. [6]. Substituting (3.13) into the right-hand side of (3.8) we infer that solutions of (3.8) are given by

$$v_{kl}(r) = \alpha_{kl} r^{\mu_k} \bar{v}_k(r) + \beta_{kl} r^{\mu_k}, \tag{3.17}$$

where α_{kl} is as before, β_{kl} is an arbitrary constant proportional to α_{kl}, i.e., $\beta_{kl} = c\alpha_{kl}$ for an arbitrary constant c, and $\bar{v}_k$ is the unique solution of the problem

$$\begin{cases} \bar{v}_k''(r) + \dfrac{a_k}{r}\bar{v}_k'(r) = -g'(\sigma_0(r))\bar{u}_k(r), \\ \bar{v}_k(0) = 0, \quad \bar{v}_k'(0) = 0. \end{cases} \tag{3.18}$$

Note that $\bar{v}_k(r)$, and, consequently, also $v_k(r)$, is defined for all $0 \le r < R^*$, and $\bar{v}_k \in C^\infty[0, R^*)$. Substituting (3.13) and (3.17) into (3.9)–(3.11) we get the following equations for α_{kl}, β_{kl}, and c_{kl}:

$$\begin{cases} R^{\mu_k}\bar{u}_k(R)\alpha_{kl} + \sigma_0'(R)c_{kl} = 0, \\ [R^{\mu_k}\bar{v}_k'(R) + \mu_k R^{\mu_k-1}\bar{v}_k(R)]\alpha_{kl} + \mu_k R^{\mu_k-1}\beta_{kl} - g(\bar{\sigma})c_{kl} = 0, \\ R^{\mu_k}\bar{v}_k(R)\alpha_{kl} + R^{\mu_k}\beta_{kl} + \dfrac{\gamma}{R^2}\Big(1 - \dfrac{\lambda_k}{n-1}\Big)c_{kl} = 0. \end{cases} \tag{3.19}$$

Hence, (3.1)–(3.5) has a nontrivial solution if and only if there exists $k \ge 2$ such that (3.19) has a nontrivial solution. In the following, we provide conditions on γ which guarantee that (3.19) has a nontrivial solution.

LEMMA 3.2. *The system* (3.19) *has a nontrivial solution if and only if* $\gamma = \gamma_k$, *where*

$$\gamma_k = \frac{(n-1)R^3}{(\lambda_k - n+1)\mu_k}\left[g(\bar{\sigma}) - \frac{\sigma_0'(R)}{\bar{u}_k(R)R^{a_k}}\int_0^R g'(\sigma_0(\rho))\bar{u}_k(\rho)\rho^{a_k}\,d\rho\right], \quad k \ge 2. \tag{3.20}$$

In this case the nontrivial solutions of (3.19) *are unique up to a constant factor. Moreover,* $\gamma_k > 0$ *for all* $k \ge 2$ *and there exists a* $k^* = k^*(f, g, R, n)$ *such that*

$$\gamma_{k+1} < \gamma_k \quad \forall \quad k \ge k^*. \tag{3.21}$$

Finally, $\lim_{k\to\infty}\gamma_k = 0$.

Proof. A simple computation shows that the determinant of the coefficient matrix of (3.19) is equal to the product of $R^{2\mu_k-1}\mu_k\bar{u}_k(R)$ with

$$\frac{R}{\mu_k}\left[g(\bar{\sigma}) + \frac{\bar{v}_k'(R)\sigma_0'(R)}{\bar{u}_k(R)}\right] + \frac{\gamma}{R^2}\left(1 - \frac{\lambda_k}{n-1}\right) \equiv D_k(\gamma).$$

Hence, (3.19) has a nontrivial solution if and only if $D_k(\gamma) = 0$, or

$$\gamma = \gamma_k \equiv \frac{(n-1)R^3}{(\lambda_k - n+1)\mu_k}\left[g(\bar{\sigma}) + \frac{\bar{v}_k'(R)\sigma_0'(R)}{\bar{u}_k(R)}\right].$$

From (3.18) we further infer that

$$\bar{v}_k'(r) = -\frac{1}{r^{a_k}}\int_0^r g'(\sigma_0(\rho))\bar{u}_k(\rho)\rho^{a_k}\,d\rho, \quad r \in [0, R^*), \tag{3.22}$$

which shows that γ_k is given by (3.20). If $\gamma = \gamma_k$, then clearly solutions of (3.19) are unique up to a constant factor. To prove $\gamma_k > 0$, we need only to show that

$$g(\bar{\sigma}) > \frac{\sigma_0'(R)}{\bar{u}_k(R)R^{a_k}} \int_0^R g'(\sigma_0(\rho))\bar{u}_k(\rho)\rho^{a_k}\,d\rho. \tag{3.23}$$

Since $a_k = 2\mu_k + n - 1$ (see (3.16)), we infer that $\rho^{a_k} \le R^{\mu_k - 1}\rho^{\mu_k + n}$ for $0 < \rho \le R$. Thus

$$\begin{aligned}\frac{\sigma_0'(R)}{\bar{u}_k(R)R^{a_k}} \int_0^R g'(\sigma_0(\rho))\bar{u}_k(\rho)\rho^{a_k}\,d\rho &\le \frac{\sigma_0'(R)}{\bar{u}_k(R)R^{\mu_k+n}} \int_0^R g'(\sigma_0(\rho))\bar{u}_k(\rho)\rho^{\mu_k+n}\,d\rho \\ &= \frac{1}{R^n}\int_0^R g'(\sigma_0(\rho))w(\rho)\rho^n\,d\rho,\end{aligned}$$

where $w(r) = \frac{\sigma_0'(R)r^{\mu_k}\bar{u}_k(r)}{R^{\mu_k}\bar{u}_k(R)}$. But (A_2) implies that $\bar{u}_k(r) > 0$, and we conclude that

$$w''(r) + \frac{n-1}{r}w'(r) - \frac{n-1}{r^2}w(r) \ge w''(r) + \frac{n-1}{r}w'(r) - \frac{\lambda_k}{r^2}w(r) = f'(\sigma_0(r))w(r),$$

and

$$w(0) = 0, \quad w(R) = \sigma_0'(R).$$

This implies, by comparison, that $w(r) < \sigma_0'(r)$ for $0 < r < R$, since by differentiating the first equation of (2.1) we get that

$$\sigma_0'''(r) + \frac{n-1}{r}\sigma_0''(r) - \frac{n-1}{r^2}\sigma_0'(r) = f'(\sigma_0(r))\sigma_0'(r),$$

and we know that $\sigma_0'(0) = 0$. Hence

$$\begin{aligned}\frac{1}{R^n}\int_0^R g'(\sigma_0(\rho))w(\rho)\rho^n\,d\rho &< \frac{1}{R^n}\int_0^R g'(\sigma_0(\rho))\sigma_0'(\rho)\rho^n\,d\rho \\ &= g(\sigma_0(R)) - \frac{n}{R^n}\int_0^R g(\sigma_0(\rho))\rho^{n-1}\,d\rho \\ &= g(\bar{\sigma}) + \frac{n}{R}p_0'(R) = g(\bar{\sigma}).\end{aligned}$$

This proves (3.23). Hence, $\gamma_k > 0$ for $k \ge 2$.

To verify (3.21), we first observe that (3.15) implies that

$$\bar{u}_k'(r) = \frac{1}{r^{a_k}}\int_0^r f'(\sigma_0(\rho))\bar{u}_k(\rho)\rho^{a_k}\,d\rho, \quad r \in [0, R^*).$$

Hence $\bar{u}_k$ is increasing. Let us introduce the notation

$$\delta_k := \frac{\sigma_0'(R)}{\bar{u}_k(R)R^{a_k}} \int_0^R g'(\sigma_0(\rho))\bar{u}_k(\rho)\rho^{a_k}\,d\rho.$$

Integration by parts shows that

$$\begin{aligned}\delta_k &= \frac{R\sigma_0'(R)g'(\bar\sigma)}{a_k+1} - \frac{\sigma_0'(R)}{(a_k+1)\bar u_k(R)R^{a_k}}\int_0^R g''(\sigma_0(\rho))\sigma_0'(\rho)\bar u_k(\rho)\rho^{a_k+1}d\rho\\ &\quad -\frac{\sigma_0'(R)}{(a_k+1)\bar u_k(R)R^{a_k}}\int_0^R g'(\sigma_0(\rho))\bar u_k'(\rho)\rho^{a_k+1}d\rho\\ &= \frac{R\sigma_0'(R)g'(\bar\sigma)}{a_k+1} - \frac{\sigma_0'(R)}{(a_k+1)\bar u_k(R)R^{a_k}}\int_0^R g''(\sigma_0(\rho))\sigma_0'(\rho)\bar u_k(\rho)\rho^{a_k+1}d\rho\\ &\quad -\frac{\sigma_0'(R)}{(a_k+1)\bar u_k(R)R^{a_k}}\int_0^R\int_0^\rho g'(\sigma_0(\rho))\rho f'(\sigma_0(\eta))\bar u_k(\eta)\eta^{a_k}d\eta d\rho\\ &\equiv \frac{R\sigma_0'(R)g'(\bar\sigma)}{a_k+1} - \frac{\sigma_0'(R)}{(a_k+1)}\varepsilon_k(R).\end{aligned}$$

Since $\bar u_k$ is increasing, we have $0 \le \bar u_k(\rho)\rho^{a_k}/\bar u_k(R)R^{a_k} \le (\rho/R)^{a_k}$ for $0 \le \rho \le R$, which implies that $\lim_{k\to\infty}\bar u_k(\rho)\rho^{a_k}/\bar u_k(R)R^{a_k} = 0$ for $0 \le \rho < R$. Hence, by dominated convergence, we see that $\lim_{k\to\infty}\varepsilon_k(R) = 0$, or

$$\delta_k = \frac{R\sigma_0'(R)g'(\bar\sigma)}{a_k+1}(1+o(1)) \quad \text{as } k\to\infty.$$

Substituting this expression into (3.20), we deduce that

$$\gamma_{k+1}-\gamma_k = -\frac{3(n-1)R^3g(\bar\sigma)}{k^4}(1+o(1)) \quad \text{as } k\to\infty,$$

so that γ_k is strictly decreasing for k sufficiently large. This completes the proof of (3.21). Finally, $\lim_{k\to\infty}c_k = \infty$ implies that (γ_k) is a null sequence. □

Remark 3.1. We note that

$$D_1(\gamma) = 0 \quad \text{for any } \gamma > 0.$$

Indeed, letting $c = \frac{n}{f(\sigma_0(0))}$, it is easy to verify that $\bar u_1(r) = c\frac{\sigma_0'(r)}{r}$ solves (3.15) in the case $k=1$. Thus using (3.22) we get

$$\begin{aligned}D_1(\gamma) &= R\Big[g(\bar\sigma) + \frac{\bar v_1'(R)\sigma_0'(R)}{\bar u_1(R)}\Big] = R[g(\bar\sigma)+c^{-1}R\bar v_1'(R)]\\ &= R\left[g(\bar\sigma) - \frac{1}{R^n}\int_0^R g'(\sigma_0(\rho))\sigma_0'(\rho)\rho^n d\rho\right] = \frac{n}{R}p_0'(R) = 0.\end{aligned}$$

This implies that for $k=1$, the system (3.19) has nontrivial solutions for any $\gamma > 0$. This nonuniqueness is a consequence of the fact that the system (1.1) is invariant under translations of Ω.

We now summarize the main result of this section.

THEOREM 3.3. *The system of* (3.1)–(3.5) *has a nontrivial solution if and only if* $\gamma = \gamma_k$ *for some* $k \ge 2$.

In the succeeding section we shall express (1.1) as a bifurcation problem in a suitable function space which can be treated by standard bifurcation theory in Banach spaces. We shall see that Theorem 3.3 implies that $\gamma_k's$ are bifurcation points of the full system (1.1), provided $k \ge k^*$.

4. Nonbifurcation at $\gamma \neq \gamma_k$. In this section we shall reduce the system of equations (1.1) into the form (1.4), and use the reduced equation and the implicit function theorem to prove that if $\gamma \neq \gamma_k$ $(k \geq 2)$, then no bifurcation occurs at γ.

For a given $R > 0$, we denote by B_R the open ball in $\mathbb{R}^n$ centered at the origin with radius R, and for a given function $\eta \in C(\mathbb{S}^{n-1})$ satisfying

$$\max_{\omega \in \mathbb{S}^{n-1}} |\eta(\omega)| < R,$$

we denote

$$\Omega_\eta = \{x \in \mathbb{R}^n :\ x = r\omega,\ 0 \leq r < R + \eta(\omega),\ \omega \in \mathbb{S}^{n-1}\}.$$

Given two Banach spaces X and Y, we denote by $L(X, Y)$ the Banach space of all bounded linear mappings from X into Y, by $C(X, Y)$ the Fréchet space of all continuous mappings from X into Y, and by $C^1(X, Y)$ the Fréchet space of all continuously Fréchet differentiable mappings from X to Y. We shall use both $DF(u)$ and $F'(u)$ to denote the Fréchet derivative. For a positive integer m, we denote by $C^m(X, Y)$ the Fréchet space of all m-times continuously Fréchet differentiable mappings from X into Y, and we write $C^\infty(X, Y) = \cap_{m\geq 1} C^m(X, Y)$. Given $0 < \mu < 1$, a positive integer m, and two open sets $\Omega_1, \Omega_2 \subseteq \mathbb{R}^n$, we denote by $\mathrm{Diff}^{m+\mu}(\overline{\Omega}_1, \overline{\Omega}_2)$ the set of all $C^{m+\mu}$-diffeomorphisms from $\overline{\Omega}_1$ onto $\overline{\Omega}_2$. We also set $\mathrm{Diff}^\infty(\overline{\Omega}_1, \overline{\Omega}_2) = \cap_{m\geq 1} \mathrm{Diff}^m(\overline{\Omega}_1, \overline{\Omega}_2)$.

We now use the so-called *Hanzawa transformation* to convert the original problem (1.1) into a nonlinear problem on the fixed reference domain B_R. For this purpose pick a function $\phi \in C^\infty[0, \infty)$ such that

$$\phi(t) = 0 \quad \text{for} \ \ 0 \leq t \leq \frac{1}{2}, \qquad \phi(1) = 1, \tag{4.1}$$

and

$$0 \leq \phi'(t) \leq C_0 \quad \text{for} \ \ t \geq 0, \qquad \phi''(t) \geq 0 \quad \text{for} \ \ 0 \leq t \leq 1. \tag{4.2}$$

By the mean value theorem we have $C_0 > 2$. Given a number $R > 0$ and a function $\eta \in C(\mathbb{S}^{n-1})$ satisfying

$$\max_{\omega \in \mathbb{S}^{n-1}} |\eta(\omega)| < \frac{R}{C_0 - 1}, \tag{4.3}$$

we consider the mapping $\Psi_\eta :\ \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$\Psi_\eta(x) = x - \eta(\omega)\phi\left(\frac{|x|}{R + \eta(\omega)}\right)\omega, \quad x \in \mathbb{R}^n,$$

where $\omega = \frac{x}{|x|}$ for $x \neq 0$. Using (4.1)–(4.3), one can easily verify that if $\eta \in C^{m+\mu}(\mathbb{S}^{n-1})$ for some $m \in \mathbb{N}$ and $0 \leq \mu < 1$, then

$$\Psi_\eta \in \mathrm{Diff}^{m+\mu}(\mathbb{R}^n, \mathbb{R}^n) \bigcap \mathrm{Diff}^{m+\mu}(\overline{\Omega}_\eta, \overline{B}_R).$$

As usual we denote by Ψ_η^* and $(\Psi_\eta)_*$ the pullback and push-forward operators, respectively, induced by Ψ_η, i.e.,

$$\Psi_\eta^* u = u \circ \Psi_\eta \quad \text{for} \ \ u \in C(\overline{B}_R), \qquad (\Psi_\eta)_* v = v \circ \Psi_\eta^{-1} \quad \text{for} \ \ v \in C(\overline{\Omega}_\eta).$$

Clearly, if $\eta \in C^{m+\mu}(\mathbb{S}^{n-1})$ for some $m \in \mathbb{N}$ and $0 \le \mu < 1$, then

$$\Psi_\eta^* \in L(C^{m+\mu}(\overline{B}_R), C^{m+\mu}(\overline{\Omega}_\eta)), \quad (\Psi_\eta)_* \in L(C^{m+\mu}(\overline{\Omega}_\eta), C^{m+\mu}(\overline{B}_R)).$$

We denote by Ψ, Ψ^*, and Ψ_*, respectively, the three mappings

$$\eta \mapsto \Psi_\eta, \quad \eta \mapsto \Psi_\eta^*, \quad \eta \mapsto (\Psi_\eta)_*.$$

Since $\phi \in C^\infty[0,\infty)$, we also have

$$\Psi \in C^\infty(C^{m+\mu}(\mathbb{S}^{n-1}) \cap M(R, C_0), \mathrm{Diff}^{m+\mu}(\overline{\Omega}_\eta, \overline{B}_R)),$$

where

$$M(R, C_0) = \left\{ \eta \in C(\mathbb{S}^{n-1}) : \max_{\omega \in \mathbb{S}^{n-1}} |\eta(\omega)| < \frac{R}{C_0 - 1} \right\}.$$

Similar mapping properties of Ψ^* and Ψ_* can easily be formulated.

Given $\eta \in C^{2+\mu}(\mathbb{S}^{n-1}) \cap M(R, C_0)$, we define operators $L(\eta): C^{2+\mu}(\overline{B}_R) \to C^\mu(\overline{B}_R)$ and $N(\eta): C^{2+\mu}(\overline{B}_R) \to C^{1+\mu}(\partial B_R)$ by

$$L(\eta) = (\Psi_\eta)_* \circ \Delta \circ \Psi_\eta^*, \qquad N(\eta) = (\Psi_\eta|_{\partial\Omega_\eta})_* \circ \frac{\partial}{\partial \nu} \circ \Psi_\eta^*,$$

where $\partial/\partial\nu$ denotes the outward normal derivative operator on $\partial\Omega_\eta$, and $\Psi_\eta|_{\partial\Omega_\eta}$ denotes the restriction of Ψ_η on $\partial\Omega_\eta$. Clearly, if $\eta \in C^{m+\mu}(\mathbb{S}^{n-1})$ $(m \ge 2)$, then

$$L(\eta) \in L(C^{m+\mu}(\overline{B}_R), C^{m+\mu-2}(\overline{B}_R)), \quad N(\eta) \in L(C^{m+\mu}(\overline{B}_R), C^{m+\mu-1}(\partial B_R)).$$

Moreover, the mappings $\eta \mapsto L(\eta)$ and $\eta \mapsto N(\eta)$ are smooth on $C^{m+\mu}(\mathbb{S}^{n-1}) \cap M(R, C_0)$. In order to handle the right-hand sides of (1.1), we define the operators $F, G : C^{m+\mu}(\overline{B}_R) \to C^{m+\mu}(\overline{B}_R)$ and $\Gamma_0 : C^{m+\mu}(\overline{B}_R) \to C^{m+\mu}(\partial B_R)$ by

$$F(u) = f \circ u, \quad G(u) = g \circ u, \quad \Gamma_0 u = u|_{\partial B_R}$$

for $u \in C^{m+\mu}(\overline{B}_R)$, as well as $B : C^{m+2+\mu}(\mathbb{S}^{n-1}) \to C^{m+\mu}(\partial B_R)$, where

$$\begin{aligned} B(\eta)(R\omega) \;=\;& \text{the mean curvature of the hypersurface } r = R + \eta(\omega) \\ & \text{at the point } x = (R + \eta(\omega))\omega \end{aligned}$$

for $\eta \in C^{m+2+\mu}(\mathbb{S}^{n-1})$. Again it is not difficult to verify that these mappings depend smoothly on their argument.

With the above notations we can now transform the problem (1.1) into the following system: Find $(u, v, \eta) \in C^{2+\mu}(\overline{B}_R) \times C^{2+\mu}(\overline{B}_R) \times C^{2+\mu}(\mathbb{S}^{n-1})$, such that

$$L(\eta)u = F(u), \tag{4.4}$$
$$\Gamma_0 u = \bar{\sigma}, \tag{4.5}$$
$$L(\eta)v = -G(u), \tag{4.6}$$
$$N(\eta)v = 0, \tag{4.7}$$
$$\Gamma_0 v = \gamma B(\eta). \tag{4.8}$$

Fix $m \in \mathbb{N}$ with $m \ge 2$ and $0 < \mu < 1$. Given $\eta \in C^{m+\mu}(\mathbb{S}^{n-1})$, it follows from Theorems 9.5 and 12.8 in [10] that the problem (4.4), (4.5) has a unique solution

$u \in C^{m+\mu}(\overline{B}_R)$, satisfying $0 < u(x) \le \bar{\sigma}$ for $x \in \overline{B}_R$. We define a mapping $A_1 : C^{m+\mu}(\mathbb{S}^{n-1}) \to C^{m+\mu}(\overline{B}_R)$ by setting $A_1(\eta) = u$.

Next, let us introduce the functional Φ on $C^{m+\mu}(\mathbb{S}^{n-1})$ by setting

$$\Phi(\eta) = \int_{B_R} G(A_1(\eta))(x) J_\eta(x) dx \qquad \text{for } \eta \in C^{m+\mu}(\mathbb{S}^{n-1}), \tag{4.9}$$

where $J_\eta(x)$ is the determinant of the Jacobian of the transformation $x \mapsto \Psi_\eta^{-1}(x)$. Substituting $u = A_1(\eta)$ into (4.6) and using the Fredholm property of (4.6) (4.7), we see that $\Phi(\eta) = 0$ is a necessary (and sufficient) condition for the solvability of problem (4.6), (4.7). Later we shall see that $\Phi \in C^\infty(C^{m+\mu}(\mathbb{S}^{n-1}) \cap M(R, C_0), \mathbb{R})$, and that $\Phi'(0) \ne 0$. Therefore, $\mathcal{M} := \Phi^{-1}(0)$ is a C^∞ Banach submanifold of codimension 1 in a small neighborhood of the origin of $C^{m+\mu}(\mathbb{S}^{n-1})$. By elliptic Schauder theory, given any $\eta \in \mathcal{M}$, the problem (4.6), (4.7) has a solution $v \in C^{m+\mu}(\overline{B}_R)$, which is unique up to a constant. Let us denote by v the particular solution which is zero at the origin, and define a mapping $A_2 : \mathcal{M} \to C^{m+\mu}(\overline{B}_R)$ by setting $A_2(\eta) = v$. Finally, we set $A = \Gamma_0 \circ A_2$. Then, clearly, $A : \mathcal{M} \to C^{m+\mu}(\partial B_R)$.

With the above notations, it is not difficult to verify that the problem (4.4)–(4.8) reduces to the following problem: Find $\eta \in C^{m+\mu}(\mathbb{S}^{n-1})$ and $c \in \mathbb{R}$ such that

$$\begin{cases} A(\eta) + c = \gamma B(\eta), \\ \Phi(\eta) = 0. \end{cases} \tag{4.10}$$

We summarize with the following lemma.

LEMMA 4.1. *The problem* (1.1) *is equivalent to the system of* (4.10). *More precisely, if* (σ, p, Ω), *where* $\sigma \in C^{m+\mu}(\overline{\Omega})$, $p \in C^{m+\mu}(\overline{\Omega})$, $\Omega = \{x \in R^n : r < \eta(\omega), r = |x|, \omega = x/|x|\}$, *and* $\eta \in C^{m+\mu}(\mathbb{S}^{n-1})$ *is a solution of* (1.1), *then there exists a unique real number c such that* (η, c) *is a solution of* (4.10). *Conversely, if* $(\eta, c) \in C^{m+\mu}(\mathbb{S}^{n-1}) \times \mathbb{R}$ *is a solution of* (4.10), *then there exists a unique pair of functions* $(\sigma, p) \in C^{m+\mu}(\overline{\Omega}) \times C^{m+\mu}(\overline{\Omega})$, *such that* (σ, p, Ω), *with* Ω *as above, is a solution of* (1.1).

Remark. Using the above equivalence and a standard bootstrapping argument, one can easily show that if (σ, p, Ω) is a solution of (1.1)–(1.5), then $\eta \in C^\infty(\mathbb{S}^{n-1})$ and, consequently, also $\sigma \in C^\infty(\overline{\Omega})$, $p \in C^\infty(\overline{\Omega})$. Besides, if (σ, p, Ω) is a solution of (1.1), then the value of c in (5.20) is clearly equal to $p(0)$.

We now prove the following preliminary result.

LEMMA 4.2. *Let* $m \in \mathbb{N} \cup \{\infty\}$ *with* $m \ge 2$ *and* $0 < \mu < 1$ *be given. Then we have the following assertions:*

(i) $\Phi \in C^\infty(C^{m+\mu}(\mathbb{S}^{n-1}) \cap M(R, C_0), \mathbb{R})$ *with* $\Phi'(0)\mathbf{1} < 0$, *where* $\mathbf{1}$ *represents the function on* $\mathbb{S}^{n-1}$ *taking identically the value* 1.

(ii) *There exists a neighborhood* $\mathcal{N}$ *of* 0 *in* $C^{m+\mu}(\mathbb{S}^{n-1})$ *such that* $\mathcal{M} := \Phi^{-1}(0) \cap \mathcal{N}$ *is a* C^∞*-Banach submanifold of codimension* 1. *Moreover, we have*

$$A \in C^\infty(\mathcal{M}, C^{m+\mu}(\partial B_R)). \tag{4.11}$$

Proof. We define a mapping

$$\mathcal{F} : (C^{m+\mu}(\mathbb{S}^{n-1}) \cap M(R, C_0)) \times C^{m+\mu}(\overline{B}_R) \to C^{m+\mu-2}(\overline{B}_R) \times C^{m+\mu}(\partial B_R)$$

by setting

$$\mathcal{F}(\eta, u) := (L(\eta)u - F(u), \Gamma_0 u - \bar{\sigma}).$$

Clearly, $\mathcal{F}$ is smooth and the system (4.4), (4.5) is equivalent to $\mathcal{F}(\eta, u) = 0$. Thus, by the definition of A_1, given any $\eta \in C^{m+\mu}(\mathbb{S}^{n-1}) \cap M(R, C_0)$, this equation has a unique solution $u = A_1(\eta)$. Consider the partial Fréchet derivative $D_u\mathcal{F}(\eta, u)$ of $\mathcal{F}(\eta, u)$ in u. An easy computation reveals that

$$D_u\mathcal{F}(\eta, u)v = (L(\eta)v - F'(u)v, \Gamma_0 v) = (L(\eta)v - f'(u)v, \Gamma_0 v) \quad \forall v \in C^{m+\mu}(\overline{B}_R).$$

Given any $(\eta, u) \in (C^{m+\mu}(\mathbb{S}^{n-1}) \cap M(R, C_0)) \times C^{m+\mu}(\overline{B}_R)$, we have that $L(\eta)$ is a uniformly elliptic operator in divergence form without the zero order term and $f'(u) \geq 0$. Hence standard elliptic theory implies that for any $(w, z) \in C^{m+\mu-2}(\overline{B}_R) \times C^{m+\mu}(\partial B_R))$, the elliptic boundary value problem $L(\eta)v - f'(u)v = w, \quad \Gamma_0 v = z$ has a unique solution $v \in C^{m+\mu}(\overline{B}_R)$. This means that the mapping $D_u\mathcal{F}(\eta, u) : C^{m+\mu}(\overline{B}_R) \to C^{m+\mu-2}(\overline{B}_R) \times C^{m+\mu}(\partial B_R)$ is an isomorphism. It follows by the Banach inverse mapping theorem that its inverse is also bounded. Thus, by the implicit function theorem we conclude that A_1 is smooth with

$$A_1'(\eta) = -(D_u\mathcal{F}(\eta, A_1(\eta)))^{-1} \circ D_\eta\mathcal{F}(\eta, A_1(\eta)).$$

Since $G \in C^\infty(C^{m+\mu}(\overline{B}_R), C^{m+\mu}(\overline{B}_R))$ and since it is obvious that the mapping $\eta \mapsto J_\eta$ belongs to $C^\infty(C^{m+\mu}(\mathbb{S}^{n-1}) \cap M(R, C_0), C^{m+\mu-1}(\overline{B}_R))$, we immediately conclude that $\Phi \in C^\infty(C^{m+\mu}(\mathbb{S}^{n-1}) \cap M(R, C_0), \mathbb{R})$.

Next, we compute $\Phi'(0)\mathbf{1}$. The change of variables $x \mapsto \Psi_\eta(x)$ in the integral on the right-hand side of (4.9) shows that

$$\Phi(\eta) = \int_{\Omega_\eta} g(\sigma_\eta(x))dx,$$

where σ_η denotes the solution of the boundary value problem

$$\Delta\sigma_\eta = f(\sigma_\eta) \quad \text{in } \Omega_\eta, \qquad \sigma_\eta = \bar{\sigma} \quad \text{on } \partial\Omega_\eta.$$

In the case $\eta = \varepsilon\mathbf{1}$, we simply denote $\sigma_{\varepsilon\mathbf{1}}$ by σ_ε, so that $\Phi(\varepsilon\mathbf{1}) = \int_{B_{R+\varepsilon}} g(\sigma_\varepsilon(x))dx$. It follows that

$$\begin{aligned}
\Phi'(0)\mathbf{1} &= \frac{d}{d\varepsilon}\Phi(\varepsilon\mathbf{1})\Big|_{\varepsilon=0} = g(\bar{\sigma})\frac{d}{d\varepsilon}|B_{R+\varepsilon}|\Big|_{\varepsilon=0} + \int_{B_R} g'(\sigma_0(|x|))\frac{d}{d\varepsilon}\sigma_\varepsilon(x)\Big|_{\varepsilon=0}dx \\
&= c_n R^{n-1}g(\bar{\sigma}) + c_n\int_0^R g'(\sigma_0(r))v_0(r)r^{n-1}dr,
\end{aligned}$$

where c_n denotes the surface measure of $\mathbb{S}^{n-1}$, and where v_0 is the solution of the boundary value problem

$$\begin{cases} v_0''(r) + \dfrac{n-1}{r}v_0'(r) = f'(\sigma_0(r))v_0(r) \quad \text{for } 0 < r < R, \\ v_0'(0) = 0, \quad v_0(R) = -\sigma_0'(R). \end{cases}$$

Using the maximum principle, one can easily show that $v_0(r) < -\sigma'(r)$ for $0 < r < R$. Hence

$$\begin{aligned}
\Phi'(0)\mathbf{1} &< c_n R^{n-1}g(\bar{\sigma}) - c_n\int_0^R g'(\sigma_0(r))\sigma_0'(r)r^{n-1}dr \\
&= \frac{c_n}{R}\int_0^R g'(\sigma_0(r))\sigma_0'(r)r^n dr - c_n\int_0^R g'(\sigma_0(r))\sigma_0'(r)r^{n-1}dr < 0.
\end{aligned}$$

Here we use the fact that

$$g(\bar{\sigma}) - \frac{1}{R^n}\int_0^R g'(\sigma_0(r))\sigma_0'(r)r^n dr = \frac{n}{R}p_0'(R) = 0.$$

This proves assertion (i).

(ii) By continuity of Φ', we infer that there exists a neighborhood $\mathcal{N}$ of 0 in $C^{m+\mu}(\mathbb{S}^{n-1})$ such that $\Phi'(\eta)\mathbf{1} \neq 0$ for all $\eta \in \mathcal{N}$. It follows that $\mathcal{M} = \Phi^{-1}(0) \cap \mathcal{N}$ is a C^∞ Banach submanifold of codimension 1 in $C^{m+\mu}(\mathbb{S}^{n-1})$. Based on the implicit function theorem one sees that, similarly as for A_1, that A_2 is smooth. Hence $A = \Gamma_0 \circ A_2 \in C^\infty(\mathcal{M}, C^{m+\mu}(\partial B_R))$. This completes the proof. □

Recall that $D_1(\gamma) = 0$ for all $\gamma > 0$; see Remark 3.1. This implies that the operator $(\eta, c) \mapsto A'(0)\eta + c - \gamma B'(0)\eta$, $(\eta, c) \in C^{m+\mu}(\mathbb{S}^{n+1}) \times \mathbb{R}$, is not injective for any $\gamma > 0$. For this reason we shall restrict the operators A, B on the subspace $C_2^{m+\mu}(\mathbb{S}^{n-1})$ of $C^{m+\mu}(\mathbb{S}^{n+1})$, which is defined as follows: We first let

$$H_1(\mathbb{S}^{n-1}) = \text{the span of all spherical harmonics of degree } 1,$$

and then set

$$C_2^{m+\mu}(\mathbb{S}^{n-1}) = \{\eta \in C^{m+\mu}(\mathbb{S}^{n-1}) :\ \eta \text{ is orthogonal to } H_1(\mathbb{S}^{n-1}) \text{ in } L^2(\mathbb{S}^{n-1})\}.$$

It can be easily shown that

$$C^{m+\mu}(\mathbb{S}^{n-1}) = H_1(\mathbb{S}^{n-1}) \oplus C_2^{m+\mu}(\mathbb{S}^{n-1}).$$

Given a point $x_0 \in \mathbb{R}$, we denote by S_{x_0} the translation on $\mathbb{R}^n$ induced by x_0, i.e., $S_{x_0}(x) = x + x_0$ for $x \in \mathbb{R}^n$. Further, let $r_0 = |x_0|$, $\omega_0 = x_0/|x_0|$, and assume that $r_0 \in (0, R/4)$. Given $\eta \in C(\mathbb{S}^{n-1})$ such that

$$\|\eta\|_{L^\infty(\mathbb{S}^{n-1})} < \min\left\{\frac{R}{C_0 - 1}, \frac{R}{4}\right\},$$

consider the image of the hypersurface $r = R + \eta(\omega)$ under the translation S_{x_0}. Clearly, this hypersurface has the equation $r = R + \tilde{\eta}(\omega)$, where $\tilde{\eta}$ is uniquely determined by η and x_0. We use the notation

$$\tilde{\eta} = S_{x_0}^*(\eta).$$

An explicit expression of $S_{x_0}^*$ is given by

$$\tilde{\eta}(\omega') = \sqrt{[R + \eta(\omega)]^2 + r_0^2 + 2r_0[R + \eta(\omega)]\omega \cdot \omega_0} - R,$$

where $\omega' \in \mathbb{S}^{n-1}$ and $\omega \in \mathbb{S}^{n-1}$ are connected by the relation

$$\omega' = \frac{[R + \eta(\omega)]\omega + r_0\omega_0}{\sqrt{[R + \eta(\omega)]^2 + r_0^2 + 2r_0[R + \eta(\omega)]\omega \cdot \omega_0}}.$$

Now, since the problem (1.1) is invariant under translations, we see that if η is a solution of (4.10), then for any $x \in \mathbb{R}^n$ with $|x|$ sufficiently small, $\tilde{\eta} = S_x^*(\eta)$ is also a solution of (4.10). This shows that if we denote by $B_\varepsilon(0)$ the ball in $\mathbb{R}^n$ with radius ε, and assume that η is a solution of (5.20), then for $\varepsilon > 0$ small enough we have

$$\left\{\begin{array}{l} A(S_x^*(\eta)) + c = \gamma B(S_x^*(\eta)) \\ \Phi(S_x^*(\eta)) = 0 \end{array}\right. \qquad \text{for any } \; x \in B_\varepsilon(0). \tag{4.12}$$

Hence, for any solution η of (4.10) which lies in $C_2^{m+\mu}(\mathbb{S}^{n-1})$, we get a new solution of (4.10) by applying S_x^* to η for any $x \in B_\varepsilon(0)$. The following lemma shows that the converse is also true, i.e., any small solution of (4.10) can be obtained by a suitable translation of a solution belonging to $C_2^{m+\mu}(\mathbb{S}^{n-1})$.

LEMMA 4.3. *For $\varepsilon > 0$ sufficiently small, let $\mathcal{B}_\varepsilon$ be the ball in $C^{m+\mu}(\mathbb{S}^{n-1})$ with radius ε. Then for any $\eta \in \mathcal{B}_\varepsilon$, there exists a unique $\xi \in C_2^{m+\mu}(\mathbb{S}^{n-1})$ and a unique $x \in B_\varepsilon$ such that $S_x^*(\xi) = \eta$.*

Proof. We introduce a mapping $F : \mathcal{B}_\varepsilon \times B_\varepsilon \to \mathbb{R}^n$ by defining the jth component of $F(\eta, x)$ to be

$$F_j(\eta, x) = nc_n^{-1} \int_{\mathbb{S}^{n-1}} S_{-x}^*(\eta)(\omega)\omega_j d\omega, \quad j = 1, 2, \dots, n,$$

where c_n denotes the surface measure of $\mathbb{S}^{n-1}$, and $d\omega$ represents the volume element on $\mathbb{S}^{n-1}$. It can be easily shown that $F \in C^{m+\mu}(\mathcal{B}_\varepsilon \times B_\varepsilon, \mathbb{R}^n)$. Besides, it is clear that $F(0,0) = 0$, and a simple calculation shows that $D_x F(0,0) = -id$. Hence, by the implicit function theorem, we infer that there exist a small neighborhood $\mathcal{O}$ of the origin of $C^{m+\mu}(\mathbb{S}^{n-1})$ and a C^m-mapping $\varphi : \mathcal{O} \to B_\varepsilon$ such that $\varphi(0) = 0$, $F(\eta, \varphi(\eta)) = 0$ for all $\eta \in \mathcal{O}$, and $x = \varphi(\eta)$ is the unique solution of the equation $F(\eta, x) = 0$ in B_ε for every fixed $\eta \in \mathcal{B}_\varepsilon$. Let $\xi = S_{-x}^*(\eta) = S_{-\varphi(\eta)}^*(\eta)$. Then, clearly, $\xi \in C^{m+\mu}(\mathbb{S}^{n-1})$, $S_x^*(\xi) = \eta$, and the condition $F(\eta, x) = 0$ implies that ξ is orthogonal to all first order spherical harmonics in $L^2(\mathbb{S}^{n-1})$, so that $\xi \in C_2^{m+\mu}(\mathbb{S}^{n-1})$, as desired. ∎

Based on Lemma 4.3, we may assume without restriction that

$$\eta \in C_2^{m+\mu}(\mathbb{S}^{n-1}). \tag{4.13}$$

Let R and $\overline{p}_0(r)$ be as in section 2. We know that R, $\overline{p}_0$ are independent of γ. By the definition of A, B it is clear that

$$A(0) = \overline{p}_0(R), \quad B(0) = \frac{1}{R}. \tag{4.14}$$

We now set

$$c_\gamma = \frac{\gamma}{R} - \overline{p}_0(R) \quad \text{for } \; \gamma > 0,$$

and remark that given any $\gamma > 0$, the pair $(\eta, c) = (0, c_\gamma)$ is a solution of (4.10).

LEMMA 4.4. *Let $\gamma_0 > 0$. If $\gamma_0 \neq \gamma_k$, $k = 2, 3, \dots,$ then there exists $\varepsilon > 0$ such that for any $\gamma \in (\gamma_0 - \varepsilon, \gamma_0 + \varepsilon)$ the pair $(0, c_\gamma)$ is the only solution of* (4.10)*, i.e., γ_0 is not a bifurcation point of* (4.10).

Proof. We set

$$X = (\mathcal{M} \cap C_2^{m+\mu}(\mathbb{S}^{n-1})) \times \mathbb{R}, \quad Y = C^{m-2+\mu}(\mathbb{S}^{n-1}),$$

and define a mapping $F : X \times \mathbb{R} \to Y$ by

$$F(\zeta, \gamma) = A(\eta) + c - \gamma B(\eta), \quad \zeta = (\eta, c) \in X, \ \ \gamma \in \mathbb{R}.$$

Here we regard X as a smooth Banach submanifold of $C_2^{m+\mu}(\mathbb{S}^{n-1}) \times \mathbb{R}$. Clearly, $F \in C^\infty(X, Y)$ and (4.10) is equivalent to

$$F(\zeta, \gamma) = 0. \tag{4.15}$$

We already know that given $\gamma > 0$, this equation has the solution $\zeta = \zeta_\gamma \equiv (0, c_\gamma)$. Thus, by the Banach inverse mapping theorem and the implicit function theorem, it suffices to prove that for any $\gamma \neq \gamma_k$, $k = 2, 3, \ldots$, the operator $D_\zeta F(\zeta_\gamma, \gamma) : T_{\zeta_\gamma}(X) \to T_0(Y)$ is an isomorphism. Here $T_{\zeta_\gamma}(X)$ denotes the tangent space of X at ζ_γ. Observe that

$$T_{\zeta_\gamma}(X) = T_0(\mathcal{M} \cap C_2^{m+\mu}(\mathbb{S}^{n-1})) \times \mathbb{R},$$

and, given $\zeta = (\eta, c) \in T_{\zeta_\gamma}(X)$, we have

$$D_\zeta F(\zeta_\gamma, \gamma)\zeta = A'(0)\eta + c - \gamma B'(0)\eta.$$

Hence, for any $\xi \in T_0(Y) = Y$, the equation $D_\zeta F(\zeta_\gamma, \gamma)\zeta = \xi$ is equivalent to the system

$$\begin{cases} A'(0)\eta + c - \gamma B'(0)\eta = \xi, \\ \Phi'(0)\eta = 0, \quad \eta \in C_2^{m+\mu}(\mathbb{S}^{n-1}). \end{cases} \tag{4.16}$$

In particular, $D_\zeta F(\zeta_\gamma, \gamma)\zeta = 0$ is equivalent to the system

$$\begin{cases} A'(0)\eta + c - \gamma B'(0)\eta = 0, \\ \Phi'(0)\eta = 0, \quad \eta \in C_2^{m+\mu}(\mathbb{S}^{n-1}). \end{cases} \tag{4.17}$$

Since (4.17) is the linearization of (4.10) with the constraint $\eta \in C_2^{m+\mu}(\mathbb{S}^{n-1})$, and (4.10) is equivalent to (1.1), whose linearization is (3.1)–(3.5), we see that (4.17) is equivalent to (3.1)–(3.5) subject to the constraint $\eta \in C_2^{m+\mu}(\mathbb{S}^{n-1})$. By Theorem 3.3, we know that if $\gamma \neq \gamma_k$, $k = 2, 3, \ldots$, then (3.1)–(3.5) does not have nontrivial solutions in $C_2^{m+\mu}(\mathbb{S}^{n-1})$. Hence, if $\gamma \neq \gamma_k$, $k = 1, 2, \ldots$, then the pair $(0, c_\gamma)$ is the only solution to (4.17), i.e., $D_\zeta F(\zeta_\gamma, \gamma)$ is injective.

Next, by assertion (i) of Lemma 4.2 we see that the curve in $C^{m+\mu}(\mathbb{S}^{n-1})$ starting from $0 \in \mathcal{M}$ with tangent $\mathbf{1}$: $t \mapsto t\mathbf{1}$ is transverse to $\mathcal{M}$. Since $\mathcal{M}$ is of codimension 1 in $C^{m+\mu}(\mathbb{S}^{n-1})$, it follows that $C^{m+\mu}(\mathbb{S}^{n-1}) = T_0(\mathcal{M}) \oplus \mathbb{R}\mathbf{1}$, and the mapping $\zeta = (\eta, c) \to \eta + c\mathbf{1}$ is an isomorphism from $T_{\zeta_\gamma}(\mathcal{M} \times \mathbb{R}) = T_0(\mathcal{M}) \times \mathbb{R}$ to $C^{m+\mu}(\mathbb{S}^{n-1})$. For $\overline{\eta} \in C^{m+\mu}(\mathbb{S}^{n-1})$, let $\overline{\eta} = \eta + c\mathbf{1}$, where $\eta \in T_0(\mathcal{M})$ and $c \in \mathbb{R}$. The above argument implies that (η, c) is uniquely determined by $\overline{\eta}$. Clearly,

$$B'(0)\overline{\eta} = B'(0)\eta - \frac{c}{R^2}.$$

We define a bounded linear operator $L_\gamma : \ C^{m+\mu}(\mathbb{S}^{n-1}) \to C^{m+\mu}(\mathbb{S}^{n-1})$ by

$$L_\gamma \overline{\eta} = A'(0)\eta + c - \frac{\gamma c}{R^2}.$$

Then (4.16) is equivalent to

$$L_\gamma \overline{\eta} - \gamma B'(0)\overline{\eta} = \xi, \quad \text{with } \ \overline{\eta} \in C_2^{m+\mu}(\mathbb{S}^{n-1}).$$

Note that $B'(0)$ is a linear second order elliptic differential operator on $\mathbb{S}^{n-1}$. Hence the well-known Fredholm theory for elliptic operators implies that, given any $\xi \in C^{m-2+\mu}(\mathbb{S}^{n-1})$, the solvability of the above equation is equivalent to the uniqueness of solutions of the corresponding homogeneous equation. Thus, by the uniqueness assertion we have just proved, it follows that if $\gamma \neq \gamma_k$, then for any given $\xi \in C^{m-2+\mu}(\mathbb{S}^{n-1})$ this equation has a solution $\overline{\eta} \in C_2^{m+\mu}(\mathbb{S}^{n-1})$. Hence, (4.16) is solvable for any $\xi \in C^{m-2+\mu}(\mathbb{S}^{n-1})$, implying that the operator $D_\zeta F(\zeta_\gamma, \gamma)$ is surjective. This completes the proof. □

By combining Lemmas 4.1 and 4.4 we get the following theorem.

THEOREM 4.5. *If $\gamma \neq \gamma_k$, $k = 2, 3, \ldots$, then γ is not a bifurcation point of the problem* (1.1).

5. Bifurcation at $\boldsymbol{\gamma_k}$. In this section we prove that γ_k is a bifurcation point, provided $k \geq k^*$; cf. Lemma 3.2. We first tackle the case $n = 2$ in some detail and then indicate how to treat the general case $n \geq 3$.

5.1. The case $\boldsymbol{n = 2}$. Throughout this subsection we use (r, θ) to indicate polar coordinates in $\mathbb{R}^2$. We shall also frequently identify a 2π-periodic function η on $\mathbb{R}$ with the function $\psi^*(\eta) = \eta \circ \psi$ on $\mathbb{S}^1$, where $\psi : \mathbb{S}^1 \to \mathbb{R}$, $\psi(\cos\theta, \sin\theta) = \theta$ for $\theta \in [0, 2\pi)$. The Banach space of all 2π-periodic functions on $\mathbb{R}$ of class $C^{m+\mu}$ is denoted by $C_{2\pi}^{m+\mu}$. Given a positive integer l, we also introduce the space

$$C_{2\pi,l}^\infty = \Big\{\eta(\theta) = a_0 + \sum_{k=1}^\infty [a_k \cos(kl\theta) + b_k \sin(kl\theta)] : \ a_0, a_k, b_k \in \mathbb{R}, \text{ and } \forall\, N > 0 \ \exists\, C_N \geq 0 : \ |a_k| + |b_k| \leq C_N (1+k)^{-N}, \ k = 1, 2, \cdots \Big\}.$$

It is well known that $C_{2\pi}^\infty = C_{2\pi,l=1}^\infty$. Furthermore, by writing $C_{2\pi,l,0}^\infty$ and $C_{2\pi,l,1}^\infty$ for the spaces obtained from $C_{2\pi,l}^\infty$ by letting $b_k = 0$ and $a_k = 0$, respectively, it is clear that $C_{2\pi,l}^\infty$, $C_{2\pi,l,0}^\infty$, $C_{2\pi,l,1}^\infty$ are closed subspaces of $C_{2\pi}^\infty$ and that

$$C_{2\pi,l}^\infty = C_{2\pi,l,0}^\infty \oplus C_{2\pi,l,1}^\infty.$$

Further, given integers $m \geq 0$, $l \geq 1$, and a number $0 \leq \mu < 1$, let

$$C_{2\pi,l}^{m+\mu} = \text{the closure of } \ C_{2\pi,l}^\infty \ \text{ in } \ C_{2\pi}^{m+\mu}.$$

The spaces $C_{2\pi,l,i}^{m+\mu}$ for $i = 0$, 1 are defined similarly.

We now collect some properties of the above spaces, which we need in what follows.

LEMMA 5.1. *The following assertions hold:*

(1) $\frac{\partial}{\partial\theta} \in L(C_{2\pi,l,0}^{m+\mu}, C_{2\pi,l,1}^{m+\mu-1}) \cap L(C_{2\pi,l,1}^{m+\mu}, C_{2\pi,l,0}^{m+\mu-1})$.

(2) $\frac{\partial^2}{\partial\theta^2} \in L(C_{2\pi,l,0}^{m+\mu}, C_{2\pi,l,0}^{m+\mu-2}) \cap L(C_{2\pi,l,1}^{m+\mu}, C_{2\pi,l,1}^{m+\mu-2})$.

(3) *If $\eta_1, \eta_2 \in C_{2\pi,l,0}^{m+\mu}$, then $\eta_1\eta_2 \in C_{2\pi,l,0}^{m+\mu}$; if $\zeta_1, \zeta_2 \in C_{2\pi,l,1}^{m+\mu}$, then $\zeta_1\zeta_2 \in C_{2\pi,l,0}^{m+\mu}$; if $\eta \in C_{2\pi,l,0}^{m+\mu}$ and $\zeta \in C_{2\pi,l,1}^{m+\mu}$, then $\eta\zeta \in C_{2\pi,l,1}^{m+\mu}$.*

(4) *If $f \in C^\infty(\mathbb{R})$, then $f \circ \eta \in C_{2\pi,l,0}^{m+\mu}$ for $\eta \in C_{2\pi,l,0}^{m+\mu}$, and $f \circ \eta \in C_{2\pi,l}^{m+\mu}$ for $\eta \in C_{2\pi,l}^{m+\mu}$.*

Proof. The first three assertions are immediate. Assertion (4) follows from the Stone–Weierstrass theorem and (3). □

In the following, we write $\overline{B}_R$ for the closed ball in $\mathbb{R}^2$ centered at 0 with radius R. We often regard a given function u defined on $\overline{B}_R$ as a mapping from $[0, R]$ to the set of 2π-periodic functions on $\mathbb{R}$ by identifying u with the mapping $r \mapsto u(r\cos\theta, r\sin\theta)$. We further set

$$C^{m+\mu}_{2\pi,l}(\overline{B}_R) = \text{the closure of } C^\infty(\overline{B}_R) \cap C^\infty([0,R], C^\infty_{2\pi,l}(\mathbb{R})) \text{ in } C^{m+\mu}(\overline{B}_R).$$

The spaces $C^{m+\mu}_{2\pi,l,i}(\overline{B}_R)$ for $i = 0$, 1 are defined similarly.

Given $\eta \in C^{m+\mu}_{2\pi}$, the mean curvature of the surface $r = R + \eta(\theta)$ is given by

$$\kappa = \frac{2[\eta'(\theta)]^2 - [R+\eta(\theta)]\eta''(\theta) + [R+\eta(\theta)]^2}{\{[R+\eta(\theta)]^2 + [\eta'(\theta)]^2\}^{\frac{3}{2}}},$$

and the outward normal derivative can be expressed as

$$\frac{\partial}{\partial\nu} = \frac{R+\eta(\theta)}{\{[R+\eta(\theta)]^2 + [\eta'(\theta)]^2\}^{\frac{1}{2}}}\left\{\frac{\partial}{\partial r} - \frac{\eta'(\theta)}{[R+\eta(\theta)]^2}\frac{\partial}{\partial\theta}\right\}.$$

Finally, shrinking $\mathcal{M}$ we may assume that $\max|\eta| \le R/(C_0 - 1)$ for all $\eta \in \mathcal{M}$. Using these notations, we prove the following properties of the operators A and B appearing in (4.10).

LEMMA 5.2. *For any integers $m \ge 2$, $l \ge 1$, and any $0 \le \mu < 1$ we have*

$$A \in C^\infty(C^{m+\mu}_{2\pi,l,0} \cap \mathcal{M}, C^{m+\mu}_{2\pi,l,0}) \cap C^\infty(C^{m+\mu}_{2\pi,l} \cap \mathcal{M}, C^{m+\mu}_{2\pi,l}),$$

$$B \in C^\infty(C^{m+\mu}_{2\pi,l,0} \cap \mathcal{M}, C^{m+\mu-2}_{2\pi,l,0}) \cap C^\infty(C^{m+\mu}_{2\pi,l} \cap \mathcal{M}, C^{m+\mu-2}_{2\pi,l}).$$

Proof. The assertion for B is an immediate consequence of Lemma 5.1. In what follows we give the proof of the assertion for A.

We first prove that A maps $C^{m+\mu}_{2\pi,l,0} \cap \mathcal{M}$ into $C^{m+\mu}_{2\pi,l,0}$. Given $\eta \in C^{m+\mu}_{2\pi,l,0} \cap \mathcal{M}$, there exists a sequence $\{\eta_j\}_{j=1}^\infty \subseteq C^\infty_{2\pi,l,0} \cap \mathcal{M}$ such that $\eta_j \to \eta$ in $C^{m+\mu}_{2\pi}$. By the continuity of $A:\ \mathcal{M} \to C^{m+\mu}_{2\pi}$, we have $A(\eta_j) \to A(\eta)$ in $C^{m+\mu}_{2\pi}$. If we can prove that $A(\eta_j) \in C^\infty_{2\pi,l,0}$, then $A(\eta) \in C^{m+\mu}_{2\pi,l,0}$ by definition of $C^{m+\mu}_{2\pi,l,0}$. Hence it suffices to prove that if $\eta \in C^\infty_{2\pi,l,0} \cap \mathcal{M}$, then $A(\eta) \in C^\infty_{2\pi,l,0}$. Let $\eta \in C^\infty_{2\pi,l,0} \cap \mathcal{M}$ and set $u = A_1(\eta)$, i.e., u is the unique solution of (4.4) and (4.5). Since $\eta \in C^\infty_{2\pi}$, we have that $u \in C^\infty(\overline{B}_R) \subseteq C^\infty([0,R], C^\infty_{2\pi}(\mathbb{R}))$. Consequently, $u(r,\theta)$ has the Fourier expansion

$$u(r,\theta) = a_0(r) + \sum_{k=1}^\infty [a_k(r)\cos(k\theta) + b_k(r)\sin(k\theta)],$$

with coefficients satisfying the following conditions: $a_0, a_k, b_k \in C^\infty[0,R]$ $(k \ge 1)$, and for any integer $m \ge 0$ and any $N > 0$ there exists a constant $C_{m,N} > 0$ such that

$$|a_k^{(m)}(r)| + |b_k^{(m)}(r)| \le C_{m,N}(1+k)^{-N}, \quad r \in [0,R], \quad k = 1, 2, \ldots .$$

We now prove that all b_k's are zero, and if k is not proportional to l, then a_k is also zero. To this end we let $H^1_{2\pi}$ be the usual $H^1(\mathbb{R})$ Sobolev spaces of 2π-periodic functions on $\mathbb{R}$. We also write $H^1_{2\pi,l,0}$ and $H^1_{2\pi,l,0}$, respectively, for the closed subspaces of $H^1_{2\pi}$ consisting of all functions having the Fourier expansions

$$\zeta(\theta) = a_0 + \sum_{k=1}^\infty a_{kl}\cos(kl\theta) \qquad \text{and} \qquad \zeta(\theta) = \sum_{k=1}^\infty b_{kl}\sin(kl\theta),$$

respectively, and let $H^1_{2\pi,l} = H^1_{2\pi,l,0} \oplus H^1_{2\pi,l,1}$. Then, clearly,

$$H^1_{2\pi} = H^1_{2\pi,l,0} \oplus H^1_{2\pi,l,1} \oplus (H^1_{2\pi,l})^{\perp}.$$

Next, we introduce an analogous splitting of the usual Sobolev space $H^1(B_R)$. Let $H^1_{l,0}(B_R)$ and $H^1_{l,1}(B_R)$ be the closed subspaces of $H^1(B_R)$ defined by

$$H^1_{l,0}(B_R) = \left\{ w \in H^1(B_R) : \ w(r,\theta) = a_0(r) + \sum_{k=1}^{\infty} a_{kl}(r)\cos(kl\theta) \right\}$$

and

$$H^1_{l,1}(B_R) = \left\{ w \in H^1(B_R) : \ w(r,\theta) = \sum_{k=1}^{\infty} b_{kl}(r)\cos(kl\theta) \right\},$$

respectively, and set $H^1_l(B_R) = H^1_{l,0}(B_R) \oplus H^1_{l,1}(B_R)$. Then, clearly,

$$H^1(B_R) = H^1_{l,0}(B_R) \oplus H^1_{l,1}(B_R) \oplus (H^1_l(B_R))^{\perp}.$$

Given $\eta \in C^{\infty}_{2\pi,l,0} \cap \mathcal{M}$, let Ψ_η be the Hanzawa transformation introduced in section 4, and define

$$\Lambda_\eta = (\Psi_\eta)_* \circ \nabla \circ \Psi^*_\eta.$$

Further, let f_1 be an antiderivative of f, i.e., $f_1' = f$. Without loss of generality, we may assume that f is asymptotically linear at infinity, because the solution of (4.4) and (4.5) satisfies $0 < u \leq \bar{\sigma}$, so that changing values of f in the interval $(\bar{\sigma}, \infty)$ without changing its monotonicity and smoothness does not change the solution. We now consider the functional I on $H^1_{l,0}(B_R) \cap H^1_0(B_R)$ defined by

$$I(w) = \frac{1}{2}\int_{B_R} |\Lambda_\eta w(y)|^2 J_\eta(y)dy + \int_{B_R} f_1(\bar{\sigma} + w(y))J_\eta(y)dy.$$

Using a standard argument, we can easily prove that I has a local minimum in $H^1_{l,0}(B_R) \cap H^1_0(B_R)$, which we denote by u_0. Since both η and f are smooth, we actually have $u_0 \in C^\infty(\overline{B}_R) \cap C^\infty([0,R], C^\infty_{2\pi,l,0})$. In what follows we prove that $u = \bar{\sigma} + u_0$.

First, since u_0 is the minimum point of I, we have

$$0 = I'(u_0)w = \int_{B_R} \Lambda_\eta u_0(y) \cdot \Lambda_\eta w(y) J_\eta(y)dy + \int_{B_R} f(\bar{\sigma} + u_0(y))w(y)J_\eta(y)dy \tag{5.1}$$

for any $w \in H^1_{l,0}(B_R) \cap H^1_0(B_R)$. Next, we set $\rho = |y|$ and write $r = r(\rho, \eta)$ for the inverse function of

$$\rho = r - \eta\phi\left(\frac{r}{R+\eta}\right)$$

for any fixed $|\eta| < R/(C_0 - 1)$. Then a simple computation shows that

$$J_\eta(y) = \frac{r(\rho,\eta(\theta))}{\rho}\frac{\partial r(\rho,\eta(\theta))}{\partial\rho}. \tag{5.2}$$

This expression, Lemma 5.1, and $\eta \in C^\infty_{2\pi,l,0} \cap M(R, C_0)$ show that $J_\eta \in C^\infty(\overline{B}_R) \cap C^\infty([0,R], C^\infty_{2\pi,l,0})$. Note that $u_0 \in C^\infty(\overline{B}_R) \cap C^\infty([0,R], C^\infty_{2\pi,l,0})$ and Lemma 5.1 also imply that $f(\bar\sigma + u_0(y)) \in C^\infty(\overline{B}_R) \cap C^\infty([0,R], C^\infty_{2\pi,l,0})$. Hence we conclude that $f(\bar\sigma + u_0(y))J_\eta(y) \in C^\infty(\overline{B}_R) \cap C^\infty([0,R], C^\infty_{2\pi,l,0}) \subseteq H^1_{l,0}(B_R)$. Therefore, given any $v \in H^1_{l,1}(B_R) \oplus (H^1_l(B_R))^\perp = (H^1_{l,0}(B_R))^\perp$, we have

$$\int_{B_R} f(\bar\sigma + u_0(y))v(y)J_\eta(y)dy = 0. \tag{5.3}$$

Finally, denoting

$$a(\rho,\eta) = 1 - \frac{\eta}{R+\eta}\phi'\Big(\frac{r}{R+\eta}\Big) \quad \text{and} \quad b(\rho,\eta) = \frac{\eta}{(R+\eta)^2}\phi'\Big(\frac{r}{R+\eta}\Big) - \phi\Big(\frac{r}{R+\eta}\Big),$$

a simple computation shows that

$$\begin{aligned} \Lambda_\eta u_0(y)\cdot\Lambda_\eta v(y) = \quad & \left[a^2(\rho,\eta) + \frac{b^2(\rho,\eta)}{r^2(\rho,\eta)}(\eta')^2\right]\frac{\partial u_0}{\partial\rho}\frac{\partial v}{\partial\rho} \\ & + \frac{b(\rho,\eta)}{r^2(\rho,\eta)}\left[\frac{\partial u_0}{\partial\rho}\frac{\partial v}{\partial\theta}\eta' + \frac{\partial v}{\partial\rho}\frac{\partial u_0}{\partial\theta}\eta'\right] + \frac{1}{r^2(\rho,\eta)}\frac{\partial u_0}{\partial\theta}\frac{\partial v}{\partial\theta}. \end{aligned} \tag{5.4}$$

From (5.2), (5.4), and a similar argument as above, we can easily deduce that

$$\int_{B_R} \Lambda_\eta u_0(y)\cdot\Lambda_\eta v(y)J_\eta(y)dy = 0 \tag{5.5}$$

for any $v \in H^1_{l,1}(B_R) \oplus (H^1_l(B_R))^\perp = (H^1_{l,0}(B_R))^\perp$. Invoking (5.1), (5.3), and (5.5), we conclude that

$$\int_{B_R} \Lambda_\eta u_0(y)\cdot\Lambda_\eta v(y)J_\eta(y)dy + \int_{B_R} f(\bar\sigma + u_0(y))v(y)J_\eta(y)dy = 0$$

for any $v \in H^1_0(B_R)$. This shows that if we use the same formula to redefine I as a functional on the larger space $H^1_0(B_R)$, then u_0 is also a critical point of this functional; see (5.1). But it is well known that the Euler–Lagrange equations of the variational problem of this redefined functional I are given by (4.4) and (4.5). Hence, $\bar\sigma + u_0$ is a solution of (4.4) and (4.5). By uniqueness, we conclude that $u = \bar\sigma + u_0$, so that $u \in C^\infty(\overline{B}_R) \cap C^\infty([0,R], C^\infty_{2\pi,l,0})$.

Now, substituting $u = A_1(\eta)$ into (4.6) and using a similar argument, we deduce that $v = A_2(\eta) \in C^\infty(\overline{B}_R) \cap C^\infty([0,R], C^\infty_{2\pi,l,0})$. This readily implies that $A(\eta) \in C^\infty_{2\pi,l,0}$, as desired.

A similar argument shows that A also maps $C^{m+\mu}_{2\pi,l} \cap \mathcal{M}$ into $C^{m+\mu}_{2\pi,l}$, so that the proof is completed. ∎

Let $\mathcal{M}_{l,0} = C^{m+\mu}_{2\pi,l,0} \cap \mathcal{M}$. It is not difficult to see that $\mathcal{M}_{l,0}$ is a smooth submanifold of $C^{m+\mu}_{2\pi,l,0}$ of codimension 1. We denote by $A_{l,0}$ and $B_{l,0}$ the restrictions of A and B to $\mathcal{M}_{l,0}$ and study the abstract problem

$$A_{l,0}(\eta) + \tilde c = \gamma B_{l,0}(\eta). \tag{5.6}$$

Clearly, (5.6) is not equivalent to (4.10), but any solution of (5.6) is of course a solution of (4.10).

Given $\varepsilon > 0$, let

$$\mathcal{N}_{l,0,\varepsilon} = \{\overline{\eta} \in C^{m+\mu}_{2\pi,l,0} \;:\; \overline{\eta} = \eta + c,\, \Phi_{l,0}(\eta) = 0,\, \|\eta\|_{C^{m+\mu}_{2\pi,l,0}(\mathbb{R})} < \varepsilon,\, c \in \mathbb{R},\, |c| < \varepsilon\}.$$

Then it is not difficult to see that the mapping $(\eta, c) \mapsto \eta + c$ is a C^∞-diffeomorphism from $\mathcal{M}_{l,0} \times (-\varepsilon, \varepsilon)$ onto $\mathcal{N}_{l,0,\varepsilon}$.

We define mappings $\overline{A}(\cdot, \gamma) : \mathcal{N}_{l,0,\varepsilon} \to C^{m+\mu}_{2\pi,l,0}$ and $\overline{B} : \mathcal{N}_{l,0,\varepsilon} \to C^{m+\mu-2}_{2\pi,l,0}$ by

$$\overline{A}(\overline{\eta}, \gamma) = A_{l,0}(\eta) + c + c_\gamma - \frac{c\gamma}{R^2}, \qquad \overline{B}(\overline{\eta}) = B_{l,0}(\eta) - \frac{c}{R^2},$$

where c_γ is as in section 4 and $\bar{\eta} = \eta + c \in \mathcal{N}_{l,0,\varepsilon}$. These mappings allow a reduction of the system (5.6) to

$$\overline{A}(\overline{\eta}, \gamma) = \gamma \overline{B}(\overline{\eta}), \tag{5.7}$$

with $\tilde{c} = c + c_\gamma$. Observe that

$$\overline{B}'(0)\overline{\eta} = B'_{l,0}(0)\eta - \frac{c}{R^2} = -\frac{1}{R^2}(\eta'' + \eta) - \frac{c}{R^2} = -\frac{1}{R^2}(\overline{\eta}'' + \overline{\eta}).$$

Since $l \geq 2$ the operator $\overline{\eta} \to \overline{\eta}'' + \overline{\eta}$ is an isomorphism mapping $C^{m+\mu}_{2\pi,l,0}$ onto $C^{m-2+\mu}_{2\pi,l,0}$, so that $\overline{B}'(0)$ is also is an isomorphism as well. We now define the mapping $F : C^{m-2+\mu}_{2\pi,l,0} \times (0,\infty) \to C^{m-2+\mu}_{2\pi,l,0}$ by

$$F(\xi, \gamma) = \overline{A}(\overline{B}'(0)^{-1}\xi, \gamma) - \gamma \overline{B}(\overline{B}'(0)^{-1}\xi), \quad \xi \in C^{m-2+\mu}_{2\pi,l,0}, \;\; \gamma > 0.$$

Setting $\overline{\eta} = \overline{B}'(0)^{-1}\xi$, (5.7) is equivalent to

$$F(\xi, \gamma) = 0. \tag{5.8}$$

Clearly, $F \in C^\infty(C^{m-2+\mu}_{2\pi,l,0} \times (0,\infty), C^{m-2+\mu}_{2\pi,l,0})$ and

$$F(0, \gamma) = 0, \quad D_\xi F(0, \gamma)\xi = K(\gamma)\xi - \gamma\xi,$$

where

$$K(\gamma)\xi = D_{\overline{\eta}}\overline{A}(0, \gamma)\overline{B}'(0)^{-1}\xi, \quad \xi \in C^{m-2+\mu}_{2\pi,l,0}.$$

Obviously, the operator $K(\gamma) \colon C^{m-2+\mu}_{2\pi,l,0} \to C^{m-2+\mu}_{2\pi,l,0}$ is compact, so that $D_\xi F(0, \gamma)$ is a Fredholm operator of index zero on $C^{m-2+\mu}_{2\pi,l,0}$. We claim that

$$\dim \mathrm{Ker} D_\xi F(0, \gamma) = 1 \quad \text{if } \; \gamma = \gamma_l. \tag{5.9}$$

Indeed, using the variable transformation $\xi = \overline{B}'(0)\overline{\eta}$, we see that $D_\xi F(0, \gamma)\xi = 0$ is equivalent to $D_{\overline{\eta}}\overline{A}(0, \gamma)\overline{\eta} = \gamma \overline{B}'(0)\overline{\eta}$. Letting $\overline{\eta} = \eta + c$ as before, this equation is equivalent to

$$A'_{l,0}(0)\eta + c = \gamma B'_{l,0}(0)\eta. \tag{5.10}$$

It follows from the discussion in section 3, in particular from Lemma 3.2, that, if $\gamma = \gamma_l$ with $l \geq k^*$, then (5.10) has nontrivial solutions, given by $\overline{\eta} = C(\cos(l\theta) + c_l)$, with $C \in \mathbb{R}$ and where c_l is a real constant uniquely determined by l. This proves

(5.9). Moreover, letting $\xi_k := \overline{B}'(0)(\cos(k\theta) + c_k)$ for $k \in \mathbb{N}$, the deduction in section 3 also yields

$$D_\xi F(0,\gamma_l)\xi_k = \frac{(\gamma_k - \gamma_l)}{R^2} k^2 \cos(k\theta), \quad k \in \mathbb{N}.$$

Furthermore, a direct calculation shows that

$$D_\gamma D_\xi F(0,\gamma_l)\xi_l = \frac{\gamma_l}{R^2}(1 - l^2)\cos(l\theta).$$

Hence we conclude that

$$\mathbb{R} \cdot D_\gamma D_\xi F(0,\gamma_l)\xi_l \oplus \operatorname{im} D_\xi F(0,\gamma_l) = C^{m-2+\mu}_{2\pi,l,0}.$$

This observation implies the nondegeneration condition in the Crandall–Rabinowitz theorem on bifurcation from simple eigenvalues; cf. [4]. From the latter result we conclude that $(\xi,\gamma) = (0,\gamma_l)$ is a bifurcation point for (5.8). The corresponding bifurcation branch of solutions is of the form

$$\xi = \frac{\varepsilon(l^2 - 1)}{R^2}\cos(l\theta) + O(\varepsilon^2), \quad \gamma = \gamma_l + O(|\varepsilon|)$$

for a small real parameter ε. Returning to the original problem, we get the result stated in Theorem 1.1 in the case $n = 2$.

5.2. The case $n \geq 3$. Let us briefly indicate how the general n-dimensional case can be treated. In polar coordinates the Laplacian Δ on $\mathbb{R}^n$ is given by

$$\Delta u = \frac{1}{r^{n-1}}\frac{\partial}{\partial r}\Big(r^{n-1}\frac{\partial u}{\partial r}\Big) + \frac{1}{r^2}\Delta_\omega u,$$

where Δ_ω represents the Laplace–Beltrami operator on $\mathbb{S}^{n-1}$. In the following, we identify a point $\omega = (\omega_1, \omega_2, \ldots, \omega_{n-1}, \omega_n)$ on $\mathbb{S}^{n-1}$ with

$$(\theta_1, \theta_2, \ldots, \theta_{n-2}, \theta_{n-1}) \in [0,\pi] \times [0,\pi] \times \cdots [0,\pi] \times [0,2\pi)$$

via the spherical coordinates

$$\begin{cases} \omega_1 = \cos\theta_1, \\ \omega_2 = \sin\theta_1 \cos\theta_2, \\ \vdots \\ \omega_{n-1} = \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{n-2}\cos\theta_{n-1}, \\ \omega_n = \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{n-2}\sin\theta_{n-1}. \end{cases}$$

Let $\xi = (\xi_1, \xi_2, \ldots, \xi_{n-1})$ be the point on $\mathbb{S}^{n-2}$ related to ω, i.e.,

$$\omega_2 = \xi_1 \sin\theta_1, \;\; \omega_3 = \xi_2 \sin\theta_1, \ldots, \omega_n = \xi_{n-1}\sin\theta_1,$$

and denote by Δ_ξ the Laplace–Beltrami operator on $\mathbb{S}^{n-2}$. Then

$$\Delta_\omega = \frac{1}{\sin^{n-2}\theta_1}\frac{\partial}{\partial\theta_1}\Big(\sin^{n-2}\theta_1 \frac{\partial u}{\partial\theta_1}\Big) + \frac{1}{\sin^2\theta_1}\Delta_\xi u. \tag{5.11}$$

To reduce the notation we omit the subscript of θ_1 and simply write θ.

Given a closed surface in the form $r = R + \eta(\theta)$, where R is a positive constant and $\eta \in C^2[0,\pi]$ with $\eta'(0) = \eta'(\pi) = 0$, the mean curvature of this surface is given by

$$\kappa = \frac{1}{n-1}\left\{\frac{(n-1)[R+\eta(\theta)]^2 + n[\eta'(\theta)]^2 - [R+\eta(\theta)]\eta''(\theta)}{\{[R+\eta(\theta)]^2+[\eta'(\theta)]^2\}^{\frac{3}{2}}} - \frac{(n-2)\eta'(\theta)\cot\theta}{[R+\eta(\theta)]\{[R+\eta(\theta)]^2+[\eta'(\theta)]^2\}^{\frac{1}{2}}}\right\}.$$

Now let $Z_k^n(\omega)$ denote the zonal spherical harmonics of degree k with pole $e = (1,0,\dots,0)$, i.e.,

$$Z_k^n(\omega) = c_{nk}C_k^{\frac{n}{2}-1}(\omega\cdot e) = c_{nk}C_k^{\frac{n}{2}-1}(\cos\theta),$$

where C_k^λ represents the ultraspherical (or Gegenbauer) polynomial of degree k and index λ, i.e.,

$$C_0^\lambda(t) = 1, \quad C_k^\lambda(t) = \frac{1}{\Gamma(\lambda)}\sum_{j=0}^{\left[\frac{k}{2}\right]}(-1)^j\frac{\Gamma(\lambda+k-j)}{j!(k-2j)!}(2t)^{k-2j} \quad (k\ge 1),$$

and c_{nk} is the normalization factor. Using (6.18) and

$$(1-t^2)(C_k^\lambda(t))'' - (n-1)t(C_k^\lambda(t))' + k(k+n-2)C_k^\lambda(t) = 0,$$

one can easily verify that $Z_k^{(n)}(\theta)$ satisfies

$$\Delta_\omega Z_k^n(\omega) = -k(k+n-2)Z_k^n(\omega).$$

We next introduce

$$P_{zon}(\mathbb{S}^{n-1}) = \text{the span of } \{Z_k^n(\omega),\, k = 0,1,2,\dots\},$$

and for any nonnegative integer m and any $0\le\mu<1$ we define

$$C_{zon}^{m+\mu}(\mathbb{S}^{n-1}) = \text{the closure of } P_{zon}(\mathbb{S}^{n-1}) \text{ in } C^{m+\mu}(\mathbb{S}^{n-1}).$$

A significant difference between the cases $n\ge 3$ and $n = 2$ is that in the latter the closure of the linear space spanned by all $Z_{kl}^2(\theta) = \cos(kl\theta)$, $k = 0,1,2,\dots$, in $C^{m+\mu}(\mathbb{S}^1)$ is an algebra for any $l\ge 2$. In the case $n\ge 3$ this is not true for $l\ge 3$. However, we have the following result. We omit the details of the proof.

Lemma 5.3. *The following assertions hold:*

(1) *If* $\eta\in C_{zon}^{m+\mu}(\mathbb{S}^{n-1})$, *then* $\frac{\partial^2\eta}{\partial\theta^2}\in C_{zon}^{m-2+\mu}(\mathbb{S}^{n-1})$ *for* $m\ge 2$.
(2) *If* $\eta\in C_{zon}^{m+\mu}(\mathbb{S}^{n-1})$, *then* $\frac{\partial\eta}{\partial\theta}\cot\theta\in C_{zon}^{m-2+\mu}(\mathbb{S}^{n-1})$ *for* $m\ge 2$.
(3) *If* η_1, $\eta_2\in C_{zon}^{m+\mu}(\mathbb{S}^{n-1})$, *then* $\eta_1\eta_2\in C_{zon}^{m+\mu}(\mathbb{S}^{n-1})$.
(4) *If* η_1, $\eta_2\in C_{zon}^{m+\mu}(\mathbb{S}^{n-1})$, *then* $\frac{\partial\eta_1}{\partial\theta}\frac{\partial\eta_2}{\partial\theta}\in C_{zon}^{m-1+\mu}(\mathbb{S}^{n-1})$ *for* $m\ge 1$.
(5) *If* $\eta\in C_{zon}^{m+\mu}(\mathbb{S}^{n-1})$, *then* $f\circ\eta\in C_{zon}^{m+\mu}(\mathbb{S}^{n-1})$ *for any* $f\in C^\infty(R)$.

Based on Lemma 5.3 one can show the following analogue to Lemma 5.2.

Lemma 5.4. *Let* A, B *be the operators defined in section* 4. *For any integer* $m\ge 2$ *and any* $0\le\mu<1$, *we have the following assertions:*

$$A\in C^\infty(C_{zon}^{m+\mu}(\mathbb{S}^{n-1})\cap\mathcal{M}, C_{zon}^{m+\mu}(\mathbb{S}^{n-1})),$$

$$B \in C^\infty(C_{zon}^{m+\mu}(\mathbb{S}^{n-1}) \cap M(R, C_0), C_{zon}^{m+\mu-2}(\mathbb{S}^{n-1})).$$

Due to Lemma 5.4, we can now slightly modify the deduction in section 5.1 to prove that every γ_k with $k \geq k^*$ is a bifurcation point. Indeed, in the present case we have to replace the space $C_{2\pi,l,0}^{m+\mu}$ by the space $C_{zon}^{m+\mu}(\mathbb{S}^{n-1})$. Since for every $k \geq k^*$ the kernel in $C_{zon}^{m+\mu}(\mathbb{S}^{n-1}) \cap C_2^{m+\mu}(\mathbb{S}^{n-1})$ of the operator $(\eta, c) \mapsto A'(0)\eta + c - \gamma_k B'(0)\eta$ is given by $\mathbb{R} \cdot (\cos(k\theta), c_k)$, we get the result after a Lyapunov–Schmidt reduction. We omit the details.

Acknowledgments. We are grateful to the anonymous referees for helpful suggestions.

REFERENCES

[1] H. M. Byrne, *A weakly nonlinear analysis of a model of avascular solid tumor growth*, J. Math. Biol., 39 (1999), pp. 59–89.

[2] H. M. Byrne and M. A. J. Chaplain, *Growth of nonnecrotic tumors in the presence and absence of inhibitors*, Math. Biosci., 130 (1995), pp. 151–181.

[3] A. Borisovich and A. Friedman, *Symmetric-breaking bifurcation for free boundary problems*, Indiana Univ. Math. J., 54 (2005), pp. 927–947.

[4] M. G. Crandall and P. H. Rabinowitz, *Bifurcation from simple eigenvalues*, J. Functional Analysis, 8 (1971), pp. 321–340.

[5] S. Cui, *Analysis of a free boundary problem modelling tumor growth*, Acta Math. Sin. (Engl. Ser.), 21 (2005), pp. 1071–1082.

[6] S. Cui and A. Friedman, *A free boundary problem for a singular system of differential equations: An application to a model of tumor growth*, Trans. Amer. Math. Soc., 355 (2003), pp. 3537–3590.

[7] M. A. Fontelos and A. Friedman, *Symmetry-breaking bifurcations of free boundary problems in three dimensions*, Asymptot. Anal., 35 (2003), pp. 187–206.

[8] A. Friedman and F. Reitich, *Symmetry-breaking bifurcation of analytic solutions to free boundary problems*, Trans. Amer. Math. Soc., 353 (2001), pp. 1587–1634.

[9] A. Friedman and F. Reitich, *On the existence of spatially patterned dormant malignancies in a model for the growth of non-necrotic vascular tumors*, Math. Models Methods Appl. Sci., 11 (2001), pp. 601–625.

[10] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.

[11] H. P. Greenspan, *On the growth and stability of cell cultures and solid tumors*, J. Theoret. Biol., 56 (1976), pp. 229–242.

[12] J. P. Ward and J. R. King, *Mathematical modelling of avascular-tumor growth*, IMA J. Math. Appl. Med. Biol., 14 (1997), pp. 39–70.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 236–262

HOMOGENIZATION IN VISCOPLASTICITY*

SERGIY NESENENKO†

Abstract. In this work we present the justification of the formally derived homogenized problem for the quasi-static initial boundary value problem with internal variables, which models the deformation behavior of viscoplastic materials with a periodic microstructure.

Key words. homogenization, plasticity, viscoplasticity, two-scale convergence, maximal monotone operator, microstructure

AMS subject classifications. 74Q15, 74C05, 74C10, 74D10, 35J25, 34G20, 34G25, 47H04, 47H05

DOI. 10.1137/060655092

1. Introduction and statement of results. During the last few decades, the rigorous mathematical investigation of homogenization has brought appreciable success in determining the macroscopic behavior from the knowledge of the microstructure in many problems from different sciences. Among them are problems from the linear and nonlinear theory of elasticity, linear viscoelasticity and electrodynamics, hydrodynamics, and porous media; see, for example, [5], [7], [9], [10], [19], [22], [23], [24], [25], [28], [29], [31]. The only rigorous results of homogenization related to problems from the theory of plasticity or viscoplasticity known to the author are [4] and [17]. This is in contrast to the importance of homogenization in solid mechanics. This circumstance motivated the further study of these problems.

In this work I deal with the homogenization of the initial boundary value problem describing the deformation behavior of inelastic materials with a periodic microstructure, in particular for plastic and viscoplastic materials. The formulation of the problem is based on the assumption that only small strains occur: Let Ω be an open bounded set, the set of material points of the body, with C^1-boundary $\partial\Omega$. T_e denotes a positive number (time of existence) and for $0 \leq t \leq T_e$

$$\Omega_t = \Omega \times (0,t).$$

Let $\mathcal{S}^3$ denote the set of symmetric 3×3 matrices, and let $u(x,t) \in \mathbb{R}^3$ be the unknown displacement of the material point x at time t; $T(x,t) \in \mathcal{S}^3$ is the unknown Cauchy stress tensor, and $z(x,t) \in \mathbb{R}^N$ denotes the unknown vector of internal variables. The model equations of the problem (a microscopic problem) are

$$-\mathrm{div}_x T(x,t) = b(x,t), \tag{1}$$

$$T(x,t) = \mathcal{D}\left[\frac{x}{\eta}\right](\varepsilon(\nabla_x u(x,t)) - Bz(x,t)), \tag{2}$$

$$\begin{aligned} \frac{\partial}{\partial t} z(x,t) &\in g\left(\frac{x}{\eta}, -\nabla_z \psi\left(\frac{x}{\eta}, \varepsilon(\nabla_x u(x,t)), z(x,t)\right)\right) \\ &= g\left(\frac{x}{\eta}, B^T T(x,t) - L\left[\frac{x}{\eta}\right] z(x,t)\right), \end{aligned} \tag{3}$$

*Received by the editors March 24, 2006; accepted for publication (in revised form) December 1, 2006; published electronically May 22, 2007.
http://www.siam.org/journals/sima/39-1/65509.html
†Department of Mathematics, Darmstadt University of Technology, Schlossgartenstr. 7, 64289 Darmstadt, Germany (nesenenko@mathematik.tu-darmstadt.de).

which must hold for $x \in \Omega$ and $t \in [0, \infty)$. The initial value for $z(x,t)$ is taken in the form

$$z(x,0) = z^{(0)}\left(x, \frac{x}{\eta}\right), \tag{4}$$

which must hold for $x \in \Omega$. We consider the Dirichlet boundary condition

$$u(x,t) = \gamma(x,t), \tag{5}$$

which must be satisfied for $(x,t) \in \partial\Omega \times [0, \infty)$.

Here

$$\varepsilon(\nabla_x u(x,t)) = \frac{1}{2}(\nabla_x u(x,t) + (\nabla_x u(x,t))^T) \in \mathcal{S}^3$$

is the strain tensor, and $B : \mathbb{R}^N \to \mathcal{S}^3$ is a linear mapping, which assigns to the vector $z(x,t)$ the plastic strain tensor $\varepsilon_p(x,t) = Bz(x,t)$. For every $y \in \mathbb{R}^3$ we denote by $\mathcal{D}[y] : \mathcal{S}^3 \to \mathcal{S}^3$ a linear, symmetric, positive definite mapping, the elasticity tensor. The mapping $y \to \mathcal{D}[y]$ is assumed to be measurable and periodic with a periodicity cell $Y \subset \mathbb{R}^3$. Suppose that there exist two positive constants $0 < \alpha \leq \beta$ satisfying

$$\alpha|\xi|^2 \leq \mathcal{D}_{ijkl}[y]\xi_{kl}\xi_{ij} \leq \beta|\xi|^2 \quad \text{for any } \xi \in \mathcal{S}^3.$$

The function $b : \Omega \times [0, \infty) \to \mathbb{R}^3$ is the volume force, and $z^{(0)} : \Omega \times \mathbb{R}^3 \to \mathbb{R}^N$ is the initial value of the vector of internal variables, periodic in y with the same periodicity cell Y. The positive semidefinite quadratic form

$$\psi(y, \varepsilon, z) = \frac{1}{2}\mathcal{D}[y](\varepsilon - Bz) \cdot (\varepsilon - Bz) + \frac{1}{2}(L[y]z) \cdot z \tag{6}$$

represents the free energy (see [1, Appendix]), and for all $y \in \mathbb{R}^3$ the function $z \to g(y,z) : \mathbb{R}^N \to 2^{\mathbb{R}^N}$ is monotone satisfying $0 \in g(y,0)$; $y \to g(y,z)$ is periodic with the periodicity cell $Y \subset \mathbb{R}^3$. The positive semidefinite $N \times N$ matrix $L[y]$ is measurable and periodic with the same periodicity cell Y.

A function $g : D(g) \subseteq \mathbb{R}^N \mapsto 2^{\mathbb{R}^N}$ is called *monotone* if

$$(z_1 - z_2, y_1 - y_2) \geq 0$$

for any $y_i \in g(z_i)$ and any $z_i \in D(g),\ i = 1, 2$. A monotone function is said to be *maximal monotone* if it has no monotone extension. In other words, g is *maximal monotone* if and only if the inequality

$$(z_1 - z_2, y_1 - y_2) \geq 0 \quad \forall\, y_1 \in g(z_1)$$

implies $y_2 \in g(z_2)$.

The number $\eta > 0$ is the scaling parameter of the microstructure.

The differential inclusion (3) with the given function g and (2) together define the material behavior. They are the constitutive relations which model the inelastic response of the body, whereas (1) is the conservation law of linear momentum. The differential inclusion (3) is called a constitutive relation (or equation) of monotone type which was first introduced in [1]. The class of constitutive relations of monotone type naturally generalizes the class of constitutive relations of generalized standard materials defined by B. Halphen and N. Q. Son, because in the last case the function

g is the gradient or the subdifferential of a convex function. For examples of models from engineering, and for the study of whether they are of monotone type, we refer the reader to [1]. It must be said here that the classical models like the Prandtl–Reuss and the Norton–Hoff laws belong to the class of constitutive equations of monotone type. However, this class is still too small to include all models used in engineering. Namely, most models describing the deformation behavior of inelastic bodies with infinitesimal strains can be written in the form (3), but often the function g is not monotone.[1]

The existence and uniqueness theory for (1)–(3) is well understood under additional assumptions: If the free energy is not only positive semidefinite but positive definite (equivalently, if the $N \times N$ matrix L is positive definite), and, additionally, if the mapping $z \to g(y, z)$ is maximal monotone for all $y \in \mathbb{R}^3$, then the initial boundary value problem has a unique solution denoted by (u_η, T_η, z_η); see [3] or [1]. We want to point out that in many cases the free energy is not positive definite but positive semidefinite. For example, the Prandtl–Reuss and Norton–Hoff laws are constitutive equations with positive semidefinite free energy, whereas models with linear hardening have positive definite free energy. For problems with semidefinite free energy, to the author's knowledge there have been no general existence results until now. But for some models with a particular choice of the function g the existence and uniqueness theory is already available; see, for example, [15], [14], [13], [30], and the literature cited there concerning this side of the investigation.

To study the asymptotic behavior of (u_η, T_η, z_η) as η tends to 0 we postulate that this function is close to the function $(\hat{u}_\eta, \hat{T}_\eta, \hat{z}_\eta)$ defined by

$$\begin{aligned}
\hat{u}_\eta(x,t) &= u_0(x,t) + \eta u_1\left(x, \frac{x}{\eta}, t\right), \\
\hat{T}_\eta(x,t) &= T_0\left(x, \frac{x}{\eta}, t\right), \\
\hat{z}_\eta(x,t) &= z_0\left(x, \frac{x}{\eta}, t\right),
\end{aligned}$$

where (u_0, u_1, T_0, z_0) solves the homogenized initial boundary value problem (see [2]):

$$-\mathrm{div}_x T_\infty(x,t) = b(x,t), \tag{7}$$

$$T_\infty(x,t) = \frac{1}{|Y|}\int_Y T_0(x,y,t)dy, \tag{8}$$

$$-\mathrm{div}_y T_0(x,y,t) = 0, \tag{9}$$

$$\begin{aligned} T_0(x,y,t) = \mathcal{D}[y](\varepsilon(\nabla_y u_1(x,y,t)) &- Bz_0(x,y,t) \\ &+\varepsilon(\nabla_x u_0(x,t))), \end{aligned} \tag{10}$$

$$\frac{\partial}{\partial t} z_0(x,y,t) \in g(y, B^T T_0(x,y,t) - L[y] z_0(x,y,t)), \tag{11}$$

$$z_0(x,y,0) = z_0^{(0)}(x,y), \tag{12}$$

[1]To the author's knowledge all models proposed in engineering sciences belong to the problems of premonotone type. The problem (1)–(3) is of a premonotone type if the multifunction g in (3) satisfies the inequality

$$\forall\ z \in D(g),\ \forall\ z^* \in g(z) \quad z^* \cdot z \geq 0.$$

which must hold for $(x, y, t) \in \Omega \times Y \times [0, \infty)$,

$$u_0(x,t) = \gamma(x,t), \quad (x,t) \in \partial\Omega \times [0,\infty). \tag{13}$$

Note. For fixed x, (9)–(12) together with the periodicity assumption on $y \mapsto (u_1, T_0)(x, y, t)$, which can be considered to be a boundary condition, define an initial boundary problem, the cell problem, in $Y \times [0, \infty)$. u_0, u_1 can be interpreted as macro- and microdisplacement, T_0 as a microstress; the macrostress T_∞ is obtained by averaging of T_0 over the representative volume element. In Theorem 2 in [4] it was shown that under some additional assumptions on the function g the homogenized initial boundary value problem (7)–(13) has a unique solution,

$$(u_0, u_1, T_\infty) \in L^2(0, T_e; H^1(\Omega, \mathbb{R}^3)) \times L^2(\Omega_{T_e}, H^1(Y, \mathbb{R}^3)) \times L^2(\Omega_{T_e}, \mathcal{S}^3)$$

$$(T_0, z_0) \in L^2((\Omega \times Y)_{T_e}, \mathcal{S}^3) \times C([0, T_e]; L^2(\Omega \times Y, \mathbb{R}^N)).$$

The main goal of this work is to prove that the solution of the microscopic problem (u_η, T_η, z_η) has as the asymptotics the function $(\hat{u}_\eta, \hat{T}_\eta, \hat{z}_\eta)$. The justification uses methods from the established homogenization theory for linear problems, but several difficulties arise not present in this theory.

First, the existence and uniqueness theory for the homogenized initial boundary value problem as well as the justification procedure are more complicated due to the impossibility of decoupling the homogenized problem (the first equation) and the so-called cell problem (the last three equations with the periodicity assumption, which can be considered as a boundary condition), unlike in linear elasticity, where the homogenized and the cell problem can be decoupled. This difficulty was successfully solved in [4].

Second, difficulties arise which are based on the fact that the solution of the homogenized problem is of low regularity because of the nonlinearity of the constitutive equation. One of the difficulties resulting from the low regularity of the solution of the homogenized problem is that the mapping $x \mapsto (T_0, z_0)(x, x/\eta, t)$ is not well defined because $x \mapsto (x, x/\eta)$ maps Ω onto a three-dimensional subspace of a six-dimensional space $\Omega \times \mathbb{R}^3$.

Indeed, $x \mapsto (x, x/\eta)$ maps Ω onto a three-dimensional subspace of $\Omega \times \mathbb{R}^3$ (see Figure 1) and by virtue of Theorem 5.2.2 in [12] the mapping $x \mapsto (T_0, z_0)(x, x/\eta, t)$ is not well defined. In other words, the function $(T_0, z_0)(t) \in L^2(\Omega \times Y, \mathcal{S}^3) \times L^2(\Omega \times Y, \mathbb{R}^N)$ has no trace on a three-dimensional subspace of a six-dimensional space.

In [4] this difficulty is overcome by imposing higher regularity on the given data in such a way that the solution of the homogenized problem becomes smoother and the existence of the trace for $(T_0(t), z_0(t))$ on the three-dimensional subspace is an easy consequence of this obtained smoothness. The higher regularity of the solution of the homogenized problem plays an essential role in that work also at another place: in order to apply the energy method of Tartar[2] (see [27, 34]) in the justification, the author needs $\partial_t \mathrm{div}_x T_0(x, y, t) \mid_{y=x/\eta}$ and $\partial_t \mathrm{rot}_x \nabla_y u_1(x, y, t) \mid_{y=x/\eta}$ to belong to a compact subset of H^{-1}_{loc}. This is provided by the smoothness of $(T_0, u_1)(x, y, t)$. Unfortunately, one cannot expect that the solution is of this higher regularity globally in time. Instead, after a certain finite time the solution is only of lower regularity. Therefore in [4] the justification of the homogenized problem is only possible locally

[2]It is also called the *oscillating test function method*.

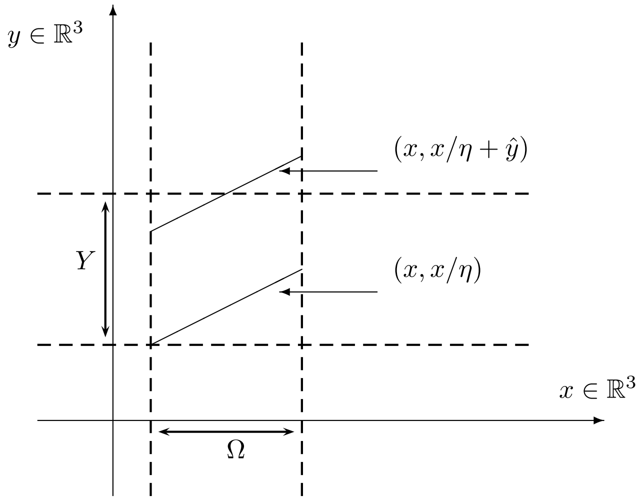


FIG. 1.

in time. In contrast, here we can justify the homogenized problem globally in time without imposing additional smoothness on the data.

To avoid the difficulty with the trace I, follow an idea,[3] proposed in [2], of introducing an additional fast variable y, which is plugged into (1)–(5):

$$-\mathrm{div}_x T(x,y,t) = b(x,t), \tag{14}$$

$$T(x,y,t) = \mathcal{D}\left[\frac{x}{\eta}+y\right](\varepsilon(\nabla_x u(x,y,t)) - Bz(x,y,t)), \tag{15}$$

$$\frac{\partial}{\partial t} z(x,y,t) \in g\left(\frac{x}{\eta}+y, B^T T(x,y,t) - L\left[\frac{x}{\eta}+y\right] z(x,y,t)\right), \tag{16}$$

$$z(x,y,0) = z^{(0)}\left(x, \frac{x}{\eta}+y\right), \tag{17}$$

which hold for $(x,y) \in \Omega \times Y$ and $t \in [0,\infty)$, and of the Dirichlet boundary condition

$$u(x,y,t) = \gamma(x,t), \tag{18}$$

which holds for $(x,y) \in \partial\Omega \times Y$ and $t \in [0,\infty)$. The function $(u_\eta, T_\eta, z_\eta)(x,y,t)$ is periodic in y.

We give the definition of a solution of the initial boundary value problem (14)–(18). $\eta > 0$ is fixed.

DEFINITION 1.1. *Let*

$$(u_\eta, T_\eta, z_\eta) : \Omega \times \mathbb{R}^3 \times \mathbb{R}^+ \mapsto \mathbb{R}^3 \times \mathcal{S}^3 \times \mathbb{R}^N$$

be a function which satisfies the initial condition (17) *for a.e.* $(x,y) \in \Omega \times \mathbb{R}^3$ *and for which the function* $(x,y) \mapsto (u_\eta, T_\eta, z_\eta)(x,y,t)$ *is a solution of* (14)–(18) *for almost*

[3]The idea of considering the family of shifted problems was also used in [16] to show that for some linear and nonlinear problems the averaging over the shifting y eliminates the rapid oscillations in the solution. For details we refer the reader to this work.

all $y \in \mathbb{R}^3$. *Then* (u_η, T_η, z_η) *is called a* family of solutions *of the initial boundary value problem* (14)–(18) *depending on the fast variable* y.

We assume now that for all $0 < \eta < \eta_0$ such a solution family (u_η, T_η, z_η) of the initial boundary value problem depending on the fast variable y exists and is close to

$$\hat{u}_\eta(x, y, t) = u_0(x, t) + \eta u_1 \left(x, \frac{x}{\eta} + y, t \right), \tag{19}$$

$$\hat{T}_\eta(x, y, t) = T_0 \left(x, \frac{x}{\eta} + y, t \right), \tag{20}$$

$$\hat{z}_\eta(x, y, t) = z_0 \left(x, \frac{x}{\eta} + y, t \right). \tag{21}$$

The functions $u_0(x,t), u_1(x,y,t), T_0(x,y,t), z_0(x,y,t)$, which are assumed to be periodic with respect to the y-argument with a periodicity cell $Y \subset \mathbb{R}^3$, solve the problem (7)–(13).

Notation. Banach spaces $W^{m,p}(\Omega, \mathbb{R}^N)$ are endowed with the norm $\| \cdot \|_{m,p,\Omega}$. $H^m(\Omega, \mathbb{R}^N) = W^{m,2}(\Omega, \mathbb{R}^N)$ are Hilbert spaces with the usual scalar product on them and the norm $\| \cdot \|_{m,\Omega}$, $\| \cdot \|_\Omega = \| \cdot \|_{0,\Omega}$.

Define the space

$$W(Y) = \left\{ v \in H^1(Y) \mid \frac{1}{|Y|} \int_Y v(y) dy = 0 \right\},$$

which becomes a Banach space due to the Poincaré–Wirtinger inequality (see [18, Proposition 3.38]) for the norm $\|u\|_{W(Y)} = \|\nabla u\|_Y$.

The symbols $\mathcal{D}_\eta$ and L_η denote the mappings $\mathcal{D}[x/\eta + y]$ and $L[x/\eta + y]$, respectively, i.e., $\mathcal{D}_\eta := \mathcal{D}[x/\eta + y]$, $L_\eta := L[x/\eta + y]$.

Now we can formulate the main result of this work.

THEOREM 1.2. *Let* $T_e > 0$. *Assume that the matrix* $L \in L^\infty(Y, \mathbb{R}^{N \times N})$ *in* (6) *is uniformly positive definite and that the mapping* $g : \mathbb{R}^3 \times \mathbb{R}^N \to 2^{\mathbb{R}^N}$ *satisfies the following three conditions:*

- $0 \in g(y, 0)$ *for almost all* $y \in Y$,
- $z \mapsto g(y, z)$ *is maximal monotone for almost all* $y \in Y$,
- *the mapping* $y \mapsto j_\lambda(y, z) : \mathbb{R}^3 \to \mathbb{R}^N$ *is measurable for all* $\lambda > 0$, *where* $j_\lambda(y, z)$ *is the inverse of* $z \mapsto z + \lambda g(y, z)$.[4]

Suppose that $b \in W^{2,1}(0, T_e; L^2(\Omega, \mathbb{R}^3))$ *and* $\gamma \in W^{2,1}(0, T_e; H^1(\Omega, \mathbb{R}^3))$. *Assume that* $z^{(0)} \in L^2(\Omega \times Y, \mathbb{R}^N)$ *and there exists* $\zeta \in L^2(\Omega \times Y, \mathbb{R}^N)$ *such that*

$$\zeta(x, y) \in g(y, B^T T^{(0)}(x, y) - L[y] z^{(0)}(x, y)) \quad \text{a.e. in } \Omega \times Y,$$

where $(u^{(0)}, T^{(0)})$ *is a weak solution of the problem of linear elasticity theory* (25)–(27) *to the data* $\hat{b} = b(0)$, $\hat{\varepsilon}_p = B z^{(0)}$, $\hat{\gamma} = \gamma(0)$.

Assume further that there exists a positive function $h \in L^2(\Omega \times Y, \mathbb{R}^N)$ *such that the inequality*

$$| \, g_\lambda(x/\eta + y, B^T T^{(0)} - L[x/\eta + y] z^{(0)})| \le C h(x, y) \tag{22}$$

[4]The mapping $z \mapsto j_\lambda(y, z)$ is single valued and well defined, since $z \mapsto g(y, z)$ is assumed to be maximal monotone.

with a constant $C = C(\lambda)$ independent of η holds, where g_λ is the Yosida approximation of g.

Then the solution (u_η, T_η, z_η) of the microscopic problem (14)–(18) with parameter y satisfies for all $0 \le t \le T_e$,

$$\lim_{\eta\to 0}(\|u_0(t) - u_\eta(t)\|_{\Omega\times Y} + \|\hat{T}_\eta(t) - T_\eta(t)\|_{\Omega\times Y} + \|\hat{z}_\eta(t) - z_\eta(t)\|_{\Omega\times Y}) = 0. \tag{23}$$

Note. We note that the inequality (22) is easily satisfied in practice. Let us take, for example, the model of Melan–Prager with the Prandtl–Reuss flow rule, i.e.,

$$g(z) = \partial I_K\left(T - k\left(\frac{\cdot}{\eta}\right)z\right),$$

where $k : \mathbb{R}^3 \to \mathbb{R}^+$ is a material function with $k(y) \ge C > 0$, and I_K is the indicator function of a closed convex set $K = \{\sigma \in \mathcal{S}^3 |\ \sigma^D \cdot \sigma^D \le c\}$ specified by the von Mises yield criterion. Here σ^D is the deviator of σ and $c > 0$ is a given constant. We assume that c is independent of η. Then the Yosida approximation of g is (see [11, p. 46])

$$g_\lambda(z) = \frac{1}{\lambda}\left(T - k\left(\frac{\cdot}{\eta}\right)z - \mathrm{proj}_K\left(T - k\left(\frac{\cdot}{\eta}\right)z\right)\right).$$

If $k \in L^\infty(Y, \mathbb{R}^+)$, then (22) is satisfied for $h = |T + \|k\|_{L^\infty(Y,\mathbb{R}^+)}z| + \|\mathrm{proj}_K(T - k(\frac{\cdot}{\eta})z)\|_{L^\infty(\Omega\times Y,\mathcal{S}^3)}$. Note that (22) is fulfilled for $k(y) \ge 0$ as well, but in this case the existence theory fails. In a similar way one gets that the inequality (22) holds for the model of Melan–Prager with the Norton–Hoff flow rule.

Note. For future use we need an estimate obtained in Theorem 2 in [4] for the time derivative of z_0. Define a function $h = -(B^T\mathcal{D}QB + L)z_0 + B^T\sigma_0$, where the operator Q is a projector in $L^2(\Omega\times Y, \mathcal{S}^3)$, and the function σ_0 solves the usual linear elasticity problem. Then the function h satisfies the inequality

$$\left\|\frac{\partial}{\partial t}h(t)\right\|_{\Omega\times Y} \le |Ch(0)| + ||B^T\sigma_{0,t}(0)||_{\Omega\times Y} + \int_0^t ||B^T\sigma_{0,tt}(s)||_{\Omega\times Y}ds, \tag{24}$$

with $|C\zeta| = \inf\{\|(B^T\mathcal{D}QB + L)\xi\|_{\Omega\times Y} \mid \xi(x, y) \in g(y, \zeta(x, y)) \text{ a.e.}\}$.
See [4] for more details.

2. Justification of the homogenized model.

2.1. Preliminaries. In this section we deal with a boundary value problem, a linear problem of elasticity theory with a parameter y, formed by the following equations:

$$-\mathrm{div}_x T(x, y) = \hat{b}(x), \tag{25}$$

$$T(x, y) = \mathcal{D}\left[\frac{x}{\eta} + y\right](\varepsilon(\nabla_x u(x, y)) - \hat{\varepsilon}_p(x, y)), \tag{26}$$

$$u(x, y) = \hat{\gamma}(x), \qquad x \in \partial\Omega. \tag{27}$$

The solution of this problem is understood in the following sense: A function $(u, T) \in L^2(Y, H^1(\Omega, \mathbb{R}^3)) \times L^2(\Omega\times Y, \mathcal{S}^3)$ is a solution of (25)–(27), if (26) is satisfied, and if for $\hat{b} \in L^2(\Omega, \mathbb{R}^3)$, $\hat{\gamma} \in H^1(\Omega, \mathbb{R}^3)$, $\hat{\varepsilon}_p \in L^2(\Omega\times Y, \mathcal{S}^3)$ and for a.e. $y \in Y$, the following identity

$$\left(\mathcal{D}\left[\frac{\cdot}{\eta} + y\right](\varepsilon(\nabla_x u(\cdot, y)) - \hat{\varepsilon}_p(\cdot, y)), \varepsilon(\nabla_x v(\cdot))\right)_\Omega = (\hat{b}, v)_\Omega \tag{28}$$

holds for all $v \in H_0^1(\Omega, \mathbb{R}^3)$, and if u can be represented in the form $u = \hat{\gamma} + w$ with $w \in L^2(Y, H_0^1(\Omega, \mathbb{R}^3))$.

By the well-known theory for elliptic boundary value problems (see [35]) one gets immediately that to $\hat{b} \in L^2(\Omega, \mathbb{R}^3)$, $\hat{\gamma} \in H^1(\Omega, \mathbb{R}^3)$, $\hat{\varepsilon}_p \in L^2(\Omega \times Y, \mathcal{S}^3)$ and for a fixed $\eta > 0$ there is a unique weak solution (u, T) satisfying

$$\|u\|_{L^2(Y,H^1(\Omega,\mathbb{R}^3))} + \|T\|_{\Omega\times Y} \leq C(\|\hat{b}\|_\Omega + \|\hat{\varepsilon}_p\|_{\Omega\times Y} + \|\hat{\gamma}\|_{1,\Omega}), \tag{29}$$

with a constant C independent of η.

In the following we need a special projection operator.

DEFINITION 2.1. *For every $\hat{\varepsilon}_p \in L^2(\Omega \times Y, \mathcal{S}^3)$ a linear operator $P_\eta : L^2(\Omega \times Y, \mathcal{S}^3) \mapsto L^2(\Omega \times Y, \mathcal{S}^3)$ is defined by*

$$P_\eta \hat{\varepsilon}_p = \varepsilon(\nabla_x u),$$

where (u, T) is a unique solution of (25)–(27) *to $\hat{b} = \hat{\gamma} = 0$. Furthermore, we define a linear operator $Q_\eta = I - P_\eta$. Here I is an identity operator.*

It is immediately seen from the estimate (29) that the operator P_η is uniformly bounded. Other properties of P_η and Q_η are delivered by the following lemma.

LEMMA 2.2. (i) *The operators P_η and Q_η are orthogonal projectors with respect to the scalar product $[\xi, \zeta]_{\Omega\times Y} = (\mathcal{D}_\eta \xi, \zeta)_{\Omega\times Y}$ on $L^2(\Omega \times Y, \mathcal{S}^3)$.*

(ii) *The operator $B^T \mathcal{D}_\eta Q_\eta B : L^2(\Omega \times Y, \mathbb{R}^N) \mapsto L^2(\Omega \times Y, \mathbb{R}^N)$ is selfajoint and nonnegative with respect to the scalar product $(\xi, \zeta)_{\Omega\times Y}$. Moreover, there exists a constant $C > 0$ such that*

$$\|B^T \mathcal{D}_\eta Q_\eta B \xi\|_{\Omega\times Y} \leq C\|\xi\|_{\Omega\times Y} \tag{30}$$

for all $\eta > 0$.

Proof. See Lemma 2.5 in [3]. ∎

Since L_η is uniformly positive definite, it follows from Lemma 2.2 that the operator $L_\eta + B^T \mathcal{D}_\eta Q_\eta B$ is uniformly positive definite and bounded. This implies that

$$\langle \xi, \zeta \rangle_{\Omega\times Y,\eta} = ((L_\eta + B^T \mathcal{D}_\eta Q_\eta B)\xi, \zeta)_{\Omega\times Y}$$

defines a scalar product on $L^2(\Omega\times Y, \mathbb{R}^N)$. Furthermore, the associated norm $||\xi||_{\Omega\times Y,\eta} = \langle \xi, \xi \rangle_{\Omega\times Y,\eta}^{1/2}$ is equivalent to the norm $|| \cdot ||_{\Omega\times Y}$.

2.2. Reduction to an evolution equation. The preparations made in the previous section enable us to reduce the initial boundary value problem (14)–(18) to an evolution equation with a monotone operator.

Note that (15) yields

$$B^T T_\eta - L_\eta z_\eta = B^T \mathcal{D}_\eta(\varepsilon(\nabla_x u_\eta) - B z_\eta) - L_\eta z_\eta. \tag{31}$$

Let (u_η, T_η, z_η) be a solution of the initial boundary value problem (14)–(18). Now we fix t. If $z(t)$ is known, then (14), (15), (18) is a boundary value problem for the components $u_\eta(t), T_\eta(t)$ of the solution, the problem from the linear elasticity theory with a parameter y. Due to linearity of these problems the functions are obtained in the form

$$(u_\eta(t), T_\eta(t)) = (\tilde{u}_\eta(t), \tilde{T}_\eta(t)) + (v_\eta(t), \sigma_\eta(t)),$$

with a solution $(v_\eta(t), \sigma_\eta(t))$ of the Dirichlet boundary value problem (25)–(27) to the data $\hat{b} = b(t)$, $\hat{\gamma} = \gamma(t)$, $\hat{\varepsilon}_p = 0$, and with a solution $(\tilde{u}_\eta(t), \tilde{T}_\eta(t))$ of the problem (25)–(27) to the data $\hat{b} = \hat{\gamma} = 0$, $\hat{\varepsilon}_p = Bz_\eta(t)$. Thus one obtains

$$\varepsilon((\nabla_x u_\eta)(t)) - Bz_\eta(t) = (P_\eta - I)Bz_\eta(t) + \varepsilon((\nabla_x v_\eta)(t)).$$

We insert this equation into (31) and obtain that (16) can be rewritten in the following form:

$$\frac{\partial}{\partial t} z_\eta(t) \in G_\eta(-(B^T \mathcal{D}_\eta Q_\eta B + L_\eta) z_\eta(t) + B^T \sigma_\eta(t)), \tag{32}$$

with the mapping $G_\eta : L^2(\Omega \times Y, \mathbb{R}^N) \mapsto 2^{L^2(\Omega\times Y,\mathbb{R}^N)}$ defined by

$$G_\eta(\xi) = \{\zeta \in L^2(\Omega \times Y, \mathbb{R}^N) | \zeta(x, y) \in g(x/\eta + y, \xi(x, y)) \text{ a.e.}\}.$$

The function σ_η can be determined from the boundary value problem (25)–(27) to the given data b, γ, $\hat{\varepsilon}_p = 0$ and can be considered as known.

The evolution equation (32) for z_η can be rewritten as a nonautonomous evolution equation in $L^2(\Omega \times Y, \mathbb{R}^N)$

$$\frac{\partial}{\partial t} z_\eta(t) + A_\eta(t) z_\eta(t) \ni 0, \tag{33}$$

with the operator

$$A_\eta(t)z(t) = -G_\eta(-(B^T \mathcal{D}_\eta Q_\eta B + L_\eta) z_\eta(t) + B^T \sigma_\eta(t)).$$

It turns out that the operator $A_\eta(t)$ is maximal monotone as the next lemma shows.

Lemma 2.3. *Operator $A_\eta(t)$ is maximal monotone on $L^2(\Omega\times Y, \mathbb{R}^N)$ with respect to the scalar product $\langle \xi, \zeta \rangle_{\Omega\times Y,\eta}$.*

Proof. Set for simplicity $M_\eta = B^T \mathcal{D}_\eta Q_\eta B + L_\eta$.

Monotonicity of $A_\eta(t)$ for all t and η with respect to the scalar product $\langle \xi, \zeta \rangle_{\Omega\times Y,\eta}$ is shown in Theorem 3.3 in [3].

Now we prove that the mapping $G_\eta : L^2(\Omega \times Y, \mathbb{R}^N) \mapsto 2^{L^2(\Omega\times Y,\mathbb{R}^N)}$, defined through the maximal monotone function $g : Y \times \mathbb{R}^N \mapsto 2^{\mathbb{R}^N}$ with $g(y, 0) \ni 0$, is maximal monotone with respect to the scalar product $(\xi, \zeta)_{\Omega\times Y}$.

It is well known that G_η is maximal monotone if and only if $I + G_\eta$ is surjective. To show the surjectivity, we must prove that for every $q \in L^2(\Omega\times Y, \mathbb{R}^N)$ the equation

$$q \in z + G_\eta z \tag{34}$$

has a solution $z \in L^2(\Omega \times Y, \mathbb{R}^N)$. Since g is maximal monotone, for a.e. (x, y) the mapping $(\cdot \ + g(x/\eta + y, \cdot)) : \mathbb{R}^N \mapsto 2^{\mathbb{R}^N}$ has an inverse $j(x/\eta + y, \cdot) : \mathbb{R}^N \mapsto \mathbb{R}^N$, which satisfies for a.e. (x, y) the inequality $|j(x/\eta + y, \xi) - j(x/\eta + y, \zeta)| \leq |\xi - \zeta|$ for all $\xi, \zeta \in \mathbb{R}^N$. This Lipschitz continuity, together with the measurability of j with respect to the first argument, yields that the function $j(x/\eta+y, q)$ is of Caratheodory type. Thus one can prove that the mapping $(x, y) \mapsto j(x/\eta+y, q(x, y))$ is measurable. From $g(\cdot, 0) \ni 0$ it follows that $j(x/\eta + y, 0) = 0$, whence

$$|j(x/\eta + y, \xi)| = |j(x/\eta + y, \xi) - j(x/\eta + y, 0)| \leq |\xi|. \tag{35}$$

For $q \in L^2(\Omega \times Y, \mathbb{R}^N)$ we define $z(x, y) = j(x/\eta + y, q(x, y))$ for all $(x, y) \in \Omega \times Y$. Obviously, such defined z solves (34) if $z \in L^2(\Omega \times Y, \mathbb{R}^N)$. Yet, (35) yields that indeed $z \in L^2(\Omega \times Y, \mathbb{R}^N)$ and therefore we conclude that $I + G_\eta$ is surjective. Hence G_η is maximal monotone.

With this result it is easy to prove that $A_\eta(t)$ is maximal monotone for all t. In other words we have to show that for all $[z, \xi] \in L^2(\Omega \times Y, \mathbb{R}^N) \times L^2(\Omega \times Y, \mathbb{R}^N)$ and all $[\hat{z}, \hat{\xi}] \in GrA_\eta(t)$ such that

$$\langle z - \hat{z}, \xi - \hat{\xi} \rangle_{\Omega \times Y, \eta} \geq 0,$$

it follows that $[z, \xi] \in GrA_\eta(t)$.

Indeed,

$$\langle z - \hat{z}, \xi - \hat{\xi} \rangle_{\Omega \times Y, \eta} = ((-M_\eta z + B^T \sigma) - (-M_\eta \hat{z} + B^T \sigma), (-\xi) + \hat{\xi})_{\Omega \times Y} \geq 0.$$

Since G_η is maximal monotone, $[-M_\eta z + B^T \sigma, -\xi] \in GrG_\eta$, which means that $[z, \xi] \in GrA_\eta(t)$. $\square$

2.3. Beginning of the justification. Now we can prove the main result of this work, Theorem 1.2.

The approximate solution $(u_0, \hat{T}_\eta, \hat{z}_\eta)(x, y, t)$, determined from the homogenized problem, solves the same initial boundary value problem as the exact solution, however, with a special right-hand side. Observing (7)–(13) we get by a simple computation that $(u_0, \hat{T}_\eta, \hat{z}_\eta)$ satisfies the equations (a special microscopic problem)

$$-\operatorname{div}_x \hat{T}_\eta = -\operatorname{div}_x T_0(x, \xi, t)|_{\xi = \frac{x}{\eta} + y}, \tag{36}$$

$$\hat{T}_\eta = \mathcal{D}\left[\frac{x}{\eta} + y\right](\varepsilon(\nabla_x u_0) - B\hat{z}_\eta + \varepsilon(\nabla_\xi u_1(x, \xi, t))|_{\xi = \frac{x}{\eta} + y}), \tag{37}$$

$$\frac{\partial}{\partial t}\hat{z}_\eta \in g\left(\frac{x}{\eta} + y, B^T \hat{T}_\eta - L\left[\frac{x}{\eta} + y\right]\hat{z}_\eta\right), \tag{38}$$

$$\hat{z}_\eta(x, y, 0) = z_0^{(0)}\left(x, \frac{x}{\eta} + y\right), \qquad (x, y) \in \Omega \times Y, \tag{39}$$

$$u_0(x, t) = \gamma(x, t) \qquad (x, t) \in \partial\Omega \times [0, \infty). \tag{40}$$

Since these equations have the same structure as the equations of the microscopic problem with a parameter y, we can again employ the procedure from the last section and obtain that if $(\hat{v}_\eta(t), \hat{\sigma}_\eta(t))$ is the solution of the linear boundary value problem (25)–(27) to the data

$$\hat{b}(x) = -\operatorname{div}_x T_0(x, \xi, t)|_{\xi = x/\eta + y}, \tag{41}$$

$$\hat{\varepsilon}_p(x) = -\varepsilon(\nabla_\xi u_1(x, \xi, t))|_{\xi = x/\eta + y}, \tag{42}$$

$$\hat{\gamma}(x) = \gamma(x, t), \tag{43}$$

then the special microscopic problem is equivalent to a nonautonomous evolution equation

$$\frac{\partial}{\partial t}\hat{z}_\eta(t) + \hat{A}_\eta(t)\hat{z}_\eta(t) \ni 0, \tag{44}$$

where

$$\hat{A}_\eta(t)v = -g\left(\frac{x}{\eta} + y, -(B^T \mathcal{D}_\eta Q_\eta B + L_\eta)v + B^T \hat{\sigma}_\eta\right).$$

It turns out that the operator $\hat{A}_\eta(t)$ is maximal monotone, as we will see in the following lemma.

Lemma 2.4. *The operator $\hat{A}_\eta(t)$ is maximal monotone on $L^2(\Omega \times Y, \mathbb{R}^N)$ with respect to the scalar product $\langle \xi, \xi \rangle_{\Omega \times Y, \eta}$.*

Proof. The proof is the same as for $A_\eta(t)$ (Lemma 2.3). □

We are going to use the results of Lemmas 2.3 and 2.4 and a special distance between two maximal monotone operators to estimate the difference of solutions of the evolution inclusions (33), (44) by the norm of a function, which solves a linear elasticity problem to a special data. This is a crucial step in the justification of the homogenized model because we are able to employ classical results from homogenization theory for linear problems. In the next section this estimate is given.

2.4. Main estimate. We define a special distance between two maximal monotone operators due to [35]. With its help we obtain an important estimate in the justification procedure of the homogenized model.

Let H be a Hilbert space.

Definition 2.5. *The distance between two maximal monotone operators on H is defined as*

$$\operatorname{dis}(A_1, A_2) = \sup \left\{ \frac{(y_1 - y_2, x_2 - x_1)}{\|y_1\| + \|y_2\| + 1} \mid x_i \in D(A_i),\ y_i \in A_i x_i,\ i = 1, 2 \right\}$$

with the value possibly equal to $+\infty$.

The distance dis is not a metric because in a general case the triangle inequality is not fulfilled.

Concerning properties and applications to the study of evolution equations in a Hilbert space, the reader is referred to the original work [35].

The following lemma plays an important role in the proof of the convergence result, since it reduces the convergence problem for the nonlinear evolution equation to the convergence problem for a linear system of elasticity.

Lemma 2.6. *For the functions $\hat{z}_\eta(t)$ and $z_\eta(t)$, the following estimate with a constant C independent of η holds:*

$$\|\hat{z}_\eta(t) - z_\eta(t)\|^2_{\Omega \times Y} \le C \int_0^t \|\hat{\sigma}_\eta(s) - \sigma_\eta(s)\|_{\Omega \times Y} ds. \tag{45}$$

Proof. We use the operator distance introduced in Definition 2.5:

$$\operatorname{dis}(\hat{A}_\eta(t), A_\eta(t)) = \sup_{\substack{z_1 \in D(\hat{A}_\eta),\ z_2 \in D(A_\eta) \\ y_1 \in \hat{A}_\eta z_1,\ y_2 \in A_\eta z_2}} \frac{\langle y_1 - y_2, z_2 - z_1 \rangle_{\Omega \times Y, \eta}}{1 + \|y_1\|_{\Omega \times Y, \eta} + \|y_2\|_{\Omega \times Y, \eta}},$$

which is well defined, since $\hat{A}_\eta$ and A_η are maximal monotone operators.

We get

$$\begin{aligned}
&\langle y_1 - y_2, z_2 - z_1 \rangle_{\Omega \times Y, \eta} = (M_\eta(y_1 - y_2), z_2 - z_1)_{\Omega \times Y} \\
&= -(-y_2 + y_1, (-M_\eta z_2 + B^T \sigma_\eta(t)) - (-M_\eta z_1 + B^T \hat{\sigma}_\eta(t)))_{\Omega \times Y} \\
&\quad -(y_2 - y_1, B^T \sigma_\eta(t) - B^T \hat{\sigma}_\eta(t))_{\Omega \times Y} \\
&\le -(y_2 - y_1, B^T(\sigma_\eta(t) - \hat{\sigma}_\eta(t)))_{\Omega \times Y},
\end{aligned}$$

since the inclusion $-y_2 + y_1 \in G_\eta(-M_\eta z_2 + B^T\sigma_\eta(t)) - G_\eta(-M_\eta z_1 + B^T\hat{\sigma}_\eta(t))$ holds and the operator G_η is monotone.

Then we have

$$\operatorname{dis}(\hat{A}_\eta(t), A_\eta(t)) \leq \sup_{\substack{z_1\in D(\hat{A}_\eta),\ z_2\in D(A_\eta)\\ y_1\in\hat{A}_\eta z_1,\ y_2\in A_\eta z_2}} \frac{|(y_2-y_1, B^T(\sigma_\eta(t)-\hat{\sigma}_\eta(t)))_{\Omega\times Y}|}{1+\|y_1\|_{\Omega\times Y,\eta}+\|y_2\|_{\Omega\times Y,\eta}}$$

$$\leq \sup_{\substack{z_1\in D(\hat{A}_\eta),\ z_2\in D(A_\eta)\\ y_1\in\hat{A}_\eta z_1,\ y_2\in A_\eta z_2}} \frac{(\|y_2\|_{\Omega\times Y,\eta}+\|y_1\|_{\Omega\times Y,\eta})\|B^T(\sigma_\eta(t)-\hat{\sigma}_\eta(t))\|_{\Omega\times Y}}{1+\|y_1\|_{\Omega\times Y,\eta}+\|y_2\|_{\Omega\times Y,\eta}}$$

$$\leq \|B^T(\sigma_\eta(t)-\hat{\sigma}_\eta(t))\|_{\Omega\times Y} \leq C_1\|\sigma_\eta(t)-\hat{\sigma}_\eta(t)\|_{\Omega\times Y}.$$

Now we use the last inequality to obtain the main estimate. If $\hat{z}_\eta(t)$ and $z_\eta(t)$ are, respectively, absolutely continuous solutions of the monotone evolution equations (33) and (44) with the same initial conditions, then one easily gets

$$\frac{d}{dt}\|\hat{z}_\eta(t)-z_\eta(t)\|^2_{\Omega\times Y,\eta} = 2\langle \hat{z}_{\eta,t}(t)-z_{\eta,t}(t), \hat{z}_\eta(t)-z_\eta(t)\rangle_{\Omega\times Y,\eta}$$

$$= 2\frac{\langle \hat{z}_{\eta,t}(t)-z_{\eta,t}(t), \hat{z}_\eta(t)-z_\eta(t)\rangle_{\Omega\times Y,\eta}}{1+\|\hat{z}_{\eta,t}(t)\|_{\Omega\times Y,\eta}+\|z_{\eta,t}(t)\|_{\Omega\times Y,\eta}}(1+\|\hat{z}_{\eta,t}(t)\|_{\Omega\times Y,\eta}+\|z_{\eta,t}(t)\|_{\Omega\times Y,\eta})$$

$$\leq 2\ \operatorname{dis}(\hat{A}_\eta(t), A_\eta(t))(1+\|\hat{z}_{\eta,t}(t)\|_{\Omega\times Y,\eta}+\|z_{\eta,t}(t)\|_{\Omega\times Y,\eta})$$

$$\leq 2C_1\|\bar{\sigma}_\eta(t)\|_{\Omega\times Y}(1+\|\hat{z}_{\eta,t}(t)\|_{\Omega\times Y}+\|z_{\eta,t}(t)\|_{\Omega\times Y}),$$

since $-\hat{z}_{\eta,t}(t)\in\hat{A}_\eta\hat{z}_\eta(t)$, $-z_{\eta,t}(t)\in A_\eta z_\eta(t)$ a.e. Here $\bar{\sigma}_\eta(t)=\sigma_\eta(t)-\hat{\sigma}_\eta(t)$.

As a result of all calculations:

$$\|\hat{z}_\eta(T_e)-z_\eta(T_e)\|^2_{\Omega\times Y} \leq 2C_1\int_0^{T_e}\|\bar{\sigma}_\eta(t)\|_{\Omega\times Y}(1+\|\hat{z}_{\eta,t}(t)\|_{\Omega\times Y}+\|z_{\eta,t}(t)\|_{\Omega\times Y})dt.$$

We have to show that $\|\hat{z}_{\eta,t}(t)\|_{\Omega\times Y}$ and $\|z_{\eta,t}(t)\|_{\Omega\times Y}$ are uniformly bounded with respect to η.

We can transform (33) into an autonomous equation by inserting

$$h_\eta = -(B^T\mathcal{D}_\eta Q_\eta B + L_\eta)z_\eta + B^T\sigma_\eta$$

into (32). This autonomous equation is

$$\frac{d}{dt}h_\eta(t) + C_\eta h(t) \ni B^T\sigma_{\eta,t}(t)$$

with the operator $C_\eta : L^2(\Omega\times Y,\mathbb{R}^N) \to 2^{L^2(\Omega\times Y,\mathbb{R}^N)}$ defined by $C_\eta = (B^T\mathcal{D}_\eta Q_\eta B + L_\eta)G_\eta$. The operator C_η is maximal monotone (see Theorem 3.3 in [3]).

The estimate (79) then implies

$$\left\|\frac{\partial}{\partial t}h_\eta(t)\right\|_{\Omega\times Y} \leq ||G_\lambda h(0)||_{\Omega\times Y} + ||B^T\sigma_{\eta,t}(0)||_{\Omega\times Y} + \int_0^t ||B^T\sigma_{\eta,tt}(s)||_{\Omega\times Y}ds.$$

From the estimate (29) with $\hat{\varepsilon}_p = 0$ we conclude that the $L^2(\Omega\times Y)$-norm of σ_η is uniformly bounded with respect to η. By virtue of the assumptions made on b, γ, we can differentiate (25)–(27) with respect to t and apply the existence theory for elliptic problems to the obtained system. It results in the inequality

$$\|v_{\eta,t}\|_{L^2(Y,H^1(\Omega,\mathbb{R}^3))} + \|\sigma_{\eta,t}\|_{\Omega\times Y} \le C(\|\hat{b}_t\|_\Omega + \|\hat{\gamma}_t\|_{1,\Omega}), \tag{46}$$

with a constant C independent of η. Equation (46) yields that the $L^2(\Omega\times Y)$-norm of $\sigma_{\eta,t}$ is uniformly bounded with respect to η. Similarly, we conclude the same result for $\sigma_{\eta,tt}$. From (22) it follows that $||G_\lambda h(0)||_{\Omega\times Y}$ is also uniformly bounded with respect to η (G_λ is the Yosida approximation of G). These conclusions imply that the function $z_{\eta,t}$ is uniformly bounded.[5]

We notice that $\|\hat{z}_{\eta,t}\|_{\Omega\times Y} = \|z_{0,t}\|_{\Omega\times Y}$, where $z_0(t)$ is a solution of the homogenized problem. Thus the required result is obtained from the estimate (24). This completes the proof of the lemma. ∎

Note. Using the special distance between two maximal monotone operators, the difference of solutions of the evolution inclusions can be estimated by the norm of a function which solves a linear elasticity problem for special data. It is a crucial step in the justification of the homogenized model. Instead of treating evolution equations (inclusions) with, in general, nonlinear multivalued operators, the problem is reduced to the linear elasticity case. It allows us the possibility of using standard methods which work perfectly for linear problems.

2.5. End of the justification. Thus, we can estimate $\hat{z}_\eta - z_\eta$ by the difference $\sigma_\eta - \hat{\sigma}_\eta$. In the next section we present an estimate for the function $\sigma_\eta - \hat{\sigma}_\eta$. Both, σ_η and $\hat{\sigma}_\eta$ solve the same boundary value problem, the problem of linear elasticity theory, but for different data. Here we only state the estimate and refer to the next section for the proof.

LEMMA 2.7. *Let $(v_\eta(t), \sigma_\eta(t))$ be a solution of the boundary value problem* (25)–(27) *for the data*

$$\hat{b} = b(t), \qquad \hat{\varepsilon}_p = 0, \qquad \hat{\gamma} = \gamma(t),$$

and let $(\hat{v}_\eta(t), \hat{\sigma}_\eta(t))$ be a solution of the problem (25)–(27) *for the data*

$$\hat{b} = -\mathrm{div}_x T_0\left(x, \frac{x}{\eta} + y, t\right), \qquad \hat{\varepsilon}_p = -\varepsilon\left(\nabla_y u_1\left(x, \frac{x}{\eta} + y, t\right)\right), \qquad \hat{\gamma} = \gamma(t).$$

Then for all $t \in [0, T_e]$

$$\|v_\eta(t) - \hat{v}_\eta(t)\|_{\Omega\times Y} + \|\sigma_\eta(t) - \hat{\sigma}_\eta(t)\|_{\Omega\times Y} \to 0 \qquad \text{as } \eta \to 0. \tag{47}$$

Moreover, there exists a constant C independent of η such that for all $t \in [0, T_e]$ and all $\eta > 0$,

$$\|\sigma_\eta(t) - \hat{\sigma}_\eta(t)\|_{\Omega\times Y} \le C. \tag{48}$$

With this lemma the proof of Theorem 1.2 can be finished.

Proof of Theorem 1.2. Lemma 2.7 and the inequality (45) yield, together with Lebesgue's convergence theorem, that for all $t \in [0, T_e]$,

$$\lim_{\eta\to 0} \|\hat{z}_\eta(t) - z_\eta(t)\|_{\Omega\times Y} = 0. \tag{49}$$

[5] Remember also that the operator $(B^T \mathcal{D}_\eta Q_\eta B + L_\eta)$ is uniformly bounded and invertible.

We observe also that (14), (15), (18) form the boundary value problem for (u_η, T_η) and (36), (37), (40) form a boundary value problem for $(u_0, \hat{T}_\eta)$. Definition 2.1 of P_η and the definitions of $(\hat{v}_\eta(t), \hat{\sigma}_\eta(t))$ and $(v_\eta(t), \sigma_\eta(t))$ thus yield the decomposition

$$u_\eta = w_\eta + v_\eta, \quad T_\eta = \mathcal{D}\left[\frac{\cdot}{\eta} + y\right](P_\eta - I)Bz_\eta + \sigma_\eta,$$
$$u_0 = \hat{w}_\eta + \hat{v}_\eta, \quad \hat{T}_\eta = \mathcal{D}\left[\frac{\cdot}{\eta} + y\right](P_\eta - I)B\hat{z}_\eta + \hat{\sigma}_\eta,$$

where $w_\eta(t), \hat{w}_\eta(t) \in L^2(Y, H_0^1(\Omega, \mathbb{R}^3))$ are the unique functions from Definition 2.1 which satisfy $\varepsilon(\nabla_x w_\eta(t)) = P_\eta B z_\eta(t)$ and $\varepsilon(\nabla_x \hat{w}_\eta(t)) = P_\eta B \hat{z}_\eta(t)$. We thus have

$$\varepsilon(\nabla_x(w_\eta - \hat{w}_\eta)) = P_\eta B(z_\eta - \hat{z}_\eta), \tag{50}$$

$$T_\eta - \hat{T}_\eta = -\mathcal{D}\left[\frac{\cdot}{\eta} + y\right] Q_\eta B(z_\eta - \hat{z}_\eta) + (\sigma_\eta - \hat{\sigma}_\eta), \tag{51}$$

$$u_0 - u_\eta = (w_\eta - \hat{w}_\eta) + (v_\eta - \hat{v}_\eta). \tag{52}$$

From Lemma 2.7, (49), (51), and the uniform boundedness of $\mathcal{D}_\eta Q_\eta B$, we infer that

$$\lim_{\eta\to 0} \|\hat{T}_\eta(t) - T_\eta(t)\|_{\Omega\times Y} = 0.$$

Since for a.e. $y \in Y$ the function $(w_\eta - \hat{w}_\eta)$ belongs to $H_0^1(\Omega, \mathbb{R}^3)$, we conclude from the first Korn's inequality that the inequality $\|(w_\eta - \hat{w}_\eta)(t,y)\|_{1,\Omega} \le C\|\varepsilon(\nabla_x(w_\eta - \hat{w}_\eta)(t,y))\|_\Omega$ holds for a.e. y and from (49), (50) that $\|w_\eta(t) - \hat{w}_\eta(t)\|_{\Omega\times Y} \to 0$ as $\eta \to 0$; from Lemma 2.7 and (52) we thus conclude that

$$\lim_{\eta\to 0} \|u_0(t) - u_\eta(t)\|_{\Omega\times Y} = 0$$

for all $t \in [0, T_e]$. These two relations and (49) together yield

$$\lim_{\eta\to 0}(\|u_0(t) - u_\eta(t)\|_{\Omega\times Y} + \|\hat{T}_\eta(t) - T_\eta(t)\|_{\Omega\times Y} + \|\hat{z}_\eta(t) - z_\eta(t)\|_{\Omega\times Y}) = 0.$$

This completes the proof of Theorem 1.2. $\square$

To finish the proof of Theorem 1.2 it thus remains to verify Lemma 2.7. The next section is devoted to the proof of this lemma.

3. Convergence results based on the two-scale convergence method. In this section we present the convergence result stated in Lemma 2.7. To do this an auxiliary function is taken into consideration. We define it to be a solution of a Dirichlet boundary value problem of the linear elasticity theory for special data. These data are chosen to be smooth enough so the two-scale convergence method can be applied. The direct application of the method seems impossible to the author because of the low regularity of the functions $T_0(x,y,t)$ and $u_1(x,y,t)$, the solutions of the homogenized problem, which are now considered as the data for the elasticity problem.

Now we give the definition of the so-called two-scale converged sequence.

We assume that $|Y| = 1$.

Definition 3.1. *A sequence of functions u_η in $L^2(\Omega, \mathbb{R}^3)$ is said to two-scale converge to a limit $u_0(x,y)$ belonging to $L^2(\Omega\times Y, \mathbb{R}^3)$ if, for any test function $\psi(x,y)$ in $L^2(\Omega, C(Y, \mathbb{R}^3))$, one has*

$$\lim_{\eta\to 0} \int_\Omega u_\eta(x)\psi\left(x, \frac{x}{\eta}\right) dx = \int_\Omega \int_Y u_0(x,y)\psi(x,y)\,dxdy. \tag{53}$$

Note. This definition makes sense for every bounded sequence u_η in $L^2(\Omega, \mathbb{R}^3)$. As it is shown in Theorem 1.2 in [6] for such a sequence u_η there exists a limit $u_0 \in L^2(\Omega \times Y, \mathbb{R}^3)$ such that, with possible expanse of extracting a subsequence, $u_\eta(x)$ two-scale converges to $u_0(x, y)$.

Note. The two-scale convergence method [27], [6] is an alternative to the energy method of Tartar [26], applied for partial differential equations with periodically oscillating coefficients. We refer the reader to [6] for main properties of two-scale converged sequences.

Let the boundary value problem be given:

$$-\mathrm{div}_x T(x, y) = \hat{b}(x), \tag{54}$$

$$T(x, y) = \mathcal{D}\left[\frac{x}{\eta} + y\right](\varepsilon(\nabla_x u(x, y)) - \hat{\varepsilon}_p(x, y)), \tag{55}$$

$$u(x, y) = \hat{\gamma}(x), \qquad x \in \partial\Omega, \tag{56}$$

with given functions $\hat{b} : \Omega \mapsto \mathbb{R}^3$, $\hat{\gamma} : \partial\Omega \mapsto \mathbb{R}^3$, $\hat{\varepsilon}_p : \Omega \mapsto \mathcal{S}^3$, a given number $\eta > 0$, and fixed $y \in Y$. This is the linear problem of elasticity with a parameter y.

We recall that for a.e. $y \in Y$ the function $(v_\eta(t), \sigma_\eta(t))$ solves the boundary value problem (54)–(56) to the data

$$\hat{b} = b(t), \qquad \hat{\varepsilon}_p = 0, \qquad \hat{\gamma} = \gamma(t),$$

and $(\hat{v}_\eta(t), \hat{\sigma}_\eta(t))$ solves the problem (54)–(56) to

$$\hat{b} = -\mathrm{div}_x T_0\left(x, \frac{x}{\eta} + y, t\right), \qquad \hat{\varepsilon}_p = -\varepsilon\left(\nabla_y u_1\left(x, \frac{x}{\eta} + y, t\right)\right), \qquad \hat{\gamma} = \gamma(t).$$

My goal now is to show that for almost all $y \in Y$ and all $t \in [0, T_e]$,

$$\lim_{\eta \to 0} \|v_\eta(\cdot, y, t) - \hat{v}_\eta(\cdot, y, t)\|_\Omega = 0,$$
$$\lim_{\eta \to 0} \|\sigma_\eta(\cdot, y, t) - \hat{\sigma}_\eta(\cdot, y, t)\|_\Omega = 0.$$

These relations follow from two auxiliary lemmas proved in the next section.

3.1. Two auxiliary lemmas.

LEMMA 3.2. *Let the function $\tau \in L^2(\Omega \times Y, \mathcal{S}^3)$ have the property $\mathrm{div}_y \tau(x, y) = 0$, and let the family $\{\tau_{\eta,n}(x) = \tau_n(x, x/\eta)\}_{\eta,n}$ with $\tau_n \in L^2(\Omega, C(Y, \mathcal{S}^3))$ be such that the sequence $\tau_n(x, y)$ converges strongly to $\tau(x, y)$ in $L^2(\Omega \times Y, \mathcal{S}^3)$. Denote*

$$\tau_{n,\infty}(x) := \frac{1}{|Y|}\int_Y \tau_n(x, y)dy \quad \text{and} \quad \tau_\infty(x) = \frac{1}{|Y|}\int_Y \tau(x, y)dy.$$

Then $\tau_{n,\infty}(x)$ converges strongly to $\tau_\infty(x)$ in $L^2(\Omega, \mathcal{S}^3)$.

Let $(v_{\eta,n}, \sigma_{\eta,n}) \in H_0^1(\Omega, \mathbb{R}^3) \times L^2(\Omega, \mathcal{S}^3)$ be a weak solution of the boundary value problem formed by the equations

$$-\mathrm{div}\ \sigma_{\eta,n} = b + \mathrm{div}_x \tau_{\eta,n}, \qquad x \in \Omega, \tag{57}$$

$$\sigma_{\eta,n} = \mathcal{D}\left[\frac{\cdot}{\eta}\right]\ \varepsilon(\nabla_x v_{\eta,n}), \qquad x \in \Omega. \tag{58}$$

If, additionally, the function τ_∞ satisfies

$$-\mathrm{div}\tau_\infty = b$$

for $b \in L^2(\Omega, \mathbb{R}^3)$, *then*

$$\lim_{n\to\infty} \lim_{\eta\to 0} (\|v_{\eta,n}\|_\Omega + \|\sigma_{\eta,n}\|_\Omega) = 0.$$

Proof. First, we observe that the symmetry of the matrices $\sigma_{\eta,n}$, $\tau_{\eta,n}$ and (57)–(58) yields

$$\begin{aligned} c\|\sigma_{\eta,n}\|_\Omega^2 &\le \int_\Omega \mathcal{D}^{-1}\left[\frac{x}{\eta}\right] \ \sigma_{\eta,n}(x)\cdot\sigma_{\eta,n}(x)dx = \int_\Omega \varepsilon(\nabla_x v_{\eta,n}(x))\cdot\sigma_{\eta,n}(x)dx \\ &= (v_{\eta,n}, b)_\Omega + \left(\operatorname{div}_x \tau_n\left(\cdot, \frac{\cdot}{\eta}\right), v_{\eta,n}\right)_\Omega \\ &= (v_{\eta,n}, b)_\Omega - \left(\tau_n\left(\cdot, \frac{\cdot}{\eta}\right), \nabla_x v_{\eta,n}\right)_\Omega . \end{aligned} \tag{59}$$

Now we notice that for a fixed n the function $\tau_n(x, x/\eta)$ can be considered as a test function in the definition of the two-scale convergence (see Definition 3.1), and by properties of these functions

$$\|\psi\left(\cdot, \frac{\cdot}{\eta}\right)\|_\Omega \le \|\psi(\cdot,\cdot)\|_{L^2(\Omega, C(Y,\mathbb{R}^3))} \equiv \left(\int_\Omega \sup_{y\in Y} |\psi(x,y)|^2 dx\right)^{1/2}, \tag{60}$$

we can easily conclude using the standard estimates for elliptic boundary value problems that the sequence $v_{\eta,n}$ is uniformly bounded in $H_0^1(\Omega,\mathbb{R}^3)$ for a fixed n. Then by virtue of the property (i) of Proposition 1.14 in [6], one gets the following result:

$$\begin{aligned} &(v_{\eta,n}, b)_\Omega - \left(\tau_n\left(\cdot, \frac{\cdot}{\eta}\right), \nabla_x v_{\eta,n}\right)_\Omega \to (v_{0,n}, b)_\Omega \\ &\qquad - \int_{\Omega\times Y} \tau_n(x,y)\cdot\nabla_x v_{0,n}(x)dxdy - \int_{\Omega\times Y} \tau_n(x,y)\cdot\nabla_y w_{1,n}(x,y)dxdy \\ &= (v_{0,n}, b)_\Omega - (\tau_{n,\infty}, \nabla_x v_{0,n})_\Omega - \int_{\Omega\times Y} \tau_n(x,y)\cdot\nabla_y w_{1,n}(x,y)dydx, \end{aligned} \tag{61}$$

where the function $(v_{0,n}, w_{1,n}) \in H_0^1(\Omega,\mathbb{R}^3)\times L^2(\Omega, W(Y,\mathbb{R}^3))$ solves the problem written in the variational form

$$\begin{aligned} &\int_Y\int_\Omega \mathcal{D}[y]\varepsilon(\nabla v_{0,n}(x) + \nabla_y w_{1,n}(x,y))\varepsilon(\nabla\psi(x) + \nabla_y\psi_1(x,y))dxdy \\ &\qquad = (b + \operatorname{div}_x \tau_{n,\infty}, \psi)_\Omega \end{aligned} \tag{62}$$

with a function $(\psi,\psi_1) \in H_0^1(\Omega,\mathbb{R}^3)\times L^2(\Omega, W(Y,\mathbb{R}^3))$. Here $v_{0,n}(x)$ is a weak limit of $v_{\eta,n}(x)$ in $H_0^1(\Omega,\mathbb{R}^3)$. Equation (62) is obtained in the same way as (89) in the proof of Theorem 4.2.

The existence and uniqueness of the solution for this problem is obtained in Theorem 4.2. If in (62) we choose $\psi = v_{0,n}$ and $\psi_1 = w_{1,n}$, then we easily obtain the following estimate for $(v_{0,n}(x), w_{1,n}(x,y))$ with constants C, C_1 independent of n:

$$\|v_{0,n}\|_{1,\Omega}^2 + \|w_{1,n}\|_{L^2(\Omega, W(Y,\mathbb{R}^3))}^2 \le C(\|b\|_\Omega + \|\tau_{n,\infty}\|_\Omega) \le C_1. \tag{63}$$

As a consequence of the last estimate, one can extract subsequences of $v_{0,n}(x)$ and $w_{1,n}(x,y)$, which converge weakly to $v_{0,\infty}(x)$ in $H_0^1(\Omega,\mathbb{R}^3)$ and to $w_{1,\infty}(x,y)$ in $L^2(\Omega, W(Y,\mathbb{R}^3))$, respectively. Taking into account this fact and the properties of τ_∞ and τ we finally obtain after the passage to the limit in (62) as $n\to\infty$ that

$$(v_{0,\infty},b)_\Omega-(\tau_\infty,\nabla_x v_{0,\infty})_\Omega-(\tau,\nabla_y w_{1,\infty})_{\Omega\times Y}$$

$$=(b+\mathrm{div}_x\tau_\infty,v_{0,\infty})_\Omega+(\mathrm{div}_y\tau,w_{1,\infty})_{\Omega\times Y}=0.$$

This means that $\|\sigma_{\eta,n}\|_\Omega\to 0$ as $\eta\to 0$ and $n\to\infty$. Together with Korn's inequality it follows that

$$\lim_{n\to\infty}\lim_{\eta\to 0}(\|\sigma_{\eta,n}\|_\Omega+\|v_{\eta,n}\|_\Omega)=0.$$

This ends the proof of Lemma 3.2. $\square$

LEMMA 3.3. *Let $\kappa_\eta(x):=\kappa(x,x/\eta)$ be a sequence of functions in $L^2(\Omega,\mathbb{R}^{3\times3})$, where $\kappa\in L^2(\Omega,C(Y,\mathbb{R}^{3\times3}))$ satisfies the relation $\kappa=\nabla_y\vartheta$ with a suitable Y-periodic in y function $\vartheta(x,y)$. Suppose that $(v_\eta,\sigma_\eta)\in H_0^1(\Omega,\mathbb{R}^3)\times L^2(\Omega,\mathcal{S}^3)$ is a weak solution of the boundary value problem formed by the equations*

$$-\mathrm{div}\ \sigma_\eta=0,\qquad x\in\Omega,\tag{64}$$

$$\sigma_\eta=\mathcal{D}\left[\frac{\cdot}{\eta}\right]\ (\varepsilon(\nabla_x v_\eta)+\varepsilon(\kappa_\eta)),\qquad x\in\Omega.\tag{65}$$

Then

$$\lim_{\eta\to 0}(\|v_\eta\|_\Omega+\|\sigma_\eta\|_\Omega)=0.$$

Proof. Similarly as in the proof of Lemma 3.2, the symmetry of σ_η and (64)–(65) yields

$$c\|\sigma_\eta\|_\Omega^2\le\int_\Omega\kappa\left(x,\frac{x}{\eta}\right)\cdot\sigma_\eta(x)dx+\int_\Omega\nabla_x v_\eta\cdot\sigma_\eta(x)dx=(\kappa_\eta,\sigma_\eta)_\Omega.\tag{66}$$

Notice that due to the regularity assumption on the function $\kappa(x,x/\eta)$ it can be taken as a test function in the sense of the definition of two-scale convergence (see Definition 3.1). Moreover, from the estimate

$$\left\|\psi\left(x,\frac{x}{\eta}\right)\right\|_{L^2(\Omega,\mathbb{R}^{3\times3})}\le\|\psi(x,y)\|_{L^2(\Omega,C(Y,\mathbb{R}^{3\times3}))},$$

which holds for every $\psi\in L^2(\Omega,C(Y,\mathbb{R}^{3\times3}))$ and from (66) we obtain

$$c\|\sigma_\eta\|_\Omega^2\le\|\kappa_\eta\|_{L^2(\Omega,C(Y,\mathbb{R}^{3\times3}))}\ \|\sigma_\eta\|_\Omega,$$

which implies that $\sigma_\eta(x)$ is uniformly bounded in $L^2(\Omega,\mathcal{S}^3)$. Simultaneously, one gets a uniform bound for v_η in $H_0^1(\Omega,\mathbb{R}^3)$. The boundness of σ_η allows us to pass to the limit in the inequality (66) as $\eta\to 0$. Using the property (iii) of Proposition 1.14 in [6] we obtain

$$\int_\Omega\kappa\left(x,\frac{x}{\eta}\right)\cdot\sigma_\eta(x)dx\to\int_{\Omega\times Y}\kappa(x,y)\cdot\sigma_0(x,y)dxdy=-(\vartheta,\mathrm{div}_y\sigma_0)_{\Omega\times Y}=0,$$

where $\sigma_0(x,y)$ is a two-scale limit of the sequence $\sigma_\eta(x)$. From the last limit relation and the inequality (66) we can conclude that σ_η converges strongly to 0 in $L^2(\Omega,\mathcal{S}^3)$.

Now we observe that the uniform boundness of v_η in $H^1_0(\Omega,\mathbb{R}^3)$ gives us the possibility of extracting a subsequence, which converges weakly to a function v_0 in $H^1_0(\Omega,\mathbb{R}^3)$, or, by the compactness result, converges strongly to the same function v_0 in $L^2(\Omega,\mathbb{R}^3)$. To finish the proof of Lemma 3.3 we have to show that $v_0(x)=0$. Notice first that well-known properties[6] of functions from $L^2(\Omega,C(Y,\mathbb{R}^{3\times 3}))$ yield the weak convergence of κ_η to 0 in $L^2(\Omega,\mathbb{R}^{3\times 3})$. Indeed, from the weak convergence of $\kappa(\cdot,\cdot/\eta)$ and periodicity of $\vartheta(x,y)$ with respect to the second variable it follows that for $\psi\in L^2(\Omega,\mathbb{R}^{3\times 3})$,

$$\int_\Omega \kappa\left(x,\frac{x}{\eta}\right)\cdot\psi(x)dx \to \int_\Omega\int_Y \nabla_y\vartheta(x,y)dy\ \psi(x)dx = 0.$$

Taking into account the last result we get from the equality

$$(\varepsilon(\nabla_x v_\eta),\psi)_\Omega + (\kappa_\eta,\psi)_\Omega = \int_\Omega \mathcal{D}^{-1}\left[\frac{x}{\eta}\right]\sigma_\eta(x)\cdot\psi(x)dx,$$

where $\psi\in L^2(\Omega,\mathcal{S}^3)$, that $\varepsilon(\nabla_x v_\eta(x))$ weakly converges to 0 in $L^2(\Omega,\mathcal{S}^3)$.[7]

Since $\nabla_x v_\eta \rightharpoonup \nabla_x v_0$ in $L^2(\Omega,\mathbb{R}^{3\times 3})$ it follows that $\varepsilon(\nabla_x v_0(x))=0$. Using that $v_0\in H^1_0(\Omega,\mathbb{R}^3)$ we conclude from the first Korn's inequality that $v_0(x)=0$.

Therefore

$$\|\sigma_\eta\|_\Omega + \|v_\eta\|_\Omega \to 0 \qquad \text{as} \quad \eta\to 0.$$

This completes the proof of Lemma 3.3. ∎

3.2. Proof of Lemma 2.7. Now we are well prepared to prove Lemma 2.7. The crucial point in the proof of Lemma 2.7 is to introduce an auxiliary function, which solves a linear elasticity problem to smoothed data. These smoothed data satisfy all requirements of Lemmas 3.2 and 3.3. The rest of the proof is the consequence of these two lemmas.

Proof. Let $T_{0,n}(x,y,t)$ be a sequence of smooth functions in $C_0^\infty(\Omega,C(Y,\mathcal{S}^3))$ that converges strongly to $T_0(x,y,t)$ in $L^2(\Omega\times Y,\mathcal{S}^3)$ for all t, and let $u_{1,n}(t)$ be another sequence of smooth functions in $C_0^\infty(\Omega,C(Y,\mathbb{R}^3))$ that converges strongly to $u_1(t)$ in $L^2(\Omega,H^1(Y,\mathbb{R}^3))$ for all t. We notice that an approximation sequence for $T_\infty(t)$ in the strong topology of $L^2(\Omega,\mathcal{S}^3)$ necessarily has to be of the form $T_{\infty,n}(x,t)=\int_Y T_{0,n}(x,y,t)dy$.

Now we fix t and introduce an auxiliary function[8] $(v_{\eta,n},\sigma_{\eta,n})$. We define it as a unique solution of an elasticity problem to the data determined by the smooth functions $T_{0,n}(x,y,t)$ and $u_{1,n}(x,y,t)$:

$$-\mathrm{div}\sigma_{\eta,n}(x) = -\mathrm{div}_x T_{0,n}\left(x,\frac{x}{\eta}+y,t\right), \tag{67}$$

[6]We use the fact (Lemma 9.1 of [18]) that for any $\psi\in L^2(\Omega,C(Y,\mathbb{R}^{3\times 3}))$ the sequence

$$\psi\left(\cdot,\frac{\cdot}{\eta}\right) \rightharpoonup \int_Y \psi(\cdot,y)dy \quad \text{weakly in } L^2(\Omega,\mathbb{R}^{3\times 3}).$$

[7]From the strong convergence of σ_η to 0 and the inequality $\|\mathcal{D}^{-1}[\cdot/\eta]\sigma_\eta\|_\Omega \le C\|\sigma_\eta\|_\Omega$, we conclude that the sequence $\mathcal{D}^{-1}[\cdot/\eta]\sigma_\eta$ also converges strongly to 0 in $L^2(\Omega,\mathcal{S}^3)$. The constant C is independent of η.

[8]A similar idea was used in [8].

$$\sigma_{\eta,n}(x) = \mathcal{D}\left[\frac{x}{\eta}+y\right]\left(\varepsilon(\nabla_x v_{\eta,n}(x)) + \varepsilon\left(\nabla_y u_{1,n}\left(x, \frac{x}{\eta}+y, t\right)\right)\right), \tag{68}$$

$$v_{\eta,n}(x) = \gamma(x,t), \qquad x \in \partial\Omega. \tag{69}$$

The existence and uniqueness of the function $(v_{\eta,n}, \sigma_{\eta,n})$ as a solution of the boundary value problem (67)–(69) can be obtained by the well-known theory for elliptic problems. Details are omitted.

We obviously have that

$$\sigma_\eta(x,t) - \hat{\sigma}_\eta(x,t) = (\sigma_\eta(x,t) - \sigma_{\eta,n}(x)) + (\sigma_{\eta,n}(x) - \hat{\sigma}_\eta(x,t)),$$

$$v_\eta(x,t) - \hat{v}_\eta(x,t) = (v_\eta(x,t) - v_{\eta,n}(x)) + (v_{\eta,n}(x) - \hat{v}_\eta(x,t)).$$

The proof of the convergence result will be separated into two steps. In the first step we show that the sequences $(\bar{v}_{\eta,n}(x),\, \bar{\sigma}_{\eta,n}(x))$ with $\bar{\sigma}_{\eta,n}(x) := \sigma_\eta(x,t) - \sigma_{\eta,n}(x)$ and $\bar{v}_{\eta,n}(x) := v_\eta(x,t) - v_{\eta,n}(x)$ converge to 0 in appropriate strong topologies. Then the same result will be shown for sequences $(\hat{v}_{\eta,n}(x),\, \hat{\sigma}_{\eta,n}(x))$, where $\hat{\sigma}_{\eta,n}(x) := \sigma_{\eta,n}(x) - \hat{\sigma}_\eta(x,t)$ and $\hat{v}_{\eta,n}(x) := v_{\eta,n}(x) - \hat{v}_\eta(x,t)$.

First step. By definition, $(\bar{v}_{\eta,n}(x),\, \bar{\sigma}_{\eta,n}(x))$ is a weak solution of the boundary value problem (54)–(56) to the data

$$\hat{b}(x) = b(x,t) + \mathrm{div}_x T_{0,n}\left(x, \frac{x}{\eta}+y, t\right), \tag{70}$$

$$\hat{\varepsilon}_p(x,y) = \varepsilon\left(\nabla_y u_{1,n}\left(x, \frac{x}{\eta}+y, t\right)\right), \tag{71}$$

$$\hat{\gamma}(x) = 0. \tag{72}$$

By linearity of the problem (54)–(56) to the data (70)–(72) the required convergence of $(\bar{v}_{\eta,n}(x),\, \bar{\sigma}_{\eta,n}(x))$ is an easy consequence of Lemmas 3.2 and 3.3, which are applied for a.e. value $y = \hat{y} \in Y$.

Indeed, to see that we set for fixed t and a.e. $\hat{y}$,

$$\tau_{\eta,n}(x) = T_{0,n}\left(x, \frac{x}{\eta}+\hat{y}, t\right), \quad \tau(x,y) = T_0(x, y+\hat{y}, t), \quad \tau_\infty(x) = T_\infty(x,t),$$
$$\kappa_{\eta,n}(x) = -\nabla_y u_{1,n}\left(x, \frac{x}{\eta}+\hat{y}, t\right), \quad \tau_{\infty,n}(x) = T_{\infty,n}(x,t), \quad b = b(t).$$

Note first that the periodicity of $T_{0,n}, T_0$ and the choice of $T_{0,n}$ yield

$$\int_{\Omega\times Y} (T_{0,n}(x, y+\hat{y}, t) - T_0(x, y+\hat{y}, t))^2 dxdy$$
$$= \int_{\Omega\times Y} (T_{0,n}(x,y,t) - T_0(x,y,t))^2 dxdy \to 0 \ \text{ as } n \to \infty.$$

From (9) one immediately gets

$$\mathrm{div}_y \tau(x,y) = \mathrm{div}_y T_0(x, y+\hat{y}, t) = 0$$

and

$$\mathrm{div}_x \tau_{\eta,n}(x) = \mathrm{div}_x T_{0,n}\left(x, \frac{x}{\eta}+\hat{y}, t\right).$$

Due to the periodicity of $T_0(x, y, t)$ we obtain also that

$$\int_Y T_0(x, y + \hat{y}, t)dy = \int_Y T_0(x, y, t)dy = T_\infty(x, t) = \tau_\infty(x),$$

and from the strong convergence of $T_{0,n}$ to T_0 we deduce that the sequence $\tau_{\infty,n}(x) = T_{\infty,n}(x, t)$ converges strongly to $\tau_\infty(x) = T_\infty(x, t)$ in $L^2(\Omega, \mathcal{S}^3)$. One needs to apply Hölder's inequality.

Moreover, (7) implies that

$$b + \operatorname{div}_x \tau_\infty(x) = b(t) + \operatorname{div}_x T_\infty(t) = 0.$$

Since by definition $(\bar{v}_{\eta,n}(x), \bar{\sigma}_{\eta,n}(x))$ is a weak solution of the boundary value problem (54)–(56) to the data

$$\begin{aligned}
\hat{b}(x) &= b(x, t) + \operatorname{div}_x T_{0,n}\left(x, \frac{x}{\eta} + \hat{y}, t\right) = b + \operatorname{div}_x \tau_{\eta,n},\\
\hat{\varepsilon}_p(x) &= \varepsilon(\nabla_y u_{1,n}(x, y, t))|_{y=\frac{x}{\eta}+\hat{y}} = -\varepsilon(\kappa_\eta)\\
\hat{\gamma}(x) &= 0,
\end{aligned}$$

we can use the linearity of the problem (54)–(56) to write the function $(\bar{v}_{\eta,n}(x), \bar{\sigma}_{\eta,n}(x))$ as a sum of two functions

$$(\bar{v}_{\eta,n}, \bar{\sigma}_{\eta,n})(x) = (\bar{v}, \bar{\sigma})(x) + (\hat{v}, \hat{\sigma})(x),$$

where the function $(\bar{v}, \bar{\sigma})(x)$ solves the boundary problem (57)–(58) and the function $(\hat{v}, \hat{\sigma})(x)$ solves the problem (64)–(65). Then the application of Lemmas 3.2 and 3.3 to $(\bar{v}, \bar{\sigma})$ and $(\hat{v}, \hat{\sigma})$ gives, for every t and a.e. $\hat{y} \in Y$,

$$\lim_{n\to\infty} \lim_{\eta\to 0} (\|\bar{v}_{\eta,n}(\cdot, \hat{y}, t)\|_\Omega + \|\bar{\sigma}_{\eta,n}(\cdot, \hat{y}, t)\|_\Omega) = 0. \tag{73}$$

Second step. We fix t and y. The function $(\hat{v}_{\eta,n}(x), \hat{\sigma}_{\eta,n}(x))$ is a weak solution of the boundary value problem (54)–(56) to the data

$$\hat{b}(x) = -\operatorname{div}_x \left(T_{0,n}\left(x, \frac{x}{\eta} + y, t\right) - T_0\left(x, \frac{x}{\eta} + y, t\right)\right) \tag{74}$$

$$\hat{\varepsilon}_p(x) = -\varepsilon\left(\nabla_y\left(u_{1,n}\left(x, \frac{x}{\eta} + y, t\right) - u_1\left(x, \frac{x}{\eta} + y, t\right)\right)\right), \tag{75}$$

$$\hat{\gamma}(x) = 0. \tag{76}$$

Using the properties of $\mathcal{D}_\eta$, we obtain an elliptic theory applied to the boundary value problem (54)–(56) to the data (74)–(76), observing that $\hat{b} \in H^{-1}(\Omega, \mathbb{R}^3)$, $\hat{\varepsilon}_p \in L^2(\Omega, \mathcal{S}^3)$, $\hat{\gamma} \in H^1(\Omega, \mathbb{R}^3)$, and $\eta > 0$, $y \in Y$, and that there is a constant C independent of η, y, t, and n, such that

$$\begin{aligned}
\|\hat{\sigma}_{\eta,n}(t)\|_\Omega^2 \le C \Bigg[&\left\| T_{0,n}\left(\cdot, \frac{\cdot}{\eta} + y, t\right) - T_0\left(\cdot, \frac{\cdot}{\eta} + y, t\right)\right\|_\Omega^2\\
&+ \left\| \nabla_y u_{1,n}\left(\cdot, \frac{\cdot}{\eta} + y, t\right) - \nabla_y u_1\left(\cdot, \frac{\cdot}{\eta} + y, t\right)\right\|_\Omega^2 \Bigg].
\end{aligned}$$

We integrate the right-hand side of this inequality with respect to the parameter y over Y. As a result we have

$$\int_\Omega \int_Y \left| T_{0,n}\left(x, \frac{x}{\eta} + y, t\right) - T_0\left(x, \frac{x}{\eta} + y, t\right) \right|^2$$

$$+ \left| \nabla_y u_{1,n}\left(x, \frac{x}{\eta} + y, t\right) - \nabla_y u_1\left(x, \frac{x}{\eta} + y, t\right) \right|^2 dydx$$

$$= \int_\Omega \int_{\frac{x}{\eta}+Y} |T_{0,n}(x,y,t) - T_0(x,y,t)|^2$$

$$+ |\nabla_y u_{1,n}(x,y,t) - \nabla_y u_1(x,y,t)|^2 \, dydx$$

$$= \|T_{0,n}(t) - T_0(t)\|^2_{\Omega\times Y} + \|\nabla_y u_{1,n}(t) - \nabla_y u_1(t)\|^2_{\Omega\times Y}.$$

Thus the function $\sigma_\eta(t) - \hat{\sigma}_\eta(t)$ must satisfy the following inequality:

$$\|\sigma_\eta(t) - \hat{\sigma}_\eta(t)\|^2_{\Omega\times Y} \le C \left(\int_Y \|\bar{\sigma}_{\eta,n}(y,t)\|^2_\Omega dy \right.$$

$$\left. + \|T_{0,n}(t) - T_0(t)\|^2_{\Omega\times Y} + \|\nabla_y u_{1,n}(t) - \nabla_y u_1(t)\|^2_{\Omega\times Y} \right).$$

We notice that the $L^2(\Omega)$-norm of $\bar{\sigma}_{\eta,n}$ is uniformly bounded with respect to η and y. Indeed, $(\bar{v}_{\eta,n}(x), \bar{\sigma}_{\eta,n}(x))$ is the solution of (54)–(56) to the data (70)–(72). Applying the standard existence theory for linear elliptic problems we obtain that

$$\|\bar{\sigma}_{\eta,n}(y,t)\|_\Omega \le C \left(\|b(t)\|_\Omega + \left\| T_{0,n}\left(\cdot, \frac{\cdot}{\eta} + y, t\right) \right\|_\Omega + \left\| \nabla_y u_{1,n}\left(\cdot, \frac{\cdot}{\eta} + y, t\right) \right\|_\Omega \right).$$

The functions $T_{0,n}$, $\nabla_y u_{1,n}$ can be considered as an "admissible" test function in the definition of the two-scale convergence. By the properties of these functions we get that

$$\|\bar{\sigma}_{\eta,n}(y,t)\|_\Omega \le C \, (\|b(t)\|_\Omega + \|T_{0,n}(t)\|_{L^2(\Omega, C(Y, \mathcal{S}^3))}$$
$$+ \|\nabla_y u_{1,n}(t)\|_{L^2(\Omega, C(Y, \mathbb{R}^{3\times 3}))}).$$

Thus we can use Lebesgue's convergence theorem to interchange the passage to the limit and the integration in the lines below.

Let us now pass to the limit as $\eta \to 0$:

$$\lim_{\eta\to 0} \|\sigma_\eta(t) - \hat{\sigma}_\eta(t)\|^2_{\Omega\times Y} \le C \quad \lim_{\eta\to 0} \left(\int_Y \|\bar{\sigma}_{\eta,n}(y,t)\|^2_\Omega dy \right.$$

$$\left. + \|T_{0,n}(t) - T_0(t)\|^2_{\Omega\times Y} + \|\nabla_y u_{1,n}(t) - \nabla_y u_1(t)\|^2_{\Omega\times Y} \right)$$

$$= C \left(\int_Y \lim_{\eta\to 0} \|\bar{\sigma}_{\eta,n}(y,t)\|^2_\Omega dy + \|T_{0,n}(t) - T_0(t)\|^2_{\Omega\times Y} \right.$$

$$\left. + \|\nabla_y u_{1,n}(t) - \nabla_y u_1(t)\|^2_{\Omega\times Y} \right).$$

Lemmas 3.2 and 3.3 imply, if we set $\tau_{\eta,n}(x) = T_{0,n}(x, x/\eta + y, t)$ and $\kappa_{\eta,n}(x) = -\nabla_y u_{1,n}(x, x/\eta + y, t)$, that the function $\lim_{\eta\to 0} \|\bar{\sigma}_{\eta,n}(y,t)\|^2_\Omega$ is uniformly bounded with respect to n and y. It follows from (59), (62), and (63) and we can now pass to the limit as $n \to \infty$:

$$\begin{aligned}
&\lim_{\eta\to 0} \|\sigma_\eta(t) - \hat{\sigma}_\eta(t)\|^2_{\Omega\times Y} \le \lim_{n\to\infty} C \left(\int_Y \lim_{\eta\to 0} \|\bar{\sigma}_{\eta,n}(y,t)\|^2_{2,\Omega} dy \right. \\
&\left. + \lim_{n\to\infty} \|T_{0,n}(t) - T_0(t)\|^2_{\Omega\times Y} + \lim_{n\to\infty} \|\nabla_y u_{1,n}(t) - \nabla_y u_1(t)\|^2_{\Omega\times Y} \right) \\
&= C \left(\int_Y \lim_{n\to\infty} \lim_{\eta\to 0} \|\bar{\sigma}_{\eta,n}(y,t)\|^2_{2,\Omega} dy + \lim_{n\to\infty} \|T_{0,n}(t) - T_0(t)\|^2_{\Omega\times Y} \right. \\
&\left. + \lim_{n\to\infty} \|\nabla_y u_{1,n}(t) - \nabla_y u_1(t)\|^2_{\Omega\times Y} \right) = 0.
\end{aligned}$$

In exactly the same way we get

$$\begin{aligned}
&\lim_{\eta\to 0} \|v_\eta(t) - \hat{v}_\eta(t)\|^2_{\Omega\times Y} \le C \left(\int_Y \lim_{n\to\infty} \lim_{\eta\to 0} \|\bar{v}_{\eta,n}(y)\|^2_{2,\Omega} dy \right. \\
&\left. + \lim_{n\to\infty} \|T_{0,n}(t) - T_0(t)\|^2_{\Omega\times Y} + \lim_{n\to\infty} \|\nabla_y u_{1,n}(t) - \nabla_y u_1(t)\|^2_{\Omega\times Y} \right) = 0.
\end{aligned}$$

This ends the proof of Lemma 2.7. ∎

4. Appendix.

4.1. Existence of solutions for the reduced equation. In this subsection we are going to show the existence result[9] in a Hilbert space H for the following Cauchy problem:

$$\frac{d}{dt} u(t) + A(u(t)) \ni f(t), \tag{77}$$

$$u(0) = u_0 \tag{78}$$

with $A = MG$, where M is a linear, bounded, positive definite, selfadjoint operator, and G is a maximal monotone operator with respect to the usual scalar product $(\cdot,\cdot)$. It is already shown in Theorem 3.3 in [3] that A is maximal monotone with respect to the scalar product $\langle\cdot,\cdot\rangle = (M^{-1}\cdot,\cdot)$.

THEOREM 4.1. *Let G_λ be a Yosida approximation of G. Assume that $u_0 \in D(A)$ and $f \in W^{1,1}(0,T_e;H)$. Then the Cauchy problem has a unique solution $u \in W^{1,\infty}(0,T_e;H)$. The solution satisfies the inequality*

$$\left\| \frac{d}{dt} u(t) \right\|_H \le C \|G_\lambda u_0\|_H + \|f(0)\|_H + \int_0^t \|f'(s)\|_H ds \tag{79}$$

with a constant C independent of λ.

[9] Actually in [11], [32] it is already shown that the problem (77) has a unique solution, but in one place in the justification proof we need other estimates for solutions than this theory delivers. We slightly modify the classical proof in order to obtain the mentioned estimates. More details can be found in [11], [32].

Proof. The uniqueness follows directly from the monotonicity of A.

For each $\lambda > 0$ let u_λ be the solution of

$$\frac{d}{dt}u_\lambda(t) + A_\lambda(u_\lambda(t)) \ni f(t), \tag{80}$$

$$u_\lambda(0) = u_0 \tag{81}$$

with a maximal monotone (with respect to $\langle\cdot,\cdot\rangle = (M^{-1}\cdot,\cdot)$) operator $A_\lambda = MG_\lambda$.

Then similarly as in Theorem IV.4.1 in [32] we get the estimate for $u_\lambda^{'}$,

$$\begin{aligned}\|u_\lambda^{'}(t)\|_H &\le \|A_\lambda u_0\|_H + \|f(0)\|_H + \int_0^t \|f^{'}(s)\|_H ds \\ &\le C\|G_\lambda u_0\|_H + \|f(0)\|_H + \int_0^t \|f^{'}(s)\|_H ds \\ &\le C|G^0 u_0| + \|f(0)\|_H + \int_0^t \|f^{'}(s)\|_H ds. \end{aligned} \tag{82}$$

From (80) and (82) it follows that $u_\lambda^{'}$, u_λ, and $A_\lambda u_\lambda$ are uniformly bounded in $C([0,T_e],H)$.

u_λ is a Cauchy sequence in $C([0,T_e],H)$. To see that, let $\lambda,\mu > 0$ and use (80) to obtain

$$\frac{1}{2}\frac{d}{dt}\|u_\lambda(t) - u_\mu(t)\|^2 = -\langle A_\lambda(u_\lambda(t)) - A_\mu(u_\mu(t)), u_\lambda(t) - u_\mu(t)\rangle.$$

With $u_\lambda = \lambda G_\lambda u_\lambda + J_\lambda u_\lambda$ and $u_\mu = \mu G_\mu u_\mu + J_\mu u_\mu$ (J_λ is a resolvent of G) we obtain as in Theorem IV.4.1 in [32]

$$\|u_\lambda(t) - u_\mu(t)\|^2 \le \frac{\lambda+\mu}{2}K^2, \qquad 0 \le t \le T_e,$$

where $K = \sup\{\|G_\lambda(u_\lambda(t))\| \mid 0 \le t \le T_e,\ \lambda > 0\}$, so u_λ is a Cauchy sequence in $C([0,T_e],H)$ with

$$\|u_\lambda(t) - u(t)\|_H \le \sqrt{\frac{\lambda}{2}}C\left(|G^0 u_0| + \|f(0)\|_H + \|f\|_{C([0,T_e],H)} + \int_0^{T_e} \|f^{'}(s)\|_H ds\right). \tag{83}$$

The proof ends similarly as in Theorem IV.4.1 in [32]. □

4.2. Homogenization of linear elasticity systems. Now we show how to apply the two-scale convergence method to the homogenization of linear elasticity systems with periodically oscillating coefficients. This example is of great importance in the rigorous justification procedure because of the frequent use of estimates obtained for the sequence of solutions of the linear elasticity problem as well as of its homogenized problem. Therefore the proof, actually a rephrasing of Theorem 2.3 in [6] for the case of linear elasticity with a slight modification, is given in detail.

Consider the following problem:

$$-\mathrm{div}\ \mathcal{D}\left[\frac{x}{\eta}\right]\varepsilon\left(\nabla_x u(x)\right) = b(x), \quad x \in \Omega, \tag{84}$$

$$u(x) = 0, \qquad x \in \partial\Omega, \tag{85}$$

with a given function $b \in L^2(\Omega, \mathbb{R}^3)$ and a Y-periodic linear positive definite mapping $\mathcal{D}[y] : \mathcal{S}^3 \mapsto \mathcal{S}^3$, the elasticity tensor. $\mathcal{D}[y]$ is such that there exist two positive constants $0 < \alpha \leq \beta$ satisfying

$$\alpha|\xi|^2 \leq \mathcal{D}_{ijkl}[y]\xi_{kl}\xi_{ij} \leq \beta|\xi|^2 \quad \text{for any } \xi \in \mathcal{S}^3. \tag{86}$$

The last assumption (86) implies that the mapping $\mathcal{D}[y]$ belongs to $L^\infty(Y, \mathcal{S}^3)$ and, consequently, by virtue of the first Korn's inequality (see, for example, [28]) that the problem (84)–(85) admits a unique solution u_η in $H^1_0(\Omega, \mathbb{R}^3)$, which satisfies the estimate

$$\|u_\eta\|_{1,\Omega} \leq C\|b\|_2, \tag{87}$$

where C is a positive constant that depends on Ω and α and not on η.

THEOREM 4.2. *The sequence u_η of solutions of the problem* (84)–(85) *converges weakly to $u(x)$ in $H^1_0(\Omega, \mathbb{R}^3)$, and the sequence ∇u_η two-scale converges to $\nabla u(x) + \nabla_y u_1(x, y)$, where $(u, u_1) \in H^1_0(\Omega, \mathbb{R}^3) \times L^2(\Omega, W(Y, \mathbb{R}^3))$ is the unique solution of the following two-scale homogenized system:*

$$\begin{aligned} -\operatorname{div}_y[\mathcal{D}[y]\varepsilon(\nabla u(x) + \nabla_y u_1(x, y))] &= 0 \quad \textit{in } \Omega \times Y, \\ -\operatorname{div}_x\left[\int_Y \mathcal{D}[y]\varepsilon(\nabla u(x) + \nabla_y u_1(x, y))dy\right] &= b(x) \quad \textit{in } \Omega, \\ y \mapsto u_1(x, y) &\quad Y\textit{-periodic.} \end{aligned}$$

Proof. By virtue of the estimate (87) and Proposition 1.14 in [6], the sequence u_η and the sequence of its gradient ∇u_η, up to a subsequence, have the weak limit $u \in H^1_0(\Omega, \mathbb{R}^3)$ and the two-scale limit $\nabla u(x) + \nabla_y u_1(x, y)$, respectively, with $u_1 \in L^2(\Omega, W(Y, \mathbb{R}^3))$.

Multiply (84) by a test function $\psi(x) + \eta\psi_1(x, x/\eta)$, with $\psi \in C^\infty_0(\Omega, \mathbb{R}^3)$ and $\psi_1 \in C^\infty_0(\Omega, C^\infty(Y, \mathbb{R}^3))$. Integration by parts of the resulting equation and some rewriting imply

$$\begin{aligned} &\int_\Omega \nabla u_\eta(x)\mathcal{D}\left[\frac{x}{\eta}\right]\varepsilon\left[\nabla\psi(x) + \nabla_y\psi_1\left(x, \frac{x}{\eta}\right)\right]dx \\ &+\eta\int_\Omega \mathcal{D}\left[\frac{x}{\eta}\right]\varepsilon(\nabla u_\eta(x))\varepsilon\left(\nabla_x\psi_1\left(x, \frac{x}{\eta}\right)\right)dx = \int_\Omega b(x)\left(\psi(x) + \eta\psi_1\left(x, \frac{x}{\eta}\right)\right)dx. \end{aligned} \tag{88}$$

Here the symmetry of $\mathcal{D}$ was used in the first term.

It is easily seen now that applying consecutively the limit relation in the definition of the admissible test functions and Theorem 1.8 in [6] justifies the passage to the two-scale limit in (88):

$$\begin{aligned} &\int_Y\int_\Omega \mathcal{D}[y]\varepsilon(\nabla u(x) + \nabla_y u_1(x, y))\varepsilon(\nabla\psi(x) + \nabla_y\psi_1(x, y))dxdy \\ &= \int_\Omega b(x)\psi(x)dx. \end{aligned} \tag{89}$$

Equation (89) holds true for any function $(\psi, \psi_1) \in H^1_0(\Omega, \mathbb{R}^3) \times L^2(\Omega, W(Y, \mathbb{R}^3))$. Equation (89) is the variational formulation associated with the two-scale homogenized problem stated above. Due to the first Korn's inequality and to the Korn's inequality for periodic functions (see [28]), the Hilbert space $\mathcal{H} = H_0^1(\Omega,\mathbb{R}^3)\times L^2(\Omega, W(Y,\mathbb{R}^3))$ can be endowed with the following norm:

$$\|\Psi\|_{\mathcal{H}}^2 = \|\psi\|_{1,\Omega}^2 + \|\varepsilon(\nabla_y\psi_1)\|_{\Omega\times Y}^2.$$

Then the application of the Lax–Milgram lemma shows that there exists a unique solution of the two-scale homogenized problem. Consequently, the entire sequences $u_\eta(x)$ and $\nabla u_\eta(x)$ converge to $u(x)$ and $\nabla u(x) + \nabla_y u_1(x,y)$. ∎

4.3. Homogenization of the second order elliptic operators with nonuniformly oscillating coefficients. For another interesting application of a parameter y consider the following problem with nonuniformly oscillating coefficients (Chapter 1, section 6 in [9]):

$$-\mathrm{div}\, \mathcal{D}\left[x, \frac{x}{\eta}\right](\nabla_x u_\eta(x)) = b(x), \quad x\in\Omega, \tag{90}$$

$$u_\eta(x) = 0, \qquad x\in\partial\Omega, \tag{91}$$

with a given $b\in L^2(\Omega)$ and with a Y-periodic in y matrix $\mathcal{D}[x,y]$. We assume that there exist two positive constants $0<\alpha\le\beta$ such that the matrix $\mathcal{D}[x,y]$ satisfies

$$\alpha|\xi|^2 \le \mathcal{D}_{ij}[x,y]\xi_i\xi_j \le \beta|\xi|^2 \quad \text{for any } \xi\in\mathbb{R}^N. \tag{92}$$

The last assumption (92) implies that the mapping $\mathcal{D}[x,y]$ belongs to $L^\infty(\Omega\times Y,\mathbb{R}^{N\times N})$. But it is not enough to ensure that the mapping $x\mapsto\mathcal{D}[x,x/\eta]$ is measurable, so in [9] it is assumed that $\mathcal{D}\in C(\Omega, L^\infty(Y,\mathbb{R}^{N\times N}))$ to obtain a solution of (90)–(91). Instead of increasing the regularity of $\mathcal{D}[x,y]$ we consider a family of shifted problems

$$-\mathrm{div}_x\, \mathcal{D}\left[x, \frac{x}{\eta}+y\right](\nabla_x u_\eta(x,y)) = b(x), \qquad (x,y)\in\Omega\times Y, \tag{93}$$

$$u_\eta(x,y) = 0, \qquad (x,y)\in\partial\Omega\times Y. \tag{94}$$

Now the problem (93)–(94) admits a unique solution u_η in $L^2(Y, H_0^1(\Omega))$, which satisfies the estimate

$$\|u_\eta\|_{L^2(Y,H_0^1(\Omega))} \le C\|b\|_\Omega,$$

where C is a positive constant independent of η. For a.e. fixed $y\in Y$ the existence of the solution is provided by the well-known result for second order elliptic operators and the integrability with respect to y is then the easy consequence of it.

It is convenient now to write (93)–(94) in the form

$$-\mathrm{div}_x\sigma_\eta(x,y) = b(x), \qquad (x,y)\in\Omega\times Y, \tag{95}$$

$$\sigma_\eta(x,y) = \mathcal{D}\left[x, \frac{x}{\eta}+y\right](\nabla_x u_\eta(x,y)), \quad (x,y)\in\Omega\times Y, \tag{96}$$

$$u_\eta(x,y) = 0, \qquad (x,y)\in\partial\Omega\times Y. \tag{97}$$

Now the solution of (95)–(97) is a function $(u_\eta,\sigma_\eta)\in L^2(Y,H_0^1(\Omega))\times L^2(\Omega\times Y,\mathbb{R}^N)$.

Then inserting the formal ansatz for the solution u_η

$$\hat{u}_\eta(x,y) = u_0(x) + \eta u_1\left(x, \frac{x}{\eta}+y\right) + \eta^2 u_2\left(x,\frac{x}{\eta}+y\right) + \cdots$$

into the boundary problem (93)–(94) and identifying powers of η lead to the homogenized problem

$$\begin{aligned}
-\mathrm{div}_x \sigma_\infty(x) &= b(x), \\
\sigma_\infty(x) &= \frac{1}{|Y|}\int_Y \sigma_0(x,y)dy, \\
-\mathrm{div}_y \sigma_0(x,y) &= 0, \\
\sigma_0(x,y,t) &= \mathcal{D}[x,y](\nabla_y u_1(x,y) + \nabla_x u_0(x)),
\end{aligned}$$

which must hold for $(x,y) \in \Omega \times Y$,

$$u_0(x) = 0,$$

which must hold for $x \in \partial\Omega$. This form of the homogenized problem is equivalent to the already obtained one in [9]. A slight modification of the proof of Lemma 2.7 yields the following convergence result:

$$\lim_{\eta\to 0}(\|u_0 - u_\eta\|_{\Omega\times Y} + \|\sigma_0 - \sigma_\eta\|_{\Omega\times Y}) = 0, \tag{98}$$

where (u_0, u_1, σ_0) is the solution of the homogenized problem and (u_η, σ_η) is the solution of the problem (95)–(97) with a parameter y.

Equation (98) holds without imposing additional regularity on $\mathcal{D}[x,y], b(x)$, and $\partial\Omega$.

Acknowledgment. The author thanks Hans-Dieter Alber for helpful suggestions.

REFERENCES

[1] H.-D. Alber, *Materials with Memory. Initial-Boundary Value Problems for Constitutive Equations with Internal Variables*, Lecture Notes in Math. 1682, Springer-Verlag, Berlin, 1998.

[2] H.-D. Alber, *Evolving microstructure and homogenization*, Contin. Mech. Thermodyn., 12 (2000), pp. 235–286.

[3] H.-D. Alber and K. Chelminski, *Quasistatic problems in viscoplasticity theory.* I. *Models with linear hardening*, in Operator Theoretical Methods and Applications to Mathematical Physics, Oper. Theory Adv. Appl. 147, Birkhäuser, Basel, 2004, pp. 105–129.

[4] H.-D. Alber, *Justification of homogenized models for viscoplastic bodies with microstructure*, in Deformation and Failure in Metallic Materials, K. Hutter, and H. Baaser, eds., Lecture Notes Appl. Comput. Mech. 10, Springer, Berlin, 2003, pp. 295–319.

[5] G. Allaire, *Shape Optimization by the Homogenization Method*, Appl. Math. Sci. 146, Springer-Verlag, New York, 2002.

[6] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[7] N. S. Bakhvalov and G. P. Panasenko, *Homogenization: Averaging Processes in Periodic Media*, Nauka, Moscow, 1984 (in Russian).

[8] N. S. Bakhvalov and A. Bourgeat, *Convergence towards the homogenized solution in case of Lipschitzian boundary*, Appl. Anal., 80 (2001), pp. 161–175.

[9] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, Stud. Math. Appl. 5, North–Holland, Amsterdam, 1978.

[10] A. Braides and A. Defranceschi, *Homogenization of Multiple Integrals*, Oxford Lecture Ser. Math. Appl. 12, Clarendon Press, Oxford University Press, New York, 1998.

[11] H. Brézis, *Operateurs maximaux monotones*, North–Holland, Amsterdam, 1973.

[12] V. I. Burenkov, *Sobolev Spaces on Domains*, Teubner-Texte Math. 137, Teubner, Stuttgart, 1998.

[13] K. Chelminski, *On monotone plastic constitutive equations with polynomal growth condition*, Math. Methods Appl. Sci., 22 (1999), pp. 547–562.

[14] K. Chelminski, *Corecive approximation of viscoplasticity and plasticity*, Asymptot. Anal., 26 (2001), pp. 105–133.

[15] K. Chelminski, *Global existence of weak-type solutions for models of monotone type in the theory of inelastic deformations*, Math. Methods Appl. Sci., 25 (2002), pp. 1195–1230.

[16] K. Cherednichenko, *Higher Order and Non-Local Effects in Homogenization of Periodic Media*, Ph.D. thesis, Bath, 2001.

[17] V. Chiadò Piat and G. V. Sandrakov, *Homogenization of some variational inequalities for elasto-plastic torsion problems*, Asymptot. Anal., 40 (2004), pp. 1–23.

[18] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Clarendon Press, Oxford University Press, New York, 1999.

[19] G. Dal Maso, *An Introduction to Γ-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser Boston, Boston, 1993.

[20] G. A. Francfort, *Homogenization and linear thermoelasticity.* SIAM J. Math. Anal., 14 (1983), pp. 696–708.

[21] G. A. Francfort and F. Murat, *Homogenization and optimal bounds in linear elasticity*, Arch. Rational Mech. Anal., 94 (1986), pp. 307–334.

[22] G. A. Francfort and P.-M. Suquet, *Homogenization and mechanical dissipation in thermoviscoelasticity*, Arch. Rational Mech. Anal., 96 (1986), pp. 265–293.

[23] U. Hornung, ed., *Homogenization and Porous Media*, Interdisciplinary Appl. Math. 6, Springer-Verlag, New York, 1997.

[24] W. Jäger and A. Mikelic, *On the effective equations of a viscous incompressible fluid flow through a filter of finite thickness*, Comm. Pure Appl. Math., 51 (1998), pp. 1073–1121.

[25] V. Jikov, S. Kozlov, and O. Oleinik, *Homogenization of Differential Operators and Integral Functions*, Springer-Verlag, Berlin, 1995.

[26] F. Murat, L. Tartar, H-*convergence*, in Topics in the Mathematical Modelling of Composite Materials, A. Cherkaev, and R.V. Kohn, eds., Progr. Nonlinear Differential Equations Appl. 31, Birkhäuser Boston, Boston, 1997, pp. 21–43.

[27] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.

[28] O. A. Oleinik, A. S. Shamaev, and G. A. Yosifian, *Mathematical Problems in Elasticity and Homogenization*, Stud. Math. Appl. 26, North–Holland, Amsterdam, 1992.

[29] A. Pankov, *G-Convergence and Homogenization of Nonlinear Partial Differential Operators*, Math. Appl. 422, Kluwer, Dordrecht, 1997.

[30] W. Pompe, *Quasistatic and dynamic problems in viscoplasticity theory—non-linearities with power growth*, Math. Methods Appl. Sci., 27 (2004), pp. 1347–1365.

[31] E. Sánchez-Palencia, *Non-Homogeneous Media and Vibration Theory*, Lecture Notes in Physics 127, Springer-Verlag, Berlin-New York, 1980.

[32] R. E. Showalter, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, Math. Surveys Monogr. 49, AMS, Providence, RI, 1997.

[33] L. Tartar, *Compensated compactness and partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium Vol. IV, Pitman, Boston, 1979, pp. 136–212.

[34] T. Valent, *Boundary Value Problems of Finite Elasticity. Local Theorems on Existence, Uniqueness, and Analytic Dependence on Data*, Springer Tracts Nat. Philos. 31, Springer-Verlag, New York, 1988.

[35] A. A. Vladimirov, *Nonstationary dissipative evolution equations in a Hilbert space*, Nonlinear Anal., 17 (1991), pp. 499–518.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 263–280

FLUX OF SUPERCONDUCTING VORTICES THROUGH A DOMAIN*

S. N. ANTONTSEV† AND N. V. CHEMETOV‡

Abstract. This article addresses the mathematical study of a mean-field model of superconducting vortices in a II-type superconductor, previously introduced in [S. J. Chapman, *SIAM J. Appl. Math.*, 55 (1995), pp. 1259–1279]. We investigate a hyperbolic-elliptic type system of PDEs in a given domain. Motivated by physical experiments, we consider nonzero and nonconstant boundary conditions, which describe a flux of superconducting vortices through the domain. We prove the existence of the regular solutions of a parabolic-elliptic approximated system and establish a uniform L_∞-bound for the vorticity and the convergence to the initial system. Finally, we analyze the regularity of weak solutions.

Key words. mean-field model, superconducting vortices, flux, solvability

AMS subject classifications. 78A25, 35D05, 76B47

DOI. 10.1137/060655146

1. The statement of the problem.

Certain materials, when cooled through a critical temperature, exhibit a superconducting state in which they have the ability to conduct electric currents without resistance. Many models have been proposed to describe the behavior of superconductors: the microscopic theory of Bardeen, Cooper, and Schrieffer, the mesoscopic theories of London, Ginzburg, and Landau, and the macroscopic critical state theories, like the Bean model. For the description and relationship of these models we refer to the excellent works of Chapman [3], [4]. In this paper we investigate a mean-field model of superconducting vortices in a II-type superconductor, formulated in [5]. In 1-D and 2-D cases the following system describes the evolution of the vorticity ω and the average magnetic field h of the superconducting sample:

$$\omega_t + \operatorname{div}(\omega\vec{v}) = 0, \quad (\vec{x},t) \in \Omega_T := \Omega \times (0,T), \tag{1.1}$$

$$-\Delta h + h = \omega, \qquad (\vec{x},t) \in \Omega_T, \tag{1.2}$$

$$\vec{v} = -\nabla h, \qquad (\vec{x},t) \in \Omega_T, \tag{1.3}$$

with the following boundary conditions:

$$\vec{v}\,\vec{n} = a(\vec{x},t), \quad (\vec{x},t) \in \Gamma_T := \Gamma \times (0,T), \tag{1.4}$$

$$\omega = b(\vec{x},t), \quad (\vec{x},t) \in \Gamma_T^- := \Gamma^- \times (0,T) \tag{1.5}$$

and the initial conditions

$$\omega(\vec{x},0) = \omega_0(\vec{x}), \quad \vec{x} \in \Omega. \tag{1.6}$$

*Received by the editors March 24, 2006; accepted for publication (in revised form) December 19, 2006; published electronically May 22, 2007.
http://www.siam.org/journals/sima/39-1/65514.html

†Departamento de Matemática, Universidade da Beira Interior, Rua Marquês d'Avila e Bolama, 6201-001 Covilha, Portugal (anton@ubi.pt). The work of this author was partially supported by the Portuguese research project POCI/MAT/61576/2004/FCT/MCES.

‡CMAF/Universidade de Lisboa, Av. Prof. Gama Pinto, 2, 1649-003 Lisboa, Portugal (chemetov@ptmat.fc.ul.pt). The work of this author was partially supported by FCT and FEDER through the Project POCTI/ISFL/209 of Centro de Matemática e Aplicações Fundamentais da Universidade de Lisboa (CMAF/UL).

Here $\vec{n}$ denotes the outward normal to the $C^{2+\gamma}$-smooth boundary $\Gamma = \Gamma^+ \cup \Gamma^0 \cup \Gamma^-$ of the bounded domain $\Omega \subset \mathbb{R}^n$, with some $\gamma > 0$; Γ^+ is the part of the boundary Γ, where $\vec{v}\,\vec{n} = a > 0$; Γ^0 is the part of Γ, where $\vec{v}\,\vec{n} = a = 0$; Γ^- is the part of Γ, where $\vec{v}\,\vec{n} = a < 0$ and $meas(\Gamma^+) \neq 0$, $meas(\Gamma^-) \neq 0$. Let us note that the physical statement of the system is valid just for $n = 1$ and $n = 2$, but in this article we shall consider the common case, when $n \geq 1$. Note that the system (1.1)–(1.3) is of hyperbolic-elliptic type for unknown functions $\{\omega, h\}$.

Many articles on (1.1)–(1.3) have been written in the last ten years. The problem has been studied through various approaches ranging from asymptotic analysis, numerical simulations, and rigorous mathematical analysis. As we know that the first positive existence result for the system (1.1)–(1.3), (1.6) was obtained by Huang and Svobodny [12] in the case of the Cauchy problem for the domain $\Omega = \mathbb{R}^2$, using the method of characteristics and potentials. In the articles [2], [17], the system (1.1)–(1.3), (1.6) was considered in the case of a bounded domain Ω with boundary conditions

$$
\begin{aligned}
h(\vec{x}, t) &= a, \quad (\vec{x}, t) \in \Gamma_T, \\
\omega(\vec{x}, t) &= b, \quad (\vec{x}, t) \in \Gamma_T^-,
\end{aligned} \tag{1.7}
$$

where a, b are given *constants*. In [17] the existence of a weak solution of (1.1)–(1.3), (1.6)–(1.7) for $b = 0$ was deduced by an approach based on a parabolic-elliptic approximation. In the 1-D case the authors have shown the uniqueness of the solution. Also in this work and in [6] the existence of a steady state solution $\{\omega, h\}$ of (1.1)–(1.3), (1.6)–(1.7) was demonstrated. In [9] the solvability of (1.1)–(1.3), (1.6)–(1.7) with an additional curvature term in (1.1) was shown. In [10], the case of flux pinning and boundary nucleation of vorticity (i.e., the *constants* $a, b > 0$) was considered, where existence and uniqueness of a solution in the 1-D case were proved. The discretization, the convergence of discretizated solutions, and numerical simulations of (1.1)–(1.3) were extensively studied in articles [7], [11], [19], [2]. In all mentioned works the boundary conditions (1.7) were taken as *zero or constant* valued. The evolution of vortices arises under penetration of the magnetic field into superconducting bodies and due to it the vortices are generated *only at the boundary*, hence the question of the boundary conditions is important in the modelling of the motion of the vortices. Chapman discussed this question in many articles (see, for instance, [3], [4], [5]). One of the possible boundary conditions suggested by him is to define the value of the vorticity ω on Γ^- as a nonlinear combination of the current $\nabla^\perp h$,

$$
\omega = \alpha \max\{|\nabla^\perp h| - J, \quad 0\} \qquad \text{on } \Gamma_T^-, \tag{1.8}
$$

where α, J are physical constants. The physical meaning of the condition (1.8) is that the nucleation of the vortices at the boundary provided the magnitude of current is larger than the *so-called* nucleation current J. The nucleation of the vorticity is a consequence of the behavior of the current $\nabla^\perp h$, i.e., the case of nonzero or nonconstant boundary conditions is really physical.

In this article we assume that the nucleation of the vorticity can be measured by a physical experiment, which is possible (we refer to the Web site: http://www.fys.uio.no/super/results/sv/index.html), that is, the normal component of the velocity $\vec{v}$ and the vorticity ω on the boundary Γ are defined by (1.4) and (1.5), respectively. The system (1.1)–(1.6) exhibits a transport effect, which plays an important role in understanding the behavior of the vortices *after the nucleation*.

In this article, we consider the problem (1.1)–(1.6) for the general case, when $a(\vec{x},t)$, $b(\vec{x},t)$ are *arbitrary* given functions of the conditions (1.4)–(1.5). We prove the existence of solution $\{\omega, h\}$ with a natural smooth restriction on the data of the problem.

1.1. Terminology and notations. In accordance with the notations introduced in the books [14], [15], we shall use Sobolev's spaces $L_q(\Omega)$, $W_q^l(\Omega)$, $W_q^{l,\,m}(\Omega_T)$, $W_q^l(\Gamma)$, $W_q^{l,\,m}(\Gamma_T)$ for $l,q \geq 1$, $m \geq 0$ and the Hölder spaces $C^l(\Omega)$, $C^{l,\,m}(\Omega_T)$, $C^{l,\,m}(\Gamma_T)$ for $l,m \geq 0$, where the index l corresponds to the variable $\vec{x}$ and m to the variable t (l,q,m are integer or noninteger).

Let B be a given Banach space. We denote by $C(0,T;B)$ the space of continuous functions from $[0,T]$ into the Banach space B with the norm

$$||u||_{C(0,T;\ B)} = \max_{t\in[0,T]} ||u(t)||_B$$

and by $L_p(0,T;B)$ for any $1 \leq p \leq \infty$, the space of measurable functions from $[0,T]$ into the Banach space B with pth power summable on $[0,T]$, with the norm

$$||u||_{L_p(0,T;\ B)} = \left(\int_0^T ||u(t)||_B^p dt\right)^{\frac{1}{p}} \qquad \Big(= ess \sup_{0<t<T} ||u(t)||_B \text{ if } p=\infty\Big).$$

We also introduce the space $V_2^{1,0}(\Omega_T)$ with the norm

$$||u||_{V_2^{1,0}(\Omega_T)} = \max_{t\in[0,T]} ||u(x,t)||_{L_2(\Omega)} + ||u_x(x,t)||_{L_2(\Omega_T)}.$$

1.2. Regularity of data. We assume that the datum a satisfies the condition

$$a \in C^{\eta,\,\theta}(\Gamma_T) \quad \text{for some } \eta,\, \theta \in (0,1) \tag{1.9}$$

and the data $b,\ \omega_0$ admit an extension $\breve{\omega}$, defined on the domain $\overline{\Omega}_T$, such that

$$\breve{\omega}(\vec{x},t) = b(\vec{x},t), \quad (\vec{x},t) \in \Gamma_T^-; \qquad \breve{\omega}(\vec{x},0) = \omega_0(\vec{x}), \quad \vec{x} \in \Omega; \tag{1.10}$$

and $\breve{\omega}$ satisfies

$$0 \leq \breve{\omega}(\vec{x},t) \leq \aleph \quad \text{for a.e. } (\vec{x},t) \in \overline{\Omega}_T, \tag{1.11}$$

$$(||\partial_t\breve{\omega}(\cdot,t)||_{L_1(\Omega)} + || \triangle \breve{\omega}(\cdot,t)||_{L_1(\Omega)} + || \bigtriangledown \breve{\omega}(\cdot,t)||_{L_\infty(\Omega)}) \in L_1(0,T), \tag{1.12}$$

where $\aleph$ is a constant.

Remark 1. The assumption (1.11) implies that the functions ω_0 and b are positive and bounded above by the constant $\aleph$ on $\overline{\Omega}$ and Γ_T^-, respectively. In section 6 we give sufficient conditions on the data $b,\ \omega_0$ that permit the existence of an extension $\breve{\omega}$ on the domain $\overline{\Omega}_T$, satisfying the conditions (1.10)–(1.12).

Let us give the definition of the weak solution of our problem.

Definition 1. *A pair of functions $\{\omega, h\}$ is said to be a weak solution of the problem* (1.1)–(1.6)*, if $\omega \in L_\infty(\Omega_T)$, $h \in L_\infty(0,T;\ C^{1+\alpha}(\overline{\Omega})) \cap L_2(0,T;\ W_2^2(\Omega'))$ for some $\alpha \in (0,1)$ and for any subdomain Ω' of Ω such that $\overline{\Omega'} \subset \Omega$ and the following equalities*:

$$\int_{\Omega_T} \omega(\psi_t + \vec{v}\,\nabla\psi)\, d\vec{x}dt + \int_\Omega \omega_0\ \psi(\vec{x},0)\ d\vec{x} = -\int_{\Gamma_T^-} a\ b\ \psi\ d\vec{x}dt, \tag{1.13}$$

$$-\Delta h + h = \omega \qquad a.e.\ in\ \Omega_T, \tag{1.14}$$

$$\vec{v} = -\nabla h \qquad a.e.\ in\ \Omega_T, \tag{1.15}$$

$$-\nabla h\,\vec{n} = a \qquad a.e.\ on\ \Gamma_T \tag{1.16}$$

hold for an arbitrary function $\psi \in H^1(\Omega_T)$, *such that*

$$\psi(\vec{x},T) = 0 \text{ for } \vec{x} \in \Omega \quad \text{and} \quad \psi(\vec{x},t) = 0 \text{ for } (\vec{x},t) \in \Gamma_T^+ \cup \Gamma_T^0. \tag{1.17}$$

Our main result in this work is the following theorem.

THEOREM 1. *If the data* a, b, ω_0 *satisfy* (1.9)–(1.12), *then there exists at least one weak solution* $\{\omega, h\}$ *of the problem* (1.1)–(1.6). *Moreover, we have*

$$\omega \in L_\infty(\Omega_T),\ \partial_t\omega \in L_\infty(0,T;H^{-1}(\Omega)) \tag{1.18}$$

and

$$\begin{gathered} h \in L_\infty(0,T;\, C^{1+\eta}(\overline{\Omega})) \cap L_\infty(0,T;\, W_q^2(\overline{\Omega'})),\\ h,\ \bigtriangledown h \in C^{0,\,\theta}(\overline{\Omega'} \times [0,T]) \end{gathered} \tag{1.19}$$

for any $q \in (1,\infty)$ *and for any* Ω', *such that* $\overline{\Omega'} \subset \Omega$.

Remark 2. A similar result is valid for both Dirichlet and Robin boundary conditions (in accordance with "www.wikipedia.org" terminology), instead of (1.16), on the function h. For simplicity, we prove this theorem for the dimension $n \geq 2$.

We divide the proof of this theorem into four steps:

Step 1: In section 2, we remember well-known results from the theory of elliptic equations that will be very useful in what follows.

Step 2: In section 3, we show that an approximate problem of (1.1)–(1.6), which is of parabolic-elliptic type, is solvable, if the initial and boundary conditions are smooth; the reasoning is based on the Schauder fixed point theorem.

Step 3: In section 4, by using the principle of maximum-minimum and the results of potential theory for elliptic equations, mentioned in section 2, we deduce L_∞-estimates and some additional a priori estimates for the solutions of the approximate problem of (1.1)–(1.6).

Step 4: In section 5, we use a compactness argument to establish the existence of a solution for the original problem.

2. Auxiliary assertions. The elliptic boundary value problem. In this section, for the convenience of the reader, we give well-known results on elliptic equations.

According to the potential theory for elliptic equations [20], the solution of the problem

$$\begin{cases} -\Delta h_1 + h_1 = F, & \vec{x} \in \Omega, \\ -\nabla h_1\, \vec{n} = 0, & \vec{x} \in \Gamma \end{cases} \tag{2.1}$$

can be written in the form $h_1(\vec{x}) = (K_1 * F)(\vec{x}) := \int_\Omega K_1(\vec{x},\vec{y})\ F(\vec{y})\ d\vec{y}$, and the solution of the problem

$$\begin{cases} -\Delta h_2 + h_2 = 0, & \vec{x} \in \Omega, \\ -\nabla h_2\, \vec{n} = g, & \vec{x} \in \Gamma \end{cases} \tag{2.2}$$

can be written in the form $h_2 = (K_2 * g)(\vec{x}) := \int_\Gamma K_2(\vec{x},\vec{y}) g(\vec{y})\ d\vec{y}$. Since Γ is $C^{2+\gamma}$-smooth, the kernels $K_1,\ K_2$ satisfy the inequalities

$$\begin{aligned} |K_i(\vec{x},\vec{y})| &\leq C|x-y|^{2-n}, \\ |\nabla_{\vec{x}} K_i(\vec{x},\vec{y})| &\leq C|x-y|^{1-n}, \quad i=1,2, \qquad \text{for any } \vec{x},\vec{y} \in \Omega \end{aligned}$$

in the n-dimensional case with $n > 2$ and

$$|K_i(\vec{x},\vec{y})| \le C|ln|x-y||,$$
$$|\nabla_{\vec{x}} K_i(\vec{x},\vec{y})| \le C|x-y|^{-1}, \quad i = 1,2, \qquad \text{for any } \vec{x},\vec{y} \in \Omega$$

in the 2-D case.

According to the theory of elliptic equations [15, pp. 169–193], the embedding theorem of Sobolev, $W_p^1(\Omega) \subset C^{\alpha}(\overline{\Omega})$ with $\alpha = 1 - \frac{n}{p}$ if $n < p < \infty$ (see, for instance, [14, p. 61]), and the potential theory [13, p. 191, Lemma 1.4], [8, p. 88, Lemma 10.1]; [16, p. 100, Theorem 3.4 and Remark], the operators $F \to K_1 * F$, $g \to K_2 * g$ possess the following properties.

Lemma 1. *For any n-dimensional case with $n \ge 2$*

(1) *the function $h_1 = K_1 * F$ satisfies the following estimates*:

$$h_1 \ge 0 \ \text{ in } \ \Omega \qquad \text{if } F \ge 0 \ \text{a.e. in } \Omega, \tag{2.3}$$

$$||h_1||_{W_p^2(\Omega)} \le C||F||_{L_p(\Omega)} \qquad \text{if } 1 < p < \infty, \tag{2.4}$$

$$||h_1||_{C^{1+\alpha}(\overline{\Omega})} \le C||F||_{L_p(\Omega)} \quad \alpha = 1 - \frac{n}{p} \quad \text{if } n < p < \infty, \tag{2.5}$$

$$||h_1||_{C^{1+\alpha}(\overline{\Omega})} \le C||F||_{L_\infty(\Omega)} \ \ \forall\, \alpha \in (0,1), \tag{2.6}$$

$$||h_1||_{W_1^1(\Omega)} \le C||F||_{L_1(\Omega)}. \tag{2.7}$$

(2) *the function $h_2 = K_2 * g$ satisfies the following estimates*:

$$||h_2||_{C^{1+\alpha}(\overline{\Omega})} \le C||g||_{C^{\alpha}(\Gamma)} \qquad \text{if } \ 0 < \alpha < 1, \tag{2.8}$$

$$||h_2||_{C^{l}(\overline{\Omega}')} \le C||g||_{L_1(\Gamma)}, \tag{2.9}$$

for any Ω', such that $\overline{\Omega}' \subset \Omega$ and any $l \ge 0$.

3. Construction of approximate solutions. Let $\breve{\omega}^{\varepsilon}$, a^{ε} for $\varepsilon > 0$ be smooth approximations of the functions $\breve{\omega}$, a, such that

$$\begin{aligned} 0 \le & \breve{\omega}^{\varepsilon}(\vec{x},t) \le \aleph && \text{for} \quad (\vec{x},t) \in \overline{\Omega}_T, \\ 0 < & a^{\varepsilon}(\vec{x},t) && \text{for} \quad (\vec{x},t) \in \Gamma_T^{+}, \\ 0 = & a^{\varepsilon}(\vec{x},t) && \text{for} \quad (\vec{x},t) \in \Gamma_T^{0}, \\ & a^{\varepsilon}(\vec{x},t) < 0 && \text{for} \quad (\vec{x},t) \in \Gamma_T^{-}. \end{aligned} \tag{3.1}$$

We require also that

$$\begin{cases} \breve{\omega}^{\varepsilon}(\vec{x},t) \underset{\varepsilon \to 0}{\longrightarrow} \breve{\omega}(\vec{x},t) & \text{a.e. in } \overline{\Omega}_T, \\ \\ ||a^{\varepsilon} - a\,||_{C^{\eta,\theta}(\Gamma_T)} \underset{\varepsilon \to 0}{\longrightarrow} 0 \end{cases} \tag{3.2}$$

and

$$\begin{aligned} \int_0^T \{ & ||\partial_t(\breve{\omega}^{\varepsilon} - \breve{\omega})||_{L_1(\Omega)} + ||\triangle\,(\breve{\omega}^{\varepsilon} - \breve{\omega})||_{L_1(\Omega)} \\ & + ||\bigtriangledown(\breve{\omega}^{\varepsilon} - \breve{\omega})||_{L_\infty(\Omega)}\}\, dt \underset{\varepsilon \to 0}{\longrightarrow} 0. \end{aligned} \tag{3.3}$$

Let us fix a positive number R. Next we construct the pair $\{\omega_{\varepsilon,R}, h_{\varepsilon,R}\}$ as a solution of auxiliary **Problem $\mathbf{P}_{\varepsilon,\mathbf{R}}$**. For the sake of simplicity, in this section we suppress the dependence $\omega_{\varepsilon,R}, h_{\varepsilon,R}$ on ε, R and write ω and h.

Let us consider the problem which is a coupling of the following two systems.

Problem $\mathbf{P}_{\varepsilon,\mathbf{R}}$. *Find $\omega \in W_2^{2,1}(\Omega_T)$, satisfying the system*

$$(3.4)\quad \begin{cases} \omega_t + \operatorname{div}(\omega\vec{v}) = \varepsilon\Delta\omega \quad \text{and} \quad \vec{v} = -\nabla h \qquad \text{for} \quad (\vec{x},t) \in \Omega_T, \\ \omega(\vec{x},t) = \breve{\omega}^{\varepsilon}(\vec{x},t),\ (\vec{x},t) \in \Gamma_T, \qquad \omega(\vec{x},0) = \breve{\omega}^{\varepsilon}(\vec{x},0),\ \vec{x} \in \Omega \end{cases}$$

and find $h \in W_2^2(\Omega)$, satisfying the system

$$(3.5)\quad \begin{cases} -\Delta h + h = [\omega]_R, \quad (\vec{x},t) \in \Omega_T, \\ -\nabla h\,\vec{n} = a^{\varepsilon}(\vec{x},t), \quad (\vec{x},t) \in \Gamma_T, \end{cases}$$

where $[\cdot]_R$ is the cut-off function defined as $[\phi]_R := \max\{0,\ \min\{R,\phi\}\}$.

To prove the solvability of **Problem $\mathbf{P}_{\varepsilon,\mathbf{R}}$** we use the Schauder fixed point argument. Let us introduce the class of functions

$$(3.6)\qquad \mathcal{M} = \{\omega(\vec{x},t) \in C(0,T;\ L_2(\Omega)) : ||\omega||_{C(0,T;\ L_2(\Omega))} \le M\},$$

where an exact value of M will be determined below. First we define the operator T_1, which transforms a "fixed" vorticity into the corresponding superconductive field

$$(3.7)\qquad \mathcal{M} \ni \widetilde{\omega} \mapsto T_1[\widetilde{\omega}] = h$$

as the solution of (3.5), where instead of ω we put the chosen $\widetilde{\omega}$. By (2.1), (2.2) the solution h can be represented in the form

$$(3.8)\qquad h(\vec{x},t) = (K_1 * [\widetilde{\omega}]_R)(\vec{x},t) + (K_2 * a^{\varepsilon})(\vec{x},t)$$

for a.e. $(\vec{x},t) \in \Omega_T$, and by (2.5), (2.8) of Lemma 1 we derive the estimate

$$(3.9)\qquad \begin{aligned} ||h||_{L_\infty(0,T;\ C^1(\overline{\Omega}))} &\le C(||\,[\widetilde{\omega}]_R\,||_{L_\infty(\Omega_T)} + ||a^{\varepsilon}||_{L_\infty(0,T;\ C^{\eta}(\Gamma))}) \\ &\le C(R+1), \end{aligned}$$

where the constant C depends on the datum a^{ε} being independent of ε, R. The second operator T_2 describes the evolution of the vorticity

$$(3.10)\qquad \vec{v} = -\nabla h \mapsto T_2[h] = \omega,$$

where ω is the solution of (3.4) for the found "superconductive field" h in (3.7). Taking into account (3.9) and the results of [14, Theorem 4.1, p. 153 and Theorem 4.2, p. 160], there exists a unique weak solution ω of (3.4) such that

$$(3.11)\qquad ||\omega||_{V_2^{1,0}(\Omega_T)} \le C_*,$$

$$(3.12)\qquad ||\omega(\cdot,t_1) - \omega(\cdot,t_2)||_{L_2(\Omega)} \le \phi_*(|t_1 - t_2|) \quad \forall t_1, t_2 \in [0,T],$$

where the constant C_* and the function $\phi_* = \phi_*(t) \in C([0,T]),\ \phi_*(0) = 0$ depend on ε, R and the data $\breve{\omega}^{\varepsilon},\ a^{\varepsilon}$. Setting $M := C_*$ in (3.6) and $T = T_2 \circ T_1$, by (3.6), (3.11), (3.12), we see that T maps the bounded set $\mathcal{M}$ into a compact subset of $\mathcal{M}$.

In order to apply the Schauder fixed point theorem we need to prove that the operator T is continuous. Let $\widetilde{\omega}_n, \widetilde{\omega} \in \mathcal{M}$ and $\widetilde{\omega}_n \to \widetilde{\omega}$ in $C(0,T;\ L_2(\Omega))$. From (3.8) and (2.5) of Lemma 1, it follows that for any $p > n$

$$\begin{aligned} ||\nabla h_n - \nabla h||_{C^{\alpha}(\overline{\Omega})} &\le C||[\widetilde{\omega}_n]_R - [\widetilde{\omega}]_R||_{L_p(\Omega)} \\ \le C\cdot(2R)^{\frac{p-2}{p}}\,||[\widetilde{\omega}_n]_R - [\widetilde{\omega}]_R||^{2/p}_{L_2(\Omega)} &\le C(R)||\widetilde{\omega}_n - \widetilde{\omega}||^{2/p}_{L_2(\Omega)} \overset{n\to\infty}{\longrightarrow} 0, \end{aligned}$$

where the constant $C(R)$ depends on R being independent of ε and the index n. Hence

$$||\nabla h_n - \nabla h||_{C(0,T;\, C^\alpha(\overline{\Omega}))} \to 0, \tag{3.13}$$

where h_n and h are the solutions of (3.5) with ω replaced by $\widetilde{\omega}_n$ and $\widetilde{\omega}$, respectively. Next we consider $\omega_n = T[\widetilde{\omega}_n]$ and $\omega = T[\widetilde{\omega}_n]$. Using (3.13) and the results of [14] we conclude that

$$\max_{t\in[0,T]} ||\omega_n(\cdot,t) - \omega(\cdot,t)||_{L_2(\Omega)} + ||\nabla(\omega_n - \omega)||_{L_2(\Omega_T)} \to 0.$$

Therefore we conclude that the sequence ω_n itself converges to ω and the continuity of the operator T is proved.

Hence there exists a fixed point ω, such that $\omega = T[\omega]$. We now show the following lemma.

Lemma 2. *There exists at least one solution* $\{\omega, h\}$ *of the systems* (3.4)–(3.5)*, such that for some* $\alpha \in (0,1)$

$$\omega \in W_2^{1,0}(\Omega_T) \cap C(0,T;\, L_2(\Omega)), \qquad h \in C(0,T;\, C^{1+\alpha}(\overline{\Omega})).$$

By the theory of parabolic and elliptic equations the constructed functions ω, h have a better regularity. In fact we deduce the following result.

Theorem 2. *For fixed* $\varepsilon, R > 0$*, there exists a unique pair of functions*

$$\omega \in W_2^{2,1}(\Omega_T) \cap C^{\alpha,\alpha/2}(\Omega_T), \qquad h \in C^{2+\alpha,\alpha/2}(\Omega_T) \tag{3.14}$$

for some $\alpha \in (0,1)$*, which is the solution of* ***Problem*** $\mathbf{P}_{\varepsilon,\mathbf{R}}$.

Proof. Because of (3.1), the function $\breve{\omega}^\varepsilon$ is bounded in $\overline{\Omega}_T$ by the constant $\aleph$. Applying Theorem 7.1, p. 181 of [14], we have

$$||\omega||_{L_\infty(\Omega_T)} \le C(R,\varepsilon)\,\aleph, \tag{3.15}$$

where the constant $C(R,\varepsilon)$ depends on R, ε. Hence ω is the solution of the equation

$$\omega_t - \varepsilon\Delta\omega = \mathcal{F}(\vec{x},t), \quad (\vec{x},t) \in \Omega_T$$

with $\mathcal{F} = -\nabla\omega\,\vec{v} - \omega([\omega]_R - h) \in L_2(\Omega_T)$ by (3.9), (3.11), (3.15). Using Theorem 6.1, p. 178 of [14], we deduce $\omega(\vec{x},t) \in W_2^{2,1}(\Omega_T)$ and also by Theorem 10.1, p. 204 of [14], we have $\omega \in C^{\alpha,\alpha/2}(\Omega_T)$ for some $\alpha \in (0,1)$. Moreover, by the theory of elliptic equations [15], we conclude that $h \in C^{2+\alpha,\alpha/2}(\Omega_T)$.

The uniqueness of the solution $\{\omega, h\}$ for **Problem** $\mathbf{P}_{\varepsilon,\mathbf{R}}$ follows in the usual way. Let $\omega_i, h_i, i = 1,2$, be different solutions of **Problem** $\mathbf{P}_{\varepsilon,\mathbf{R}}$ and $\omega = \omega_1 - \omega_2, h = h_1 - h_2$. Then the pair $\{\omega, h\}$ satisfies

$$\begin{cases} \omega_t - \varepsilon\Delta\omega = \operatorname{div}(\nabla h_1 \omega + \omega_2 \nabla h), & (\vec{x},t) \in \Omega_T, \\ \omega(\vec{x},t) = 0, \ \ (\vec{x},t) \in \Gamma_T, & \omega(\vec{x},0) = 0, \ \ \vec{x} \in \Omega, \end{cases}$$

$$\begin{cases} -\Delta h + h = [\omega_1]_R - [\omega_2]_R, & (\vec{x},t) \in \Omega_T, \\ -\nabla h\, \vec{n} = 0, & (\vec{x},t) \in \Gamma_T. \end{cases}$$

Multiplying the first equation by ω, the second one by h, and integrating them over Ω, we obtain the following relations:

$$\frac{1}{2}\frac{d}{dt}\int_\Omega \omega^2\, d\vec{x} + \varepsilon\int_\Omega |\nabla\omega|^2\, d\vec{x} = -\int_\Omega (\nabla h_1\omega + \omega_2\nabla h)\,\nabla\omega\, d\vec{x},$$
$$\int_\Omega \left(|\nabla h|^2 + |h|^2\right)\, d\vec{x} = \int_\Omega ([\omega_1]_R - [\omega_2]_R)\, h\, d\vec{x}.$$

Applying the property of *the cut-off function*, that $|[\omega_1]_R - [\omega_2]_R| \le |\omega_1 - \omega_2|$, and Cauchy's inequality, we come to the inequality

$$\frac{d}{dt}\int_\Omega \omega^2\, d\vec{x} + \varepsilon\int_\Omega |\nabla\omega|^2\, d\vec{x} \le C\int_\Omega \omega^2\, d\vec{x} \quad \text{and} \quad \int_\Omega \omega^2(\vec{x},0)\, d\vec{x} = 0$$

with some constant $C = C(\varepsilon, \|\nabla h_1\|_{L_\infty(\Omega_T)}, \|\omega_2\|_{L_\infty(\Omega_T)})$. Application of the standard Gronwall inequality completes the proof of the uniqueness of solution. □

4. A priori estimates independent of $\boldsymbol{\varepsilon, R}$. In this section we derive a priori estimates of the solution $\{\omega_{\varepsilon,R},\ h_{\varepsilon,R}\}$ for **Problem $\mathbf{P}_{\varepsilon,\mathbf{R}}$**, which do not depend on ε, R.

In sections 4, 5, and 6, all constants C, which are used in a priori estimates, do not depend on ε and R.

4.1. Maximum-minimum principle for $\boldsymbol{\omega}$ and $\boldsymbol{h}$. In this section, we show that the solution $\{\omega_{\varepsilon,R},\ h_{\varepsilon,R}\}$ is bounded in $L_\infty(\Omega_T)$, independently of ε, R. Throughout the section, for simplicity of presentation, we continue to suppress the dependence of $\omega_{\varepsilon,R}$ and $h_{\varepsilon,R}$ on ε, R and write ω and h.

The proof of boundedness of the functions ω, h is divided into a few lemmas. First let us show the positivity of ω.

LEMMA 3. *For all $(\vec{x},t) \in \Omega_T$*

$$\omega(\vec{x},t) \ge 0. \tag{4.1}$$

Proof. By (3.14) we have

$$\sup_{\vec{x}\in\Omega}\left|\operatorname{div}\vec{v}(\vec{x},t)\right| \le \lambda(t) = \max_{\vec{x}\in\Omega}\left|[\omega(\vec{x},t)]_R - h(\vec{x},t)\right| \in C(0,T). \tag{4.2}$$

Let us denote $\omega_- = \min(\omega, 0)$. Then using (3.1), the first equation of (3.4), and the boundary condition $\omega_-(\vec{x},t) = 0$ for all $(\vec{x},t) \in \Gamma_T$, it is easy to verify that the function ω_- satisfies the inequality

$$\frac{d}{dt}\int_\Omega \omega_-^2\, d\vec{x} + \varepsilon\int_\Omega \left|\nabla\omega_-\right|^2\, d\vec{x} = -\int_\Omega \operatorname{div}\vec{v}\,\omega_-^2\, d\vec{x} \le \lambda(t)\int_\Omega \omega_-^2\, d\vec{x}.$$

Since $\omega_-(\vec{x},0) = 0$ for all $\vec{x} \in \Omega$, from the Gronwall inequality we have $\omega_- = 0$ for all $(x,t) \in \Omega_T$, which implies (4.1). □

Now we show that ω is bounded in the space $L_1(\Omega)$.

LEMMA 4. *There exists a constant Υ_0, independent of R, ε, such that*

$$||\omega(\cdot,t)||_{L_1(\Omega)} \le \Upsilon_0 \quad \forall t \in [0,T]. \tag{4.3}$$

Proof. The function $z = \omega - \breve{\omega}^\varepsilon$ satisfies the problem

$$\begin{cases} z_t + \operatorname{div}(z\vec{v}) = \varepsilon\Delta z + F, & (\vec{x},t) \in \Omega_T, \\ z|_{\Gamma_T} = 0,\ (\vec{x},t) \in \Gamma_T, & z|_{t=0} = 0,\ \vec{x} \in \Omega \end{cases} \tag{4.4}$$

with

$$F = \varepsilon \triangle \breve{\omega}^\varepsilon - \partial_t \breve{\omega}^\varepsilon - \operatorname{div}(\vec{v}\,\breve{\omega}^\varepsilon). \tag{4.5}$$

By (3.1)–(3.3) and (3.5), we have

$$\begin{aligned} &||\operatorname{div}(\vec{v}\,\breve{\omega}^\varepsilon)(\cdot,t)||_{L_1(\Omega)} \le ||\breve{\omega}^\varepsilon||_{L_\infty(\Omega_T)}\,||[\omega]_R - h||_{L_1(\Omega)} \\ &+||\nabla\breve{\omega}^\varepsilon(\cdot,t)||_{L_\infty(\Omega)}\,||\vec{v}||_{L_1(\Omega)} \le \lambda(t)(||[\omega]_R - h||_{L_1(\Omega)} + ||\vec{v}||_{L_1(\Omega)}) \end{aligned} \tag{4.6}$$

with $\lambda(t) \in L_1(0,T)$. Using (3.8) and the estimates (2.7), (2.8) of Lemma 1, we derive that for a.e. $t \in [0,T]$,

$$\begin{aligned} &\left.\begin{matrix} ||h(\cdot,t)||_{L_1(\Omega)} \\ \\ ||\vec{v}(\cdot,t)||_{L_1(\Omega)} \end{matrix}\right\} \le C(||\omega(\cdot,t)||_{L_1(\Omega)} + ||a^\varepsilon(\cdot,t)||_{C^\eta(\Gamma)}) \\ &\qquad\qquad\qquad\qquad \le C||z(\cdot,t)||_{L_1(\Omega)} + C. \end{aligned} \tag{4.7}$$

Therefore, according to (3.2), (3.3), we deduce that

$$||F(\cdot,t)||_{L_1(\Omega)} \le \lambda(t)\left(||z(\cdot,t)||_{L_1(\Omega)} + 1\right) \quad \text{with} \quad \lambda(t) \in L_1(0,T), \tag{4.8}$$

where $\lambda(t)$ does not depend on ε, R. Multiplying (4.4) by $sgn_\delta z := \frac{z}{\sqrt{z^2+\delta}}$ with some $\delta \in (0,1)$, we obtain that

$$\begin{aligned} &\partial_t((z^2+\delta)^{1/2}) + \operatorname{div}(\vec{v}\,(z^2+\delta)^{1/2}) - \frac{\delta}{(z^2+\delta)^{1/2}}\operatorname{div}\vec{v} \\ &= \varepsilon \operatorname{div}\left(\nabla z \frac{z}{(z^2+\delta)^{1/2}}\right) - \varepsilon\delta \frac{|\nabla z|^2}{(z^2+\delta)^{3/2}} + F\, sgn_\delta z. \end{aligned} \tag{4.9}$$

Taking into account that $z = 0$ on Γ_T and integrating this equality over Ω, we have

$$\frac{d}{dt}\left(||\sqrt{z^2+\delta}||_{L_1(\Omega)}\right) + I_\delta(t) \le \lambda(t)\ \left(||z||_{L_1(\Omega)} + 1\right) \text{ for a.e. } t \in [0,T],$$

with $I_\delta(t) = \int_\Omega [\operatorname{div}(\vec{v}\sqrt{z^2+\delta}) - \frac{\delta}{(z^2+\delta)^{1/2}}\operatorname{div}\vec{v}]\,d\vec{x}$. Since

$$|I_\delta| \le \delta^{1/2}\left\{\int_\Omega (|\omega| + |h|)\,d\vec{x} + \left|\int_\Gamma a^\varepsilon\,d\vec{x}\right|\right\} \underset{\delta\to 0}{\longrightarrow} 0,$$

we deduce the inequality

$$\frac{d}{dt}||z||_{L_1(\Omega)} \le \lambda(t)\,(||z||_{L_1(\Omega)} + 1), \text{ a.e. } t \in [0,T],$$

where the function $\lambda(t)$ does not depend on ε, R. Hence applying the Gronwall inequality, we obtain $||z(\cdot,t)||_{L_1(\Omega)} \le C$ for $t \in [0,T]$ with the constant C independent of ε, R. This immediately implies the estimate (4.3). □

Now we prove an auxiliary lemma.

LEMMA 5. *There exists a positive constant R_*, such that*

$$\max_{\overline{\Omega}_T} \omega(\vec{x},t) \le \max\left\{\max_{\overline{\Omega}_T} h(\vec{x},t),\ \aleph\right\} \tag{4.10}$$

for any fixed $R \geq R_*$. The constant R_* depends on $\Upsilon_0, \Gamma, \Omega, ||a||_{L_\infty(0;T;\, C^\eta(\Gamma))}$ and $\aleph$ being independent of ε.

Proof. (A) If ω attains its maximum on the parabolic boundary Γ_T, then (4.10) is obvious.

(B) If ω attains a positive maximum $\overline{\omega} = \omega(\vec{x}_0, t_0)$ inside Ω_T, then we come to the relation

$$0 \leq \omega_t(\vec{x}_0, t_0) - \varepsilon \bigtriangleup \omega(\vec{x}_0, t_0) = \overline{\omega}(h(\vec{x}_0, t_0) - [\overline{\omega}]_R) \tag{4.11}$$

and hence

$$[\overline{\omega}]_R = \min\{R, \overline{\omega}\} \leq \max_{\overline{\Omega}_T} h(\vec{x}, t).$$

Let us consider two possibilities.

(B.1) If $\overline{\omega} \leq R$, we have $[\overline{\omega}]_R = \overline{\omega} \leq \max_{\overline{\Omega}_T} h(\vec{x}, t)$, and therefore (4.10) is proved.

(B.2) If $\overline{\omega} > R$, we have

$$R \leq \max_{\overline{\Omega}_T} h(\vec{x}, t). \tag{4.12}$$

Let us prove that this case is impossible for large enough R. Using the representation (3.8), Lemma 4, and (2.3), (2.5), (2.8) of Lemma 1, we have

$$\begin{aligned}
\max_{\overline{\Omega}} |h| &\leq \max_{\overline{\Omega}} |K_1 * ([\omega]_R{}^{1/q}[\omega]_R{}^{1-1/q})| \;+\; \max_{\overline{\Omega}} |h_2| \\
&\leq C(q, \Omega)\; ||\omega||_{L_1(\Omega)}\; R^{1-1/q} \;+\; C(\Gamma)||a^\varepsilon||_{C^\eta(\Gamma)} \\
&\leq C(q, \Omega)\; \Upsilon_0\; R^{1-1/q} \;+\; C(\Gamma)||a^\varepsilon||_{C^\eta(\Gamma)}
\end{aligned}$$

with $n < q < \infty$. The previous inequality implies that

$$\max_{\overline{\Omega}_T} |h| \leq C_* \left(\max_{\overline{\Omega}_T} R^{1-1/q} + 1 \right)$$

with some constant $C_* = C(q, \Omega, \Gamma, \Upsilon_0, ||a||_{L_\infty(0;T;\, C^\eta(\Gamma))})$, which does not depend on ε, R. Since $0 < 1 - 1/q < 1$, there exists a constant R_*, depending on C_*, such that $C_*(\, R^{1-1/q} + 1) \leq \frac{R}{2}$ for any $R \geq R_*$. Therefore we conclude that

$$\max_{\overline{\Omega}_T} |h| \leq \frac{R}{2} \quad \text{for any} \quad R \geq R_*.$$

But this inequality contradicts with (4.12). □

Remark 3. Let us note that if we consider that the system (3.4)–(3.5) with the right side of (3.5) equals ω, instead of $[\omega]_R$, then the desired assertion of Lemma 5 follows immediately from an analog of the relation (4.11),

$$0 \leq \omega_t(\vec{x}_0, t_0) - \varepsilon \bigtriangleup \omega(\vec{x}_0, t_0) = \overline{\omega}(h(\vec{x}_0, t_0) - \overline{\omega}). \tag{4.13}$$

Now we are able to show the boundedness of $\{\omega, h\}$ in $L_\infty(\Omega_T)$.

LEMMA 6. *There exist constants Υ_1, Υ_2 depending on the data $\breve{\omega}$, a, but independent of R, ε, such that*

$$||\omega||_{L_\infty(\Omega_T)} \leq \Upsilon_1, \tag{4.14}$$

$$||h||_{L_\infty(\Omega_T)},\; ||\nabla h||_{L_\infty(\Omega_T)} \leq \Upsilon_2. \tag{4.15}$$

Proof. Using the representation (3.8), Lemma 4, and (2.3), (2.5), (2.8) of Lemma 1, we have

$$\max_{\overline{\Omega}} |h| \le \max_{\overline{\Omega}} |K_1 * ([\omega]_R{}^{1/q}[\omega]_R{}^{1-1/q})| \; + \; \max_{\overline{\Omega}} |h_2|$$
$$\le C(q,\Omega)\ \Upsilon_0\ \max_{\overline{\Omega}}\ \omega^{1-1/q} \; + \; C(\Gamma)||a^\varepsilon||_{C^\eta(\Gamma)}$$

with $n < q < \infty$. The previous inequality implies that

$$\max_{\overline{\Omega}_T} |h| \le C\left(\max_{\overline{\Omega}_T} \omega^{1-1/q} + ||a||_{L_\infty(0,T;\ C^\eta(\Gamma))}\right) \tag{4.16}$$

with some constant $C = C(q,\Omega,\Gamma,\Upsilon_0)$, which does not depend on ε, R. Joining (4.10) and (4.16), we obtain the inequality

$$\max_{\overline{\Omega}_T} \omega \le C \max\left\{\max_{\overline{\Omega}_T} \omega^{1-1/q} + ||a||_{L_\infty(0,T;\ C^\eta(\Gamma))},\ \aleph\right\}, \tag{4.17}$$

which leads to (4.14), because $0 < 1-1/q < 1$. The estimates (4.15) follow immediately from (2.5), (2.8) of Lemma 1. This completes the proof. □

Choosing $R := \max\{R_*, \Upsilon_1\}$ in (3.5), we see that the cut-off function $[\cdot]_R$ in (3.5) can be omitted. In the following, we consider the solution of (3.4), (3.5) as ω_ε, h_ε, depending only on ε (not R).

4.2. Estimates of derivatives. From this section we shall write ω_ε, h_ε, and $\vec{v}_\varepsilon = -\bigtriangledown h_\varepsilon$. Let us show the following lemma.

LEMMA 7. *There exist constants C independent of ε, such that*

$$||\sqrt{\varepsilon}\nabla\omega_\varepsilon||_{L_2(\Omega_T)} \le C, \tag{4.18}$$
$$||\partial_t(\omega_\varepsilon)||_{L_2(0,T;\ H^{-1}(\Omega))} \le C, \tag{4.19}$$
$$||h_\varepsilon||_{L_\infty(0,T;\ C^{1+\eta}(\overline{\Omega}))} \le C, \tag{4.20}$$
$$||h_\varepsilon(\cdot,t)||_{L_\infty(0,T;\ W^2_p(\overline{\Omega'}))} \le C, \tag{4.21}$$

for any $p \in [1,\infty)$ and for any Ω', such that $\overline{\Omega'} \subset \Omega$.

Proof. According to Lemma 6, we have that

$$||h_\varepsilon||_{L_\infty(\Omega_T)},\ ||\nabla h_\varepsilon||_{L_\infty(\Omega_T)},\ ||\omega_\varepsilon||_{L_\infty(\Omega_T)} \le \Upsilon_3, \tag{4.22}$$

where Υ_3 is independent of ε. Multiplying (4.4) by $z_\varepsilon = \omega_\varepsilon - \breve{\omega}^\varepsilon$ and using (3.1), (4.22), $\operatorname{div}\vec{v}_\varepsilon = \omega_\varepsilon - h_\varepsilon$, we easily get

$$\frac{1}{2}\frac{d}{dt}||z_\varepsilon||^2_{L_2(\Omega)} + \varepsilon||\nabla z_\varepsilon||_{L_2(\Omega)}$$
$$\le -\int_\Omega \operatorname{div}\vec{v}_\varepsilon \frac{z_\varepsilon^2}{2}\, d\vec{x} + \int_\Omega F_\varepsilon\ z_\varepsilon\ d\vec{x} \le C + C||F_\varepsilon||_{L_1(\Omega)}.$$

Integrating on $(0,T)$ and taking into account (1.12), (3.3), we have

$$||\sqrt{\varepsilon}\nabla z_\varepsilon||_{L_2(\Omega_T)} \le C, \tag{4.23}$$

which implies (4.18).

Let us choose an arbitrary function $\phi \in H^1(\Omega_T)$, such that $\phi(\vec{x},T) = \phi(\vec{x},0) = 0$ for all $\vec{x} \in \Omega$ and $\phi(\vec{x},t) = 0,\ (\vec{x},t) \in \Gamma_T$. Multiplying the first equation in (3.4) by ϕ and integrating it over Ω_T, from (4.22)–(4.23) we obtain

$$\left| \int\limits_{\Omega_T} \omega_\varepsilon\, \phi_t\ d\vec{x}dt \right| \leq C\, \|\phi\|_{L_2(0,T;\ H_0^1(\Omega))},$$

which gives (4.19).

Taking into account the representation (3.8), we have

$$h_\varepsilon(\vec{x},t) = (K_1 * \omega_\varepsilon)(\vec{x},t) + (K_2 * a^\varepsilon)(\vec{x},t). \tag{4.24}$$

Hence by (2.6), (2.8) of Lemma 1 we have for all $t \in (0,T)$,

$$||h_\varepsilon(\cdot,t)||_{C^{1+\eta}(\overline{\Omega})} \leq C\left(||\omega_\varepsilon(\cdot,t)||_{L_\infty(\Omega)} + ||a^\varepsilon(\cdot,t)||_{C^\eta(\Gamma)}\right)$$

and by (2.4), (2.9) of Lemma 1

$$||h_\varepsilon(\cdot,t)||_{W_p^2(\overline{\Omega'})} \leq C\left(||\omega_\varepsilon(\cdot,t)||_{L_\infty(\Omega)} + ||a^\varepsilon(\cdot,t)||_{L_1(\Gamma)}\right)$$

for all $p \in [1,\infty)$ and for any Ω' such that $\overline{\Omega'} \subset \Omega$. With the help of (1.9), (3.2), and (4.22), we derive the assertions (4.20), (4.21). $\square$

5. Limit transition. In this section we prove Theorem 2.

From (4.21), (4.22) we conclude that there exists a subsequence of $\{\omega_\varepsilon, h_\varepsilon\}$ such that

$$\begin{aligned} h_\varepsilon &\rightharpoonup h \text{ weakly} - * \text{ in } L_\infty(\Omega_T) \cap L_\infty(0,T;\, W_p^2(\overline{\Omega'})), \\ \nabla h_\varepsilon &\rightharpoonup \nabla h \text{ weakly} - * \text{ in } L_\infty(\Omega_T), \end{aligned} \tag{5.1}$$

for any $p \in (1,\infty)$ and for any Ω', such that $\overline{\Omega}' \subset \Omega$. By results obtained in [1], [18], the compact embedding of $L_2(\Omega)$ in $H^{-1}(\Omega)$ and (4.19) imply

$$\omega_\varepsilon \to \omega \quad \text{strongly in} \quad L_2(0,T;\ H^{-1}(\Omega)). \tag{5.2}$$

In view of (4.18), (4.22), we have

$$\begin{aligned} \omega_\varepsilon &\rightharpoonup \omega \quad \text{weakly} - * \quad \text{in } L_\infty(\Omega_T), \\ \varepsilon\nabla\omega_\varepsilon &\rightharpoonup 0 \quad \text{weakly} \qquad \text{in } L_2(\Omega_T). \end{aligned} \tag{5.3}$$

With the help of (3.2), (5.1), (5.3), the limit transition in the representation (4.24) on $\varepsilon \to 0$ yields

$$h(\vec{x},t) = (K_1 * \omega)(\vec{x},t) + (K_2 * a)(\vec{x},t) \quad \text{for a.e. } (\vec{x},t) \in \Omega_T. \tag{5.4}$$

From the elliptic theory [15], [20] and (2.6), (2.8) of Lemma 1, the function h satisfies (1.14) with the boundary conditions (1.16), such that

$$h \in L_\infty(0,T;\, C^{1+\eta}(\overline{\Omega})). \tag{5.5}$$

Definition 2. *Let the distance between any given point $\vec{x} \in \mathbb{R}^n$ and any subset $A \subseteq \mathbb{R}^n$ be defined by $d(\vec{x},A) := \inf_{\vec{y}\in A}|\vec{x}-\vec{y}|$. Let $d = d(\vec{x})$ be the distance function on Γ, defined by*

$$d(\vec{x}) := d(\vec{x}, \mathbb{R}^n \backslash \Omega) - d(\vec{x},\Omega) \quad \textit{for any } \vec{x} \in \mathbb{R}^n.$$

Let A be an arbitrary subdomain of Ω. We also introduce the distance between A and the boundary Γ by $d(A,\Gamma) := inf_{\vec{y}\in A} d(\vec{y})$.

The set of all points of $\overline{\Omega}$, whose distance to Γ (respectively, to Γ^- and to Γ^+) is less than σ, is denoted by $U_\sigma(\Gamma)$ (respectively, $U_\sigma(\Gamma^-)$ and $U_\sigma(\Gamma^+)$). Since $\Gamma \in C^{2+\gamma}$, the function $d = d(\vec{x})$ belongs to C^2 in a neighborhood $U_{\sigma_0}(\Gamma)$ of Γ for some $\sigma_0 > 0$. For $0 < 2\sigma < \sigma_0$, we introduce the approximation of the unit function for all $\vec{x} \in \overline{\Omega}$ by

$$\mathbf{1}_\sigma(\vec{x}) := \begin{cases} 1 & \text{if } \vec{x} \in \Omega \backslash U_{2\sigma}(\Gamma), \\ \frac{d-\sigma}{\sigma} & \text{if } \sigma < d(\vec{x}) < 2\sigma, \\ 0 & \text{if } 0 \le d(\vec{x}) < \sigma. \end{cases} \tag{5.6}$$

Now we show an auxiliary lemma, playing the crucial role in the proof that the function ω satisfies the boundary condition (1.5) in the sense of the equality (1.13).

LEMMA 8. *For any positive $\psi \in C^{1,1}(\Omega_T)$, such that* $\operatorname{supp}(\psi) \subset \Omega_T \cup \Gamma_T^-$, *and $\psi(\vec{x},T) = 0,\ \vec{x} \in \Omega$, we have*

$$\lim_{\sigma\to 0} \left(\overline{\lim_{\varepsilon\to 0}} \frac{1}{\sigma} \int_0^T \int_{[\sigma<d<2\sigma]} |\omega_\varepsilon - \breve{\omega}_\varepsilon|\ \vec{v}_\varepsilon \bigtriangledown d\ \psi\ d\vec{x}dt \right) = 0. \tag{5.7}$$

Proof. Taking into account that $z_\varepsilon = \omega_\varepsilon - \breve{\omega}_\varepsilon = 0$ on Γ_T and multiplying the equality (4.9) by an arbitrary nonnegative function $\eta \in H^1(\Omega_T)$ with $\eta(\vec{x},T) = 0$, $\vec{x} \in \Omega$, we derive that

$$\begin{aligned} -\int\int_{\Omega_T} \eta_t (z_\varepsilon^2+\delta)^{1/2}\, d\vec{x}dt + \int_\Omega \eta(\vec{x},0)(z_\varepsilon^2(\vec{x},0)+\delta)^{1/2}\, d\vec{x} \\ + \int_{\Gamma_T} a^\varepsilon \sqrt{\delta}\,\eta\, d\vec{x}dt - \int\int_{\Omega_T} (z_\varepsilon^2+\delta)^{1/2}\, \vec{v}_\varepsilon \nabla\eta\, d\vec{x}dt \\ -\int\int_{\Omega_T} \frac{\delta}{(z_\varepsilon^2+\delta)^{1/2}}\,(\omega_\varepsilon - h_\varepsilon)\eta\, d\vec{x}dt = -\varepsilon \int\int_{\Omega_T} \nabla z_\varepsilon \nabla\eta\, \frac{z_\varepsilon}{(z_\varepsilon^2+\delta)^{1/2}}\, d\vec{x}dt \\ -\varepsilon\delta \int\int_{\Omega_T} \frac{|\nabla z_\varepsilon|^2}{(z_\varepsilon^2+\delta)^{3/2}}\,\eta\, d\vec{x}dt + \int\int_{\Omega_T} F_\varepsilon\, sgn_\delta z_\varepsilon\, \eta\, d\vec{x}dt. \end{aligned}$$

Let us denote $G_\varepsilon := |\partial_t \breve{\omega}^\varepsilon| + |\bigtriangledown \breve{\omega}^\varepsilon|$ and $G := |\partial_t \breve{\omega}| + |\bigtriangledown \breve{\omega}|$. Since the functions $\{z_\varepsilon, \omega_\varepsilon, h_\varepsilon\}$ and a^ε are uniformly bounded with respect to ε in $L_\infty(\Omega_T)$ and in $L_\infty(\Gamma_T)$, using (4.5) and (4.18), we obtain

$$\begin{aligned} -\int\int_{\Omega_T} (z_\varepsilon^2+\delta)^{1/2}\, \vec{v}_\varepsilon \nabla\eta\, d\vec{x}dt \le C \int\int_{\Omega_T} (|\eta_t| + |\eta| + |\eta|\,|G_\varepsilon|)\, d\vec{x}dt \\ + C\int_\Omega |\eta(\vec{x},0)|\, d\vec{x} + C\sqrt{\delta} \int_{\Gamma_T} |\eta|\, d\vec{x}dt + C\sqrt{\varepsilon}||\nabla\eta||_{L_2(\Omega_T)}. \end{aligned}$$

Taking $\delta \to 0$, we get

$$\begin{aligned} -\int\int_{\Omega_T} |z_\varepsilon|\ \vec{v}_\varepsilon \nabla\eta\, d\vec{x}dt \le C \int\int_{\Omega_T} (|\eta_t| + |\eta| + |\eta|\,|G_\varepsilon|)\, d\vec{x}dt \\ + C\int_\Omega |\eta(\vec{x},0)|\, d\vec{x} + C\sqrt{\varepsilon}||\nabla\eta||_{L_2(\Omega_T)}. \end{aligned} \tag{5.8}$$

Let us choose an arbitrary function $\psi \in C^{1,1}(\Omega_T)$, satisfying the conditions of the present lemma and take the test function in (5.8). As $\eta_\sigma = (1-\mathbf{1}_\sigma)\,\psi$, we derive that

$$\int\int_{\Omega_T} |z_\varepsilon|\,\psi\,\frac{\vec{v}_\varepsilon \nabla d}{\sigma}\,\chi_{[\sigma<d<2\sigma]}\;d\vec{x}dt$$
$$\leq C\int\int_{\Omega_T}(|\partial_t\eta_\sigma| + |\eta_\sigma| + |\eta_\sigma|\,|G_\varepsilon| + (1-\mathbf{1}_\sigma)|\nabla\psi|)\;d\vec{x}dt$$
$$+\,C\int_\Omega |\eta_\sigma(\vec{x},0)|\;d\vec{x} + C\sqrt{\varepsilon}||\nabla\eta_\sigma||_{L_2(\Omega_T)}.$$

By (4.20) the set $\{\vec{v}_\varepsilon(\cdot,t)\}$ is uniformly continuous on $\overline{\Omega}$, independent of ε and $t \in [0,T]$. Also we have that the function $d(\vec{x}) \in C^2$ in the neighborhood $U_{\sigma_0}(\Gamma)$ and $\bigtriangledown d = -\vec{n}$ on Γ. Hence there exists $\sigma_1 < \sigma_0$, independent of ε and $t \in [0,T]$, such that

$$\left\{\psi\;\vec{v}_\varepsilon \bigtriangledown d\right\}(\vec{x},t) = \begin{cases} >0 & \text{if } (\vec{x},t) \in U_{\sigma_1}(\Gamma^-)\times[0,T], \\ 0 & \text{if } (\vec{x},t)\in U_{\sigma_1}(\Gamma^+\cup\Gamma^0)\times[0,T]. \end{cases} \tag{5.9}$$

Therefore for any $2\sigma \in (0,\sigma_1)$, according to (3.3), we have

$$0 \leq \overline{\lim_{\varepsilon\to 0}}\,\frac{1}{\sigma}\int_0^T\int_{[\sigma<d<2\sigma]}\psi\;|z_\varepsilon|\,\vec{v}_\varepsilon\nabla d\;d\vec{x}dt \leq C\int_\Omega|\eta_\sigma(\vec{x},0)|\;d\vec{x}$$
$$+\,C\int\int_{\Omega_T}(|\partial_t\eta_\sigma| + |\eta_\sigma| + |\eta_\sigma|\,|G| + (1-\mathbf{1}_\sigma)|\nabla\psi|)\;d\vec{x}dt\;d\vec{x}.$$

This implies the property (5.7), since $\eta_\sigma = (1-\mathbf{1}_\sigma)\psi$, $\partial_t\eta_\sigma = (1-\mathbf{1}_\sigma)\psi_t$, and $(1-\mathbf{1}_\sigma) \to 0$ in Ω_T, when $\sigma \to 0$. $\square$

Now we are able to prove that the pair $\{\omega, h\}$ satisfies (1.13). For any $2\sigma \in (0,\sigma_1)$, we put $\eta_\sigma = \mathbf{1}_\sigma\,\psi$, where ψ is an arbitrary function, satisfying the conditions of Lemma 8. Clearly, $\eta_\sigma \in H^1(\Omega_T),\ \eta_\sigma(\vec{x},T) = 0$ for $\vec{x}\in\Omega$, and $\eta_\sigma = 0$ on Γ_T. Multiplying the first equation in (3.4) by η_σ and integrating it over Ω_T, we derive that

$$0 = \left\{\int_{\Omega_T}[\omega_\varepsilon(\psi_t + \vec{v}_\varepsilon\nabla\psi)]\mathbf{1}_\sigma - \varepsilon\nabla\omega_\varepsilon\nabla\eta_\sigma\;d\vec{x}dt + \int_\Omega \breve{\omega}^\varepsilon(\vec{x},0)\,\eta_\sigma(\vec{x},0)\;d\vec{x}\right\}$$
$$+\,\frac{1}{\sigma}\int_0^T\int_{[\sigma<d<2\sigma]}\omega_\varepsilon\,(\vec{v}_\varepsilon\nabla d)\,\psi\;d\vec{x}dt = I^{\varepsilon,\sigma} + J^{\varepsilon,\sigma}. \tag{5.10}$$

Using (5.1), (5.2) and $\vec{v}_\varepsilon\nabla\psi\,\mathbf{1}_\sigma \underset{\varepsilon\to 0}{\rightharpoonup} \vec{v}\nabla\psi\,\mathbf{1}_\sigma$ weakly in $L_2(0,T;\ H_0^1(\Omega))$ with $\vec{v} = -\nabla h$, we have

$$\lim_{\varepsilon\to 0}\int_{\Omega_T}\omega_\varepsilon\,\vec{v}_\varepsilon\nabla\psi\,\mathbf{1}_\sigma\;d\vec{x}dt = \int_{\Omega_T}\omega\,\vec{v}\nabla\psi\,\mathbf{1}_\sigma\;d\vec{x}dt.$$

Because $\mathbf{1}_\sigma \underset{\sigma\to 0}{\longrightarrow} 1$ in Ω_T and Ω, we get

$$\lim_{\sigma\to 0}\left(\lim_{\varepsilon\to 0} I^{\varepsilon,\sigma}\right) = \int_{\Omega_T}\omega(\psi_t + \vec{v}\nabla\psi)\;d\vec{x}dt + \int_\Omega \omega_0\psi(\vec{x},0)\;d\vec{x}.$$

Since

$$J^{\varepsilon,\sigma} = \left[\frac{1}{\sigma}\int_0^T \int_{[\sigma<d<2\sigma]} z_\varepsilon\,(\vec{v}_\varepsilon \nabla d)\,\psi\;d\vec{x}dt\right] + \left[\frac{1}{\sigma}\int_0^T \int_{[\sigma<d<2\sigma]} \breve{\omega}^\varepsilon\,(\vec{v}_\varepsilon \nabla d)\,\psi\;d\vec{x}dt\right] = J_1^{\varepsilon,\sigma} + J_2^{\varepsilon,\sigma},$$

according to Lemma 8, we have

$$\lim_{\sigma\to 0}\,\left(\overline{\lim_{\varepsilon\to 0}}|J_1^{\varepsilon,\sigma}|\right) = 0.$$

By (5.5) the set of functions $\vec{v}_\varepsilon \bigtriangledown d$ is uniformly continuous on $\overline{\Omega}$ for all $t \in [0,T]$, independent of ε, and the trace of $\vec{v}_\varepsilon \bigtriangledown d$ on Γ_T satisfies, in the usual sense,

$$\vec{v}_\varepsilon \bigtriangledown d = -a^\varepsilon \;\text{ for } (\vec{x},t) \in \Gamma_T.$$

By the standard theory of traces and (1.10), (1.12), (3.3), the function $\breve{\omega}_\varepsilon$ has a trace on the boundary Γ_T^-, which converges to b, when $\varepsilon \to 0$, in the space $L_1(0,T;\,L_q(\Gamma^-))$ for any $q \in [1,\infty)$. Therefore the convergences (3.2), (3.3) imply

$$\lim_{\sigma\to 0}\,\left(\lim_{\varepsilon\to 0}\;J_2^{\varepsilon,\sigma}\right) = -\int_0^T \int_{\Gamma_T^-} ab\,\psi\;d\vec{x}dt.$$

Therefore the pair $\{\omega,\,h\}$ satisfies (1.13) for any ψ, satisfying the conditions of Lemma 8. In view of the linearity of (1.13) with respect to ψ, we see that this equation is also fulfilled for any ψ, satisfying the conditions (1.17) of Definition 1.

Let us show that the limit functions $\omega, h, \bigtriangledown h$ admit additional regularity with respect to the time variable t.

Lemma 9. *For any Ω', such that $\overline{\Omega}' \subset \Omega$, we have*

$$\partial_t\,\omega \in L_\infty(0,T;\;H^{-1}(\Omega)), \tag{5.11}$$

$$h,\;\bigtriangledown h \in C^{0,\,\theta}(\overline{\Omega'} \times [0,T]). \tag{5.12}$$

Proof. Let us choose in (1.13) the test function $\phi(\vec{x},t) := \psi(\vec{x})\,\varphi(t)$, such that $\psi(\vec{x}) \in H_0^1(\Omega)$ and $\varphi(t) \in W_1^1([0,T]):\;\; \varphi(0) = \varphi(T) = 0$, which yields

$$\left|\int_0^T \left(\int\limits_\Omega \omega\,\psi\;d\vec{x}\right)\varphi_t dt\right| \le C\,\|\psi\|_{H_0^1(\Omega)}||\varphi||_{L_1(0,T)},$$

which is equivalent to (5.11).

Now we show the Hölder continuity of h with respect to the time $t \in [0,T]$. By the representation (5.4) this can be done for $h_1 := (K_1 * \omega)$ and $h_2 := (K_2 * a)$, separately. Let

$$\rho(s) \in C_0^\infty(\mathbb{R}) \qquad \text{with} \qquad \rho(s) := \begin{cases} 1 \text{ if } & |s| \le 1, \\ 0 \text{ if } & |s| > 2. \end{cases}$$

We introduce the functions $\rho_\sigma(\vec{x}) := \rho(|\vec{x}|/\sigma)$ and $\rho_\sigma^\Gamma(\vec{x}) := \rho(d(\vec{x})/\sigma)$ for $\sigma > 0$. Let Ω' be a subdomain of Ω, such that $\delta := \operatorname{dist}(\overline{\Omega'},\Gamma) > 0$. Then, for any fixed $\vec{x} \in \overline{\Omega'}$

and for any $\sigma < \frac{\delta}{4}$, the function $\bigtriangledown h_1$ can be written in the form

$$\begin{aligned}\nabla h_1(\vec{x},t) &= \int_\Omega \nabla_{\vec{x}} K_1(\vec{x},\vec{y})\ \omega(\vec{y},t)\ d\vec{y}\\ &= \int_\Omega \nabla_{\vec{x}} K_1(\vec{x},\vec{y})\ \rho_\sigma^\Gamma(\vec{y})\ \omega(\vec{y},t)\ d\vec{y}\\ &+ \int_\Omega \nabla_{\vec{x}} K_1(\vec{x},\vec{y})\ \rho_\sigma(\vec{x}-\vec{y})\ \omega(\vec{y},t)\ d\vec{y}\\ &+ \int_\Omega \nabla_{\vec{x}} K_1(\vec{x},\vec{y})\ (1-\rho_\sigma^\Gamma(\vec{y})-\rho_\sigma(\vec{x}-\vec{y}))\ \omega(\vec{y},t)\ d\vec{y}. \end{aligned} \tag{5.13}$$

Hence for any $t_1, t_2 \in [0,T]$ we have that

$$\nabla h_1(\vec{x},t_2) - \nabla h_1(\vec{x},t_1) = I_1 + I_2 + I_3.$$

Using $\omega \in L_\infty(\Omega_T)$, the terms $I_i,\ i = 1,2,3$ are estimated by

$$\begin{aligned}|I_1| &\le ||\rho_\sigma^\Gamma\, \nabla_{\vec{x}} K_1(\vec{x},\cdot)||_{L_1(\Omega)}\ ||\omega(\cdot,t_2)-\omega(\cdot,t_1)||_{L_\infty(\Omega)} \le C\sigma,\\ |I_2| &\le ||\rho_\sigma\, \nabla_{\vec{x}} K_1(\vec{x},\cdot)||_{L_1(\Omega)}\ ||\omega(\cdot,t_2)-\omega(\cdot,t_1)||_{L_\infty(\Omega)} \le C\sigma,\end{aligned} \tag{5.14}$$

and

$$|I_3| \le ||(1-\rho_\sigma^\Gamma-\rho_\sigma)\,\nabla_{\vec{x}} K_1(\vec{x},\cdot)||_{H_0^1(\Omega)} ||\omega(\cdot,t_2)-\omega(\cdot,t_2)||_{H^{-1}(\Omega)}. \tag{5.15}$$

Since the function $(1-\rho_\sigma^\Gamma-\rho_\sigma)\,\nabla_{\vec{x}} K_1(\vec{x},\vec{y}),\ \ \vec{y}\in\Omega$, has a singularity just at the point $\vec{x}\in\overline{\Omega'}$, we obtain

$$||(1-\rho_\sigma^\Gamma-\rho_\sigma)\,\nabla_{\vec{x}} K_1(\vec{x},\cdot)||_{H_0^1(\Omega)} \le C\,|\,ln(\sigma)| \tag{5.16}$$

and by (5.11)

$$\begin{aligned}||\,\omega(\cdot,t_2)-\omega(\cdot,t_1)||_{H^{-1}(\Omega)} &\le \left\| \int_{t_1}^{t_2} \partial_t\omega(\cdot,t)\ dt \right\|_{H^{-1}(\Omega)}\\ &\le |t_2-t_1|\ \max_{t\in[0,T]} ||\,\partial_t\omega\,||_{L_\infty(0,T;\ H^{-1}(\Omega))} \le C|t_2-t_1|.\end{aligned} \tag{5.17}$$

Therefore choosing $\sigma := |t_2-t_1|$, from (5.13)–(5.17), we derive

$$|\nabla h_1(\vec{x},t_2) - \nabla h_1(\vec{x},t_1)| \le C|t_2-t_1|\,|\,ln|t_2-t_1|\,| \tag{5.18}$$

for all $t_1,t_2\in[0,T]$. Using this approach we can deduce the same estimate for h_1.

In view of (2.9) of Lemma 1, we have

$$||h_2(\cdot,t_2)-h_2(\cdot,t_1)||_{C^1(\overline{\Omega'})} \le C||a(\cdot,t_2)-a(\cdot,t_1)||_{L_1(\Gamma)} \le C|t_2-t_1|^\theta \tag{5.19}$$

for any $t_1,t_2\in[0,T]$. Therefore, accounting for (5.18) and (5.19), we derive (5.12). This completes the proof of the lemma. ∎

6. Appendix. Let us suppose that initially the boundary and initial data $b,\ \omega_0$ of the problem (1.1)–(1.6) are known. In this section we give sufficient conditions on these data under which there exists at least one extension $\breve{\omega}$ on the domain $\overline{\Omega}_T$, satisfying the conditions (1.10)–(1.12).

Let us assume that the data $b,\ \omega_0$ admit the following regularity:

$$\omega_0 \in W_p^{(2-\frac{1}{p})}(\Omega) \cap L_\infty(\Omega), \qquad b \in W_p^{(2-\frac{1}{p})}(\Gamma_T^-) \cap L_\infty(\Gamma_T^-) \tag{6.1}$$

with some $p > n$, such that

$$\omega_0(\vec{x}) \geq 0 \text{ a.e. in } \Omega, \quad b(\vec{x}, t) \geq 0 \text{ a.e. on } \Gamma_T^-. \tag{6.2}$$

Since $\Gamma \in C^{2+\gamma}$ for $\gamma > 0$, there exists a function $\breve{b} = \breve{b}(\vec{x}, t)$, defined on the boundary Γ_T, such that

$$\breve{b} \ \in W_p^{(2-\frac{1}{p})}(\Gamma_T) \cap L_\infty(\Gamma_T) \quad \text{and} \quad \breve{b}\big|_{\Gamma_T^-} = b, \quad \breve{b} \geq 0 \ \text{ on } \ \Gamma_T \tag{6.3}$$

and

$$\begin{cases} ||\breve{b}||_{L_\infty(\Gamma_T)} & \leq \quad ||b||_{L_\infty(\Gamma_T^-)}, \\ ||\breve{b}||_{W_p^{(2-\frac{1}{p})}(\Gamma_T)} & \leq \quad ||b||_{W_p^{(2-\frac{1}{p})}(\Gamma_T^-)}. \end{cases} \tag{6.4}$$

Let $\breve{\omega}$ be the solution of the system

$$\begin{cases} \breve{\omega}_t - \Delta\breve{\omega} = 0, \quad (\vec{x}, t) \in \Omega_T, \\ \breve{\omega}\big|_{\Gamma_T} = \breve{b}; \qquad \breve{\omega}\big|_{t=0} = \omega_0(\vec{x}), \quad \vec{x} \in \Omega. \end{cases}$$

According to (6.1)–(6.4), by [14, Theorem 9.1, p. 341 and Theorem 7.1, p. 181], there exists a unique solution $\breve{\omega} \in W_p^{2,1}(\Omega_T) \cap L_\infty(\Omega_T)$, such that

$$\begin{aligned} \breve{\omega} \geq 0 \quad \text{a.e. on } \ \Omega_T, \\ ||\breve{\omega}||_{L_\infty(\Omega_T)} &\leq C(||\breve{b}||_{L_\infty(\Gamma_T)} + ||\omega_0||_{L_\infty(\Omega)}) \\ &\leq C(||b||_{L_\infty(\Gamma_T^-)} + ||\omega_0||_{L_\infty(\Omega)}) \end{aligned} \tag{6.5}$$

and

$$\begin{aligned} ||\breve{\omega}||_{W_p^{2,1}(\Omega_T)} &\leq C\left(||\breve{b}||_{W_p^{\left(2-\frac{1}{p}\right)}(\Gamma_T)} + ||\omega_0||_{W_p^{\left(2-\frac{2}{p}\right)}(\Omega)}\right) \\ &\leq C\left(||b||_{W_p^{\left(2-\frac{1}{p}\right)}(\Gamma_T^-)} + ||\omega_0||_{W_p^{\left(2-\frac{2}{p}\right)}(\Omega)}\right), \end{aligned} \tag{6.6}$$

where the constants C depend only on p and Ω. The previous inequality implies that

$$\int_0^T || \bigtriangledown \breve{\omega}(\cdot, t)||^p_{L_\infty(\Omega)} \, dt \leq C||\breve{\omega}||^p_{W_p^{2,1}(\Omega_T)}. \tag{6.7}$$

Combining the estimates (6.5)–(6.7), we see that the constructed function $\breve{\omega}$ is an extension of $b,\ \omega_0$ on the domain $\overline{\Omega}_T$, which satisfies the conditions (1.10)–(1.12).

REFERENCES

[1] J. P. Aubin, *Un théorème de compacité*, C. R. Acad. Sci. Paris, 56 (1963), pp. 5042–5044.

[2] A. J. Briggs, J. R. Claisse, C. M. Elliott, and V. M. Styles, *Computation of vorticity evolution for a cylindrical type-*II *superconductor subject to parallel and transverse applied magnetic fields*, Ser. Adv. Math. Appl. Sci. 59, World Sci. Publ., River Edge, NJ, 2001, pp. 21–45.

[3] S. J. Chapman, *A hierarchy of models for type-*II *superconductors*, SIAM Rev., 42 (2000), pp. 555–598.

[4] S. J. Chapman, S. D. Howison, and J. R. Ockendon, *Macroscopic models of superconductivity*, SIAM Rev., 34 (1992), pp. 529–560.

[5] S. J. Chapman, *A mean-field model of superconducting vortices in three dimensions*, SIAM J. Appl. Math., 55 (1995), pp. 1259–1274.

[6] S. J. Chapman, J. Rubinstein, and M. Schatzman, *A mean-field model of superconducting vortices*, European J. Appl. Math., 7 (1996), pp. 97–111.

[7] X. Chen, C. M. Elliott, and Q. Tang, *Shooting method for vortex solutions of a complex-valued Ginsburg–Landau equation*, Proc. Roy. Soc. Edinburgh, 124 (1994), pp. 1075–1088.

[8] E. DiBenedetto, *Partial Differential Equations*, Birkhäuser Boston, Boston, 1995.

[9] C. M. Elliott, R. Schätzle, and B. Stoth, *Viscosity solutions of a degenerate parabolic elliptic system arising in the mean-field theory of superconductivity*, Arch. Ration. Mech. Anal., 145 (1998), pp. 99–127.

[10] C. M. Elliott and V. Styles, *Flux pinning and boundary nucleation of vorticity in a mean field model of superconducting vortices*, Interfaces Free Bound., 2 (2000), pp. 143–180.

[11] C. M. Elliott and V. M. Styles, *Numerical analysis of a mean field model of superconducting vortices*, IMA J. Numer. Anal., 21 (2001), pp. 1–51.

[12] C. Huang and T. Svobodny, *Evolution of mixed-state regions in type-*II *superconductors*, SIAM J. Math. Anal., 29 (1998), pp. 1002–1021.

[13] T. Kato, *On classical solutions of the two-dimensional nonstationary Euler equation*, Arch. Rational Mech. Anal., 25 (1967), pp. 188–200.

[14] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasilinear Equations of Parabolic type*, AMS, Providence, RI, 1968.

[15] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York-London, 1968.

[16] T.-C. Lin, *Smoothness results of single and double layer solutions of the Helmholtz equations*, J. Integral Equations Appl., 1 (1988), pp. 83–121.

[17] R. Schätzle and V. Styles, *Analysis of a mean field model of superconducting vortices*, European J. Appl. Math., 10 (1999), pp. 319–352.

[18] J. Simon, *Compact sets in the space* $L^p(0,T;B)$, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.

[19] V. M. Styles, *Error estimates for a finite-difference approximation of a mean field model of superconducting vortices in one-dimension*, IMA J. Numer Anal., 21 (2001), pp. 667–701.

[20] V. S. Vladimirov, *Equations of Mathematical Physics* (Uravneniya matematicheskoj fiziki), Moskua: Fiziko-Mathematicheskaya Literatura, Moscow, 2000, (in Russian).

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 281–297

CONVERGENCE OF SCHRÖDINGER OPERATORS*

JOHANNES F. BRASCHE† AND KATEŘINA OŽANOVÁ‡

Abstract. We prove two limit relations between Schrödinger operators perturbed by measures. First, weak convergence of finite real-valued Radon measures $\mu_n \longrightarrow m$ implies that the operators $-\Delta + \varepsilon^2\Delta^2 + \mu_n$ in $L^2(\mathbb{R}^d, dx)$ converge to $-\Delta + \varepsilon^2\Delta^2 + m$ in the norm-resolvent sense, provided $d \leq 3$. Second, for a large family, including the Kato class, of real-valued Radon measures m, the operators $-\Delta + \varepsilon^2\Delta^2 + m$ tend to the operator $-\Delta + m$ in the norm-resolvent sense as ε tends to zero. Explicit upper bounds for the rates of convergences are derived. Since one can choose point measures μ_n with mass at only finitely many points, a combination of both convergence results leads to an efficient method for the numerical computation of the eigenvalues in the discrete spectrum and corresponding eigenfunctions of Schrödinger operators.

Key words. point interaction, eigenvalue, eigenfunction, singular Schrödinger operator

AMS subject classifications. 81-08, 35P15, 47B25

DOI. 10.1137/060651598

1. Introduction. In this paper we are going to analyze convergence of Schrödinger operators perturbed by measures. It is known that weak convergence of potentials implies norm-resolvent convergence of the corresponding one-dimensional Schrödinger operators. This result from [6] may be interesting for several reasons. For instance, every finite real-valued Radon measure on $\mathbb{R}$ is the weak limit of a sequence of point measures with mass at only finitely many points. There exist efficient numerical methods for the computation of the eigenvalues and corresponding eigenfunctions of one-dimensional Schrödinger operators with a potential supported by a finite set; actually the effort for the computation grows at most linearly with the number of points of the support [9]. Since the norm-resolvent convergence implies convergence of the eigenvalues in the discrete spectra and corresponding eigenspaces, we get an efficient method for the numerical calculation of the discrete spectra of one-dimensional Schrödinger operators. Norm-resolvent convergence also has other important consequences: locally uniform convergence of the associated unitary groups and semigroups, convergence of the spectral projectors (which implies the mentioned results on the discrete spectra), etc.

Let us also mention a completely different motivation for studying the convergence of operators with point potentials. In quantum mechanics neutron scattering is often described via so-called zero-range Hamiltonians (the monograph [1] is an excellent standard reference to this research area). In a wide variety of models the positions of the neutrons are described via a family $(X_j)_{j=1}^n$ of independent random variables with joint distribution μ. Usually the number n of neutrons is large, and one is interested in the limit when n tends to infinity and the strengths of the single size potentials tend to zero. In the one-dimensional case this motivates one to investigate the limits

*Received by the editors February 6, 2006; accepted for publication (in revised form) January 4, 2007; published electronically May 22, 2007. This work is partially supported by the Marie Curie fellowship MEIF-CT-2004-009256.
http://www.siam.org/journals/sima/39-1/65159.html

†Institute of Mathematics, TU Clausthal, D-38678 Clausthal-Zellerfeld, Germany (johannes.brasche@tu-clausthal.de).
‡Department of Mathematics, CTH & GU, 41296 Göteborg, Sweden (nemco@math.chalmers.se).

of operators of the form

$$-\frac{d^2}{dx^2}+\frac{a}{n}\sum_{j=1}^{n}\delta_{X_j(\omega)}, \qquad \omega\in\Omega,$$

$a\neq 0$ being a real constant and $(\Omega,\mathcal{F},\mathbb{P})$ a probability space. By the theorem of Glivenko–Cantelli, for $\mathbb{P}$-almost all $\omega\in\Omega$ the sequence $(\frac{a}{n}\sum_{j=1}^{n}\delta_{X_j(\omega)})_{n\in\mathbb{N}}$ converges to the measure $a\mu$ weakly. By the mentioned result from [6], this implies that

$$-\frac{d^2}{dx^2}+a\mu=\lim_{n\longrightarrow\infty}\left(-\frac{d^2}{dx^2}+\frac{a}{n}\sum_{j=1}^{n}\delta_{X_j(\omega)}\right)$$

in the norm-resolvent sense $\mathbb{P}$ a.s.

It is the purpose of the present note to derive analogous results in the two- and three-dimensional cases. It was shown in [6] and [8] that one can approximate Schrödinger operators perturbed by suitable measures by a point potential Hamiltonian. However, the convergence there was in the strong resolvent sense, which is of course a weaker result than the norm-resolvent convergence.

If the dimension is higher than one, then it seems to be impossible to work directly with operators of the form $-\Delta+\mu$, μ being a point measure. In fact, while the operators $-\frac{d^2}{dx^2}+\sum_{j=1}^{n}a_j\delta_{x_j}$ can be defined in dimension one via Kato's quadratic form method as the unique lower semibounded self-adjoint operator associated to the energy form

$$\begin{aligned} D(\mathcal{E}) &:= H^1(\mathbb{R}),\\ \mathcal{E}(f,f) &:= \int |f'(x)|^2dx+\sum_{j=1}^{n}a_j|\tilde{f}(x_j)|^2, \quad f\in D(\mathcal{E}), \end{aligned}$$

$\tilde{f}$ being the unique continuous representative of $f\in H^1(\mathbb{R})$, in higher dimension $d>1$, the quadratic form

$$\begin{aligned} D(\mathcal{E}) &:= \{f\in H^1(\mathbb{R}^d) : f \text{ has a continuous representative } \tilde{f}\},\\ \mathcal{E}(f,f) &:= \int |\nabla f(x)|^2dx+\sum_{j=1}^{n}a_j|\tilde{f}(x_j)|^2, \quad f\in D(\mathcal{E}), \end{aligned}$$

is not lower semibounded and closable if at least one coefficient a_j is different from zero.

The strategy to overcome the mentioned problem in higher dimensions is based on two simple observations:

1. The lower semibounded self-adjoint operator $\Delta^2+\mu$ can be defined via Kato's quadratic form method for every real-valued finite Radon measure μ on $\mathbb{R}^d$ (if $d\in\{1,2,3\}$), including point measures.
2. $-\Delta+\varepsilon^2\Delta^2\longrightarrow-\Delta$ in the norm-resolvent sense, as $\varepsilon>0$ tends to zero.

We show the convergence claim in two steps. In section 2 we shall prove that the sequence $(-\Delta+\varepsilon^2\Delta^2+\mu_n)_{n\in\mathbb{N}}$ converges to $-\Delta+\varepsilon^2\Delta^2+m$ in the norm-resolvent sense provided $d\leq 3$, $\varepsilon>0$, and the finite real-valued Radon measures μ_n on $\mathbb{R}^d$ converge to the finite real-valued Radon measure m weakly. Then for a large class of measures m we shall prove that

$$-\Delta+\varepsilon^2\Delta^2+m\longrightarrow-\Delta+m$$

in the norm-resolvent sense as ε tends to zero; cf. section 3. Actually, we will not only prove convergence but also give explicit error estimates.

As approximating measures μ_n we can, in particular, choose point measures with mass at only finitely many points. In section 4 we will present formulas which make it possible to calculate the eigenvalues and corresponding eigenspaces of operators perturbed by a finite number point measures. Then similarly to [1, Chapter II.2], the spectral problem means to solve an implicit equation and the effort for these computations grows at most as $\mathcal{O}(n^3)$.

Putting both convergence results from sections 2 and 3 and formulas from section 4 together, we get an efficient method to calculate the eigenvalues in the discrete spectrum and corresponding eigenspaces of Schrödinger operators $-\Delta + m$ numerically.

Our method covers not only the case when m is absolutely continuous w.r.t. the $(d-1)$-dimensional volume measure of a manifold with codimension one but also a fairly large class of measures m containing the set of all finite real-valued measures belonging to the Kato class. In particular, the absolutely continuous case $dm = V dx$, where $-\Delta + m = -\Delta + V$ is a regular Schrödinger operator, is contained in our approach. We refer to [10] for related convergence results in the regular case.

Notation and auxiliary results. Let μ be a real-valued Radon measure on $\mathbb{R}^d$. By the Hahn–Jordan theorem, there exist unique positive Radon measures $\mu^{\pm}$ on $\mathbb{R}^d$ such that

$$\mu = \mu^+ - \mu^- \text{ and } \mu^+(\mathbb{R}^d \setminus B) = 0 = \mu^-(B)$$

for some suitably chosen Borel set B. We put

$$\| \mu \| := \mu^+(\mathbb{R}^d) + \mu^-(\mathbb{R}^d) \text{ and } |\mu| := \mu^+ + \mu^-.$$

If μ is finite, then we define its Fourier transform $\hat{\mu}$ as

$$\hat{\mu}(p) := (2\pi)^{-d/2} \int e^{ipx} \mu(dx), \qquad p \in \mathbb{R}^d.$$

Similarly, $\hat{f}$ also denotes the Fourier transform of $f \in L^2(dx) := L^2(\mathbb{R}^d, dx)$, dx being the Lebesgue measure.

For $s > 0$ we denote the Sobolev space of order s by $H^s(\mathbb{R}^d)$; i.e.,

$$H^s(\mathbb{R}^d) := \left\{ f \in L^2(dx) : \int (1+p^2)^s |\hat{f}(p)|^2 dp < \infty \right\},$$
$$\| f \|_{H^s} := \left(\int (1+p^2)^s |\hat{f}(p)|^2 dp \right)^{1/2}, \qquad f \in H^s(\mathbb{R}^d).$$

We shall use occasionally the abbreviations $L^2(\mu) := L^2(\mathbb{R}^d, \mu)$ and $H^s := H^s(\mathbb{R}^d)$.

$\| T \|_{\mathcal{H}_1, \mathcal{H}_2}$ denotes the operator norm of T as an operator from $\mathcal{H}_1$ to $\mathcal{H}_2$, and $\| T \|_{\mathcal{H}} := \| T \|_{\mathcal{H},\mathcal{H}}$. $\| f \|_{\mathcal{H}}$ and $(f,h)_{\mathcal{H}}$ represent the norm and the scalar product in the Hilbert $\mathcal{H}$, respectively. If the reference to a measure is missing, then we tacitly refer to the Lebesgue measure dx. For instance, "integrable" means "integrable w.r.t. dx" if not stated otherwise; $\| T \|$, (f,h), and $\| f \|$ denote the operator norm of T, the scalar product, and the norm in the Hilbert space $L^2(dx)$, respectively. We denote by $C_0^\infty(\mathbb{R}^d)$ the space of smooth functions with compact support.

For arbitrary $\varepsilon \geq 0$ ($\varepsilon = 0$ will be admitted only in section 3) let $\mathcal{E}_\varepsilon$ be the nonnegative closed quadratic form in the Hilbert space $L^2(dx)$ associated to the nonnegative self-adjoint operator $-\Delta + \varepsilon^2\Delta^2$ in $L^2(dx)$. Obviously we have

$$D(\mathcal{E}_\varepsilon) = H^2(\mathbb{R}^d),$$
$$\mathcal{E}_\varepsilon(f,f) = \varepsilon^2\,(\Delta f, \Delta f) + (\nabla f, \nabla f) \geq \varepsilon^2\,(\Delta f, \Delta f), \qquad f \in D(\mathcal{E}_\varepsilon),$$

for every $\varepsilon > 0$. Note that for $\varepsilon = 0$ the form domain is $H^1(\mathbb{R}^d)$, and $\mathcal{E}_0$ is the classical Dirichlet form. For any $\alpha > 0$ we put

$$\mathcal{E}_{\varepsilon,\alpha}(f,h) := \mathcal{E}_\varepsilon(f,h) + \alpha(f,h), \qquad f,h \in D(\mathcal{E}_\varepsilon).$$

2. Operator norm convergence. Throughout this section let $d \leq 3$, and let μ be a finite real-valued Radon measure on $\mathbb{R}^d$. Then, by Sobolev's embedding theorem, for every $s > 3/2$, and, in particular, for $s = 2$, every $f \in H^s(\mathbb{R}^d)$ has a unique continuous representative $\tilde{f}$ and

$$\| \tilde{f} \|_\infty := \sup\{|\tilde{f}(x)| : x \in \mathbb{R}^d\} \leq c_s \| f \|_{H^s}, \qquad f \in H^s(\mathbb{R}^d), \tag{1}$$

for some finite constant c_s. Note that $c_s \leq 1$ if $s = 2$. It follows that for every $\varepsilon > 0$ and every $\eta > 0$ there exists an $\alpha = \alpha(\varepsilon, \eta) < \infty$ such that

$$\| \tilde{f} \|_\infty^2 \leq \eta\, \mathcal{E}_\varepsilon(f,f) + \alpha(f,f), \qquad f \in H^2(\mathbb{R}^d). \tag{2}$$

Since μ is finite, for arbitrary $\varepsilon, \eta > 0$ and some finite α we get

$$\left| \int |\tilde{f}|^2 d\mu \right| \leq \eta \| \mu \| \mathcal{E}_\varepsilon(f,f) + \alpha \| \mu \| (f,f), \qquad f \in H^2(\mathbb{R}^d). \tag{3}$$

We put

$$D(\mathcal{E}_\varepsilon^\mu) := H^2(\mathbb{R}^d),$$
$$\mathcal{E}_\varepsilon^\mu(f,f) := \mathcal{E}_\varepsilon(f,f) + \int |\tilde{f}|^2 d\mu, \qquad f \in D(\mathcal{E}_\varepsilon^\mu).$$

By (3) and the Kato-Lax-Milgram-Nelson (KLMN) theorem [7, Theorem 1.5], $\mathcal{E}_\varepsilon^\mu$ is a lower semibounded closed quadratic form in $L^2(dx)$. We denote the lower semibounded self-adjoint operator in $L^2(dx)$ associated to $\mathcal{E}_\varepsilon^\mu$ by $-\Delta + \varepsilon^2\Delta^2 + \mu$.

Our main tool to prove convergence results will be a Krein-like formula which expresses the resolvent $(-\Delta + \varepsilon^2\Delta^2 + \mu + \alpha)^{-1}$ by means of the resolvent

$$G_{\varepsilon,\alpha} := (-\Delta + \varepsilon^2\Delta^2 + \alpha)^{-1}.$$

The operator $G_{\varepsilon,\alpha}$ has the integral kernel $g_{\varepsilon,\alpha}(x-y)$ with the Fourier transform

$$\hat{g}_{\varepsilon,\alpha}(p) := \frac{1}{\varepsilon^2 p^4 + p^2 + \alpha}, \qquad p \in \mathbb{R}^d.$$

For every $\varepsilon \geq 0$ and $\alpha > 0$, the function $g_{\varepsilon,\alpha}(x)$ is continuous on $\mathbb{R}^d \setminus \{0\}$, and, if $d = 1$ or if $d \leq 3$ and $\varepsilon > 0$, it is continuous on whole $\mathbb{R}^d$. Moreover, it is radially symmetric. Finally, $g_{0,\alpha}$ is the Green function of the free Laplacian in $\mathbb{R}^d$, and it is nonnegative. By the dominated convergence theorem,

$$\| g_{\varepsilon,\alpha} \|_{H^2}^2 = \int \frac{(1+p^2)^2}{|\varepsilon^2 p^4 + p^2 + \alpha|^2}\, dp \longrightarrow 0 \qquad \text{as } |\alpha| \longrightarrow \infty, \tag{4}$$

which, by Sobolev's inequality, implies that

$$\| g_{\varepsilon,\alpha} \|_\infty \longrightarrow 0 \qquad \text{as } |\alpha| \longrightarrow \infty. \tag{5}$$

The fact that $g_{\varepsilon,\alpha}$ is the Green function of $-\Delta + \varepsilon^2\Delta^2$ means that

$$\int g_{\varepsilon,\alpha}(x-y)(-\Delta + \varepsilon^2\Delta^2 + \alpha)h(y)dy = h(x) \qquad dx \text{ a.e.}$$

for all $h \in D(-\Delta + \varepsilon^2\Delta^2) = H^4(\mathbb{R}^d)$. The equation above not only holds almost everywhere w.r.t. the Lebesgue measure dx but even pointwise everywhere, as the following lemma states.

LEMMA 1. *Let the Green function $g_{\varepsilon,\alpha}$ and the operator $-\Delta+\varepsilon^2\Delta^2+\alpha$ be defined as above. Then one has*

$$\int g_{\varepsilon,\alpha}(x-y)(-\Delta + \varepsilon^2\Delta^2 + \alpha)h(y)dy = \tilde{h}(x), \qquad x \in \mathbb{R}^d, \tag{6}$$

for all $h \in H^4(\mathbb{R}^d)$.

Proof. In fact, we have only to show that the integral on the left-hand side is a continuous function of $x \in \mathbb{R}^d$. We choose any sequence $(f_n)_{n\in\mathbb{N}}$ of continuous functions with compact support converging to $(-\Delta + \varepsilon^2\Delta^2 + \alpha)h$ in $L^2(dx)$. By (4), $g_{\varepsilon,\alpha} \in H^2(\mathbb{R}^d) \subset L^2(dx)$; therefore, we can write

$$\int g_{\varepsilon,\alpha}(x-y)(-\Delta + \varepsilon^2\Delta^2 + \alpha)h(y)dy = \lim_{n\longrightarrow\infty} \int g_{\varepsilon,\alpha}(x-y)f_n(y)dy, \qquad x \in \mathbb{R}^d.$$

Obviously the mapping $x \mapsto \int g_{\varepsilon,\alpha}(x-y)f_n(y)dy$, $\mathbb{R}^d \longrightarrow \mathbb{C}$, is the unique continuous representative $\widetilde{G_{\varepsilon,\alpha}f_n}$ of $G_{\varepsilon,\alpha}f_n$ for every $n \in \mathbb{N}$. Since $G_{\varepsilon,\alpha}$ is a bounded operator from $L^2(dx)$ to $H^2(\mathbb{R}^d)$ (even to $H^4(\mathbb{R}^d)$), the sequence $(G_{\varepsilon,\alpha}f_n)_{n\in\mathbb{N}}$ converges in $H^2(\mathbb{R}^d)$ to $G_{\varepsilon,\alpha}(-\Delta + \varepsilon^2\Delta^2 + \alpha)h = h$. By Sobolev's inequality (1), this implies that the sequence $(\widetilde{G_{\varepsilon,\alpha}f_n})_{n\in\mathbb{N}}$ of the unique continuous representatives converges to a continuous function uniformly. By the last equality, $x \mapsto \int g_{\varepsilon,\alpha}(x-y)(-\Delta+\varepsilon^2\Delta^2+\alpha)h(y)dy$, $\mathbb{R}^d \longrightarrow \mathbb{C}$, is this continuous uniform limit, and we have proved (6). □

We introduce the following integral operator:

$$G^\mu_{\varepsilon,\alpha} f(x) := \int g_{\varepsilon,\alpha}(x-y)\tilde{f}(y)\mu(dy) \qquad dx \text{ a.e., } f \in H^2(\mathbb{R}^d).$$

We can prove several estimates of its operator norm.

LEMMA 2. *The operator $G^\mu_{\varepsilon,\alpha}$ is bounded on $H^2(\mathbb{R}^d)$, and its operator norm $\| G^\mu_{\varepsilon,\alpha} \|_{H^2}$ decays with $\alpha \longrightarrow \infty$. The operator is bounded also w.r.t. other operator norms; in particular, there are finite real numbers c_i, $i = 1, 2, 3$, such that*

$$\begin{aligned} \| G^\mu_{\varepsilon,\alpha} f \|_{H^2} &\le c_1(\alpha) \| \tilde{f} \|_\infty, \\ \| G^\mu_{\varepsilon,\alpha} f \|_{L^2} &\le c_2(\alpha) \| \tilde{f} \|_{L^2(|\mu|)}, \qquad f \in H^2(\mathbb{R}^d), \\ \| \widetilde{G^\mu_{\varepsilon,\alpha} f} \|_{L^2(|\mu|)} &\le c_3(\alpha) \| \tilde{f} \|_{L^2(|\mu|)}, \end{aligned}$$

and all three numbers c_i vanish in the limit $\alpha \longrightarrow \infty$.

Proof. Using Sobolev's inequality we have for arbitrary $f \in H^2(\mathbb{R}^d)$

$$|\widehat{\tilde{f}\mu}(p)|^2 \le (2\pi)^{-d} \| \tilde{f} \|^2_\infty \| \mu \|^2 \le (2\pi)^{-d} \| \tilde{f} \|^2_{H^2} \| \mu \|^2, \qquad p \in \mathbb{R}^d.$$

Then the convolution theorem yields

$$
\begin{aligned}
\| G^{\mu}_{\varepsilon,\alpha} f \|^2_{H^2} &= \int |(1+p^2)^2| \, |(g_{\varepsilon,\alpha} * \tilde{f}\mu)\hat{}(p)|^2 dp \\
&= (2\pi)^d \int \frac{(1+p^2)^2}{|\varepsilon^2 p^4 + p^2 + \alpha|^2} |\widehat{\tilde{f}\mu}(p)|^2 dp \\
&\leq \int \frac{(1+p^2)^2}{|\varepsilon^2 p^4 + p^2 + \alpha|^2} \ \| \tilde{f} \|^2_{\infty} \| \mu \|^2 \, dp \\
&\leq \int \frac{(1+p^2)^2}{|\varepsilon^2 p^4 + p^2 + \alpha|^2} dp \ \| \tilde{f} \|^2_{H^2} \| \mu \|^2 < \infty, \qquad f \in H^2(\mathbb{R}^d).
\end{aligned}
$$

Therefore $G^{\mu}_{\varepsilon,\alpha}$ is an everywhere defined bounded operator on $H^2(\mathbb{R}^d)$, and we get an upper bound for the norm

$$
\| G^{\mu}_{\varepsilon,\alpha} \|_{H^2,H^2} \leq \| \mu \| \left(\int \frac{(1+p^2)^2}{|\varepsilon^2 p^4 + p^2 + \alpha|^2} dp \right)^{1/2}, \tag{7}
$$

and the expression on the right-hand side (R.H.S.) is also the uniform upper bound c_1.

To determine the remaining upper bounds c_2 and c_3, we can write

$$
\begin{aligned}
& \int |G^{\mu}_{\varepsilon,\alpha} f(x)|^2 dx \\
= \ & \int \left| \int g_{\varepsilon,\alpha}(x-y)\tilde{f}(y)\mu^+(dy) - \int g_{\varepsilon,\alpha}(x-y)\tilde{f}(y)\mu^-(dy) \right|^2 dx \\
\leq \ & 2 \int \left| \int g_{\varepsilon,\alpha}(x-y)\tilde{f}(y)\mu^+(dy) \right|^2 dx + 2 \int \left| \int g_{\varepsilon,\alpha}(x-y)\tilde{f}(y)\mu^-(dy) \right|^2 dx \\
\leq \ & 2 \int \int |g_{\varepsilon,\alpha}(x-y)|^2 \mu^+(dy) \int |\tilde{f}(y)|^2 \mu^+(dy) \, dx \\
& + 2 \int \int |g_{\varepsilon,\alpha}(x-y)|^2 \mu^-(dy) \int |\tilde{f}(y)|^2 \mu^-(dy) \, dx \\
\leq \ & 2 \int |g_{\varepsilon,\alpha}(x)|^2 dx \ \| \mu \| \int |\tilde{f}(y)|^2 |\mu|(dy), \qquad f \in H^2(\mathbb{R}^d). \quad (8)
\end{aligned}
$$

In a similar way we arrive at

$$
\int |\widetilde{G^{\mu}_{\varepsilon,\alpha} f}(x)|^2 |\mu|(dx) \leq 2 \| g_{\varepsilon,\alpha} \|^2_{\infty} \| \mu \|^2 \int |\tilde{f}(y)|^2 |\mu|(dy).
$$

Finally, from (4) and (5) one concludes that all of the upper bounds of the operator norms tend to zero in the limit $\alpha \longrightarrow \infty$. □

General results of [3] (cf. also section 3 below) provide, in particular, an explicit formula for the resolvent of the operator $-\Delta + \varepsilon^2\Delta^2 + \mu$. In this resolvent formula there occur operators acting in different Hilbert spaces. This is inconvenient when we investigate the convergence of sequences of such operators, and we shall use a slightly different resolvent formula:

$$
(-\Delta + \varepsilon^2\Delta^2 + \mu + \alpha)^{-1} = G_{\varepsilon,\alpha} - G^{\mu}_{\varepsilon,\alpha}(I + G^{\mu}_{\varepsilon,\alpha})^{-1} G_{\varepsilon,\alpha}. \tag{9}
$$

For the sake of completeness we present the proof of the above Krein's formula in the appendix. According to Lemma 2, we can choose $\alpha > 0$ such that $\| G^{\mu}_{\varepsilon,\alpha} \|_{H^2,H^2} < 1$.

Then the operator $I + G^{\mu}_{\varepsilon,\alpha}$ is invertible, and its inverse is everywhere defined on $H^2(\mathbb{R}^d)$ and bounded; here I denotes the identity on $H^2(\mathbb{R}^d)$. By (3), we can choose $\alpha > 0$ such that, in addition,

$$\mathcal{E}^{\mu}_{\varepsilon,\alpha}(f,f) := \mathcal{E}^{\mu}_{\varepsilon}(f,f) + \alpha(f,f) \geq (f,f), \quad f \in D(\mathcal{E}^{\mu}_{\varepsilon}). \tag{10}$$

We are now prepared for the proof of the main theorem of this section.

THEOREM 3. *Let m and μ_n, $n \in \mathbb{N}$, be finite real-valued Radon measures on $\mathbb{R}^d$. Suppose that the sequence $(\mu_n)_{n\in\mathbb{N}}$ converges to m weakly and* $\sup_{n\in\mathbb{N}} \| \mu_n \| < \infty$. *Let $\varepsilon, \alpha > 0$ and $d \in \{1,2,3\}$. Then the operators $-\Delta + \varepsilon^2\Delta^2 + \mu_n$ converge to $-\Delta + \varepsilon^2\Delta^2 + m$ in the norm-resolvent sense.*

Proof. Let $\varepsilon > 0$ be arbitrary. We choose $0 < c < 1$ and $\alpha > 0$ such that

$$\| \mu_n \|^2 \int \frac{(1+p^2)^2}{|\varepsilon^2 p^4 + p^2 + \alpha|^2}\, dp \leq c^2, \quad n \in \mathbb{N}, \tag{11}$$

and

$$\| m \|^2 \int \frac{(1+p^2)^2}{|\varepsilon^2 p^4 + p^2 + \alpha|^2}\, dp \leq c^2. \tag{12}$$

According to (3), we can choose $\alpha > 0$ such that, in addition,

$$\mathcal{E}^{\mu_n}_{\varepsilon,\alpha}(f,f) \geq (f,f), \qquad f \in H^2(\mathbb{R}^d), \quad n \in \mathbb{N}. \tag{13}$$

Since $(\mu_n)_{n\in\mathbb{N}}$ converges to m weakly, (13) also holds when we replace μ_n by m. By Lemma 2, in particular estimate (7), inequalities (11) and (12) yield

$$\begin{aligned}
&\| G^{\mu_n}_{\varepsilon,\alpha} \|_{H^2,H^2} \leq c, \qquad n \in \mathbb{N},\\
&\| G^{m}_{\varepsilon,\alpha} \|_{H^2,H^2} \leq c,\\
&\| G^{m}_{\varepsilon,\alpha} f \|_{H^2} \leq c \, \| \tilde{f} \|_{\infty}, \qquad f \in H^2(\mathbb{R}^d).
\end{aligned} \tag{14}$$

Hence the resolvent formula (9) is valid both for $\mu = m$ and for $\mu = \mu_n$, $n \in \mathbb{N}$. By Lemma 2, we can choose α sufficiently large so that also

$$\int |G^m_{\varepsilon,\alpha} h(x)|^2 dx \leq c^2 \int |\tilde{h}|^2 d|m| \text{ and } \int |\widetilde{G^m_{\varepsilon,\alpha} h}(x)|^2 |m|(dx) \leq c^2 \int |\tilde{h}|^2 d|m| \tag{15}$$

for every $h \in H^2(\mathbb{R}^d)$.

For notational brevity we put

$$g_0 := g_{0,1}, \quad g := g_{\varepsilon,\alpha}, \quad G := G_{\varepsilon,\alpha}, \quad G^{\mu_n} := G^{\mu_n}_{\varepsilon,\alpha}, \quad \text{and} \quad G^m := G^m_{\varepsilon,\alpha}.$$

With this notation we have

$$\begin{aligned}
&(-\Delta + \varepsilon^2\Delta^2 + \mu_n + \alpha)^{-1} - (-\Delta + \varepsilon^2\Delta^2 + m + \alpha)^{-1}\\
&\quad = G^m[I + G^m]^{-1}G - G^{\mu_n}[I + G^{\mu_n}]^{-1}G\\
&\quad = (G^m - G^{\mu_n})[I + G^m]^{-1}G + (G^{\mu_n} - G^m)[I + G^m]^{-1}(G^{\mu_n} - G^m)[I + G^{\mu_n}]^{-1}G\\
&\qquad + G^m[I + G^m]^{-1}(G^{\mu_n} - G^m)[I + G^{\mu_n}]^{-1}G.
\end{aligned}$$

Since G is a bounded operator from $L^2(dx)$ to $H^2(\mathbb{R}^d)$, we have only to show that

$$\| G^m - G^{\mu_n} \|_{H^2,L^2(dx)} \longrightarrow 0 \qquad \text{as } n \longrightarrow \infty, \tag{16}$$

$$\| G^m[I + G^m]^{-1}(G^m - G^{\mu_n}) \|_{H^2,L^2(dx)} \longrightarrow 0 \qquad \text{as } n \longrightarrow \infty. \tag{17}$$

We introduce

$$\nu_n := m - \mu_n,$$
$$\nu_{nx}(dy) := g(x-y)\,\nu_n(dy), \qquad x \in \mathbb{R}^d, \quad n \in \mathbb{N}.$$

As $d \leq 3$, the function

$$y \mapsto \int g_0(y-a)\, f(a)\, da$$

is continuous and bounded for every $f \in L^2(dx)$; this well-known fact can be proved in the same way as (6). Since the function g is bounded and g_0 is nonnegative, it follows that

$$\left| \int |g(x-y)| \int |g_0(y-a)|\, |(-\Delta+1)h(a)|\, da\, \nu_n^{\pm}(dy) \right| < \infty$$

for all $x \in \mathbb{R}^d$ and $h \in H^2(\mathbb{R}^d)$. Hence by Fubini's theorem, the function $k_{\nu_{nx}} : \mathbb{R}^d \longrightarrow \mathbb{R}$, defined by

$$k_{\nu_{nx}}(a) := \begin{cases} \int g_0(y-a)\, g(x-y)\, \nu_n(dy) & \text{if defined,} \\ 0 & \text{otherwise,} \end{cases}$$

is Borel measurable, the integral on the R.H.S. is defined and finite for almost all $a \in \mathbb{R}^d$ (almost all w.r.t. the Lebesgue measure) and

$$\begin{aligned} |(G^{\nu_n} \tilde{h})(x)|^2 &= \left| \int g(x-y)\, h(y)\, \nu_n(dy) \right|^2 \\ &= \left| \int g(x-y) \int g_0(y-a)(-\Delta+1)h(a)\, da\, \nu_n(dy) \right|^2 \\ &\leq \int |k_{\nu_{nx}}(a)|^2 da \cdot \int |(-\Delta+1)h(a)|^2 da \\ &\leq 2 \parallel h \parallel_{H^2}^2 \int |k_{\nu_{nx}}(a)|^2 da, \qquad h \in H^2(\mathbb{R}^d),\ n \in \mathbb{N}. \end{aligned} \tag{18}$$

Thus in order to prove (16) we have only to show that

$$\int \int |k_{\nu_{nx}}(a)|^2 da\, dx \longrightarrow 0 \qquad \text{as } n \longrightarrow \infty. \tag{19}$$

We have

$$\begin{aligned} &\int \int |k_{\nu_{nx}}(a)|^2 da\, dx = (2\pi)^d \int \int |\widehat{g_0}(p)|^2 |\widehat{\nu_{nx}}(p)|^2 dp\, dx \\ &= \int \int \frac{1}{|1+p^2|^2} \int e^{ipy} g(x-y)\, \nu_n(dy) \int e^{-ipz} g(x-z) \nu_n(dz)\, dp\, dx. \end{aligned} \tag{20}$$

Since $|1+p^2|^{-2}$ and g are integrable w.r.t. the Lebesgue measure, g is bounded, and the Radon measures ν_n are finite, we can change the order of integration. Let us rewrite (20) as

$$\int f(y,z)\, h(y,z)\, \nu_n \otimes \nu_n(dy\, dz).$$

The function

$$f(y,z) := \int e^{ipy} e^{-ipz} \frac{1}{|1+p^2|^2}\, dp, \qquad y,z \in \mathbb{R}^d,$$

is bounded and continuous. It follows from the fact that it is (up to multiplication by $(2\pi)^{d/2}$) the inverse Fourier transform of the integrable function $|1+p^2|^{-2}$ at the point $z-y$.

Also the function

$$h(y,z) := \int g(x-y)\, g(x-z)\, dx$$

is bounded and continuous for $y,z \in \mathbb{R}^d$. This can be shown using the following observation. Let $y \in \mathbb{R}^d$ and K be any compact neighborhood of y. Since $|x|^j g_{\varepsilon,\alpha}(x) \longrightarrow 0$ for every $j \in N$ as $|x| \longrightarrow \infty$, there exists a constant $a < \infty$ such that

$$|g(x-y)\, g(x-z)| \le a \parallel g \parallel_\infty \mathrm{dist}(x,K)^{-4}, \qquad x \in \mathbb{R}^d \setminus K, \quad z \in \mathbb{R}^d, \quad y \in K.$$

By the Stone–Weierstrass theorem, the set of functions of the form $\sum_{j=1}^N f_j(x) g_j(y)$, $N \in \mathbb{N}$, where f_j, g_j are bounded and continuous, is dense in the space of bounded continuous functions w.r.t. the supremum norm. Since the measures ν_n tend to zero weakly and $\sup_{n\in\mathbb{N}} \parallel \nu_n \parallel < \infty$, this implies that the product measures $\nu_n \otimes \nu_n$ tend to zero weakly, too. Hence by (20), we have proved (19) and therefore also (16).

It only remains to prove (17). For this purpose we first note that

$$c_n := \int\int |k_{\nu_{nx}}(a)|^2\, da\, |m|(dx) \longrightarrow 0 \qquad \text{as } n \longrightarrow \infty.$$

This can be shown by mimicking the proof of (19). By (18), it follows that

$$\int |(G^{\nu_n} h\tilde{)}(x)|^2\, |m|(dx) \le 2c_n \parallel h \parallel_{H^2}^2, \qquad h \in H^2(\mathbb{R}^d).$$

Thus, in order to prove (17), we have only to show that there exists a finite constant C such that

$$\parallel G^m (I+G^m)^{-1} h \parallel_{L^2(dx)} \le C \left(\int |\tilde{h}|^2 d|m| \right)^{1/2}, \qquad h \in H^2(\mathbb{R}^d). \tag{21}$$

Using the estimates (14), we have

$$G^m (I+G^m)^{-1} = -\sum_{j=1}^{\infty} (-G^m)^j. \tag{22}$$

According to (15),

$$\parallel (G^m)^{j+1} h \parallel_{L^2(dx)} \le c \left(\int |\widetilde{(G^m)^j h}|^2 d|m| \right)^{1/2} \le c \cdot c^j \left(\int |\tilde{h}|^2 d|m| \right)^{1/2}$$

for every $j \in \mathbb{N}$, and hence

$$\left\| \sum_{j=1}^{\infty} (-G^m)^j h \right\|_{L^2(dx)} \le \sum_{j=1}^{\infty} c^j \left(\int |\tilde{h}|^2 d|m| \right)^{1/2} = \frac{c}{1-c} \left(\int |\tilde{h}|^2 d|m| \right)^{1/2}.$$

By (22), this implies (21) and the proof of the theorem is complete. □

Remark 4. We have shown that

$$
\begin{aligned}
&\| (-\Delta + \varepsilon^2\Delta^2 + \mu_n + \alpha)^{-1} - (-\Delta + \varepsilon^2\Delta^2 + m + \alpha)^{-1} \|^2 \\
\leq \quad & C_1 \int\int \left| \int g_{0,1}(y-a) g_{\varepsilon,\alpha}(x-y)(m-\mu_n)(dy) \right|^2 da\, dx \\
&+ C_2 \int\int \left| \int g_{0,1}(y-a) g_{\varepsilon,\alpha}(x-y)(m-\mu_n)(dy) \right|^2 da\, |m|(dx)
\end{aligned}
$$

for some finite constants $C_j = C_j(\varepsilon,\alpha)$, $j = 1, 2$, which can be computed with the aid of the proof of Theorem 3. Thus the proof provides explicit upper bounds for the error one makes when one replaces the operator $-\Delta+\varepsilon^2\Delta^2+m$ by $-\Delta+\varepsilon^2\Delta^2+\mu_n$.

Remark 5. The essential spectrum of $-\Delta + \varepsilon^2\Delta^2 + m$ remains the same for any finite real-valued Radon measure m on $\mathbb{R}^d$; i.e.,

$$
\sigma_{ess}(-\Delta + \varepsilon^2\Delta^2 + m) = \sigma_{ess}(-\Delta + \varepsilon^2\Delta^2) = [0,\infty). \tag{23}
$$

By Sobolev's inequality and [4, Lemma 19], the mapping $f \mapsto \tilde{f}$ from $H^2(\mathbb{R}^d)$ to $L^2(|m|)$ is compact. Therefore using estimate (8), one may conclude that $G^{\mu}_{\varepsilon,\alpha}$ is compact if regarded as an operator from $H^2(\mathbb{R}^d)$ to $L^2(dx)$. According to the resolvent formula (9), this implies that the resolvent difference $G^m_{\varepsilon,\alpha} - G_{\varepsilon,\alpha}$ is compact, and hence the corresponding essential spectra coincide.

3. Dependence on the coupling constant. In this section we are going to prove that

$$
-\Delta + \varepsilon^2\Delta^2 + m \longrightarrow -\Delta + m \qquad \text{as } \varepsilon \downarrow 0, \tag{24}
$$

in the norm-resolvent sense. Here m denotes a real-valued Radon measure on $\mathbb{R}^d$, and we assume, in addition, that for every $\eta > 0$ there exists a $\beta_\eta < \infty$ such that

$$
\int |f|^2 d|m| \leq \eta \left(\int |\nabla f|^2 dx + \beta_\eta \int |f|^2 dx \right), \qquad f \in C_0^\infty(\mathbb{R}^d). \tag{25}
$$

Note that we neither require that m is finite nor that $d \leq 3$. On the other hand, the condition (25) implies that $m(B) = 0$ for every Borel set B with classical capacity zero, and, for instance, it is excluded that m is a point measure if $d > 1$.

The inequality (25) holds, in particular, provided m belongs to the Kato class; i.e.,

$$
\begin{aligned}
\sup_{n\in\mathbb{Z}} |m|([n,n+1]) < \infty, &\qquad d = 1, \\
\lim_{\varepsilon\to 0} \sup_{x\in\mathbb{R}^2} \int_{B(x,\varepsilon)} |\log(|x-y|)|\, |m|(dy) = 0, &\qquad d = 2, \\
\lim_{\varepsilon\to 0} \sup_{x\in\mathbb{R}^3} \int_{B(x,\varepsilon)} \frac{1}{|x-y|}\, |m|(dy) = 0, &\qquad d = 3,
\end{aligned}
$$

with $B(x,\varepsilon)$ denoting the ball of radius ε centered at x (cf. [11, Theorem 3.1]). We refer to [7, Chapter 1.2] for additional examples of measures satisfying (25).

In general, the elements f in the form domain of $-\Delta$ do not possess a continuous representative $\tilde{f}$. Therefore we shall give a definition of $\mathcal{E}^m_\varepsilon$ different from the one in section 2 so that it works for all $\varepsilon \geq 0$. Of course, both definitions are equivalent in the special case of positive ε.

Since the space $C_0^\infty(\mathbb{R}^d)$ of smooth functions with compact support is dense in the Sobolev space $H^1(\mathbb{R}^d)$, there exists a unique bounded linear mapping $J_m : H^1(\mathbb{R}^d) \longrightarrow L^2(|m|)$ satisfying

$$J_m f = f, \qquad f \in C_0^\infty(\mathbb{R}^d)$$

(strictly speaking J_m maps the dx-equivalence class of the continuous function $\tilde{f} \in C_0^\infty(\mathbb{R}^d)$ to the $|m|$-equivalence class of $\tilde{f}$). We put

$$\begin{aligned} D(\mathcal{E}_\varepsilon^m) &:= D(\mathcal{E}_\varepsilon), \\ \mathcal{E}_\varepsilon^m(f,f) &:= \mathcal{E}_\varepsilon(f,f) + (A_m J_m f, J_m f)_{L^2(|m|)}, \qquad f \in D(\mathcal{E}_\varepsilon^m), \end{aligned}$$

where $D(\mathcal{E}_\varepsilon) = H^1(\mathbb{R}^d)$ for $\varepsilon = 0$, $D(\mathcal{E}_\varepsilon) = H^2(\mathbb{R}^d)$ otherwise, and

$$A_m h(x) := \begin{cases} h(x), & x \in B, \\ -h(x), & x \in \mathbb{R}^d \setminus B, \end{cases} \qquad h \in L^2(|m|),$$

with B being any Borel set such that $m^+(\mathbb{R}^d \setminus B) = 0 = m^-(B)$. By (25) and the KLMN theorem, the quadratic form $\mathcal{E}_\varepsilon^m$ in $L^2(dx)$ is lower semibounded and closed, and

$$\mathcal{E}_{\varepsilon,\beta_1}^m(f,f) \geq 0, \qquad f \in D(\mathcal{E}_\varepsilon^m).$$

Again, $-\Delta + \varepsilon^2\Delta^2 + m$ denotes the lower semibounded self-adjoint operator associated to $\mathcal{E}_\varepsilon^m$, and we put

$$R_{\varepsilon,\alpha}^m := (-\Delta + \varepsilon^2\Delta^2 + m + \alpha)^{-1},$$

provided the inverse operator exists. $G_{\varepsilon,\alpha}$ is defined the same way as in section 2.

One key for the proof of the convergence result (24) is the observation that one can decompose

$$\hat{g}_{\varepsilon,\alpha}(p) = \frac{c(\varepsilon)}{p^2 + \alpha(\varepsilon)} - \frac{c(\varepsilon)}{p^2 + \beta(\varepsilon)},$$

whenever $c(\varepsilon)$ is defined. The coefficients $-\alpha(\varepsilon)$ and $-\beta(\varepsilon)$ are the roots of the polynomial $\varepsilon^2 x^2 + x + \alpha$; a simple calculation yields

$$\begin{aligned} c(\varepsilon) &:= \frac{1}{\sqrt{1 - 4\varepsilon^2\alpha}} && \longrightarrow 1 \qquad \text{as } \varepsilon \downarrow 0, \\ \alpha(\varepsilon) &:= \frac{2\alpha}{1 + \sqrt{1 - 4\varepsilon^2\alpha}} && \longrightarrow \alpha \qquad \text{as } \varepsilon \downarrow 0, \\ \beta(\varepsilon) &:= \frac{1 + \sqrt{1 - 4\varepsilon^2\alpha}}{2\varepsilon^2} && \longrightarrow \infty \qquad \text{as } \varepsilon \downarrow 0. \end{aligned} \tag{26}$$

Using the parameters introduced above, we arrive at

$$G_{\varepsilon,\alpha} = c(\varepsilon) G_{0,\alpha(\varepsilon)} - c(\varepsilon) G_{0,\beta(\varepsilon)}. \tag{27}$$

In the proof of the convergence result (24) we will use again a Krein-like resolvent formula, this time using the one from [3]; cf. (29) below. First we need some preparation. Let $\alpha > 0$ and $\varepsilon \geq 0$. We introduce the operator $J_{m,\varepsilon,\alpha}$ from the Hilbert space $(D(\mathcal{E}_\varepsilon), \mathcal{E}_{\varepsilon,\alpha})$ to $L^2(|m|)$ as follows:

$$\begin{aligned} D(J_{m,\varepsilon,\alpha}) &:= D(\mathcal{E}_\varepsilon), \\ J_{m,\varepsilon,\alpha} f &:= J_m f, \qquad f \in D(J_{m,\varepsilon,\alpha}). \end{aligned}$$

By (25), the operator norm of $J_{m,\varepsilon,\alpha}$ is less than or equal to η provided $\alpha \geq \beta_\eta$. Thus we can choose $\alpha_0 > 0$ and $c < 1$ such that

$$\| J_{m,\varepsilon,\alpha} \|_{(D(\mathcal{E}_\varepsilon),\mathcal{E}_{\varepsilon,\alpha}),L^2(|m|)} \leq \sqrt{c}, \qquad \alpha \geq \alpha_0. \tag{28}$$

Due to (28), the hypothesis of Theorem 3 in [3] is satisfied, and the theorem implies that $-\alpha$ belongs to the resolvent set of $-\Delta + \varepsilon^2\Delta^2 + m$ and

$$R^m_{\varepsilon,\alpha} = G_{\varepsilon,\alpha} - (J_{m,\varepsilon,\alpha})^* A_m (1 + J_m J^*_{m,\varepsilon,\alpha} A_m)^{-1} J_m G_{\varepsilon,\alpha}, \qquad \alpha \geq \alpha_0. \tag{29}$$

In fact, we can write

$$J^*_{m,\varepsilon,\alpha'} = (J_m G_{\varepsilon,\alpha'})^*, \qquad \alpha' > 0, \tag{30}$$

since we have

$$(J^*_{m,\varepsilon,\alpha'} f, h) = \mathcal{E}_{\varepsilon,\alpha'}(J^*_{m,\varepsilon,\alpha'} f, G_{\varepsilon,\alpha'} h) = (f, J_{m,\varepsilon,\alpha'} G_{\varepsilon,\alpha'} h)_{L^2(|m|)} = ((J_m G_{\varepsilon,\alpha'})^* f, h)$$

for every $h \in L^2(dx)$, $\varepsilon \geq 0$, and $\alpha' > 0$.

THEOREM 6. *Let m be a real-valued Radon measure on $\mathbb{R}^d$ satisfying* (25)*. Then the operators $-\Delta + \varepsilon^2\Delta^2 + m$ converge to $-\Delta + m$ in the norm-resolvent sense as $\varepsilon \downarrow 0$.*

Proof. Both resolvents are written by means of Krein's formula (29), so we can compare the first and second terms separately. To see that $\| G_{\varepsilon,\alpha} - G_{0,\alpha} \|_{L^2(dx)}$ vanishes in the limit $\varepsilon \downarrow 0$ is simple. It is enough to use the first resolvent formula

$$G_{0,\alpha(\varepsilon)} - G_{0,\alpha} = (\alpha - \alpha(\varepsilon)) G_{0,\alpha} G_{0,\alpha(\varepsilon)} \tag{31}$$

and the fact that

$$\| G_{0,\alpha'} \|^2_{L^2(dx),H^1} \leq k(\alpha'), \qquad \alpha' > 0,$$

for some continuous function k vanishing at infinity (actually, $k(x) = 1/x^2$ for $x \leq 2$ and $k(x) = 1/(4(x-1))$ for $x > 2$). Then the decomposition (27) of $G_{\varepsilon,\alpha}$ and the asymptotic behavior (26) of $\alpha(\varepsilon), \beta(\varepsilon)$, and $c(\varepsilon)$ finish the argument.

The proof that the difference of second terms in Krein's formula also tends to zero as $\varepsilon \to 0$ can be reduced into two tasks

$$\begin{aligned} \| J_m G_{\varepsilon,\alpha} - J_m G_{0,\alpha} \|_{L^2(dx),L^2(|m|)} &\longrightarrow 0 \qquad \text{as } \varepsilon \downarrow 0, \\ \| (1 + J_m J^*_{m,\varepsilon,\alpha} A_m)^{-1} - (1 + J_m J^*_{m,0,\alpha} A_m)^{-1} \|_{L^2(|m|)} &\longrightarrow 0 \qquad \text{as } \varepsilon \downarrow 0. \end{aligned}$$

The argument for the first line is similar to the one we have presented above for $G_{\varepsilon,\alpha} - G_{0,\alpha}$; we only have to add that, by hypothesis (25), it follows that

$$\| J_m G_{0,\alpha'} \|^2_{L^2(dx),L^2(|m|)} \leq \max(1, \beta_1) k(\alpha'), \qquad \alpha' > 0, \tag{32}$$

where the function $k(\alpha')$ is defined as above.

To show the second line we choose any $\alpha > \alpha_0$; then from (28) we get

$$\| (1 + J_m J^*_{m,\varepsilon,\alpha} A_m)^{-1} \|_{L^2(|m|)} \leq \frac{1}{1-c}, \qquad \varepsilon \geq 0.$$

By the second resolvent identity

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

it is sufficient to prove that

$$\| J_m J_{m,\varepsilon,\alpha}^* - J_m J_{m,0,\alpha}^* \|_{L^2(|m|)} \longrightarrow 0 \qquad \text{as } \varepsilon \downarrow 0. \tag{33}$$

From (27) and (30) follows that

$$J_m J_{m,\varepsilon,\alpha}^* = c(\varepsilon) J_m (J_m G_{0,\alpha(\varepsilon)})^* - c(\varepsilon) J_m (J_m G_{0,\beta(\varepsilon)})^*;$$

note that $c(\varepsilon)$ is real for sufficiently small ε. Using this expression and (31) and (30), we get

$$\begin{aligned} &\| J_m J_{m,\varepsilon,\alpha}^* - J_m J_{m,0,\alpha}^* \|_{L^2(|m|)} \\ &\le \| (c(\varepsilon)-1) J_m (J_m G_{0,\alpha(\varepsilon)})^* \|_{L^2(|m|)} + \| J_m (J_m G_{0,\alpha(\varepsilon)})^* - J_m (J_m G_{0,\alpha})^* \|_{L^2(|m|)} \\ &\qquad + \| c(\varepsilon) J_m (J_m G_{0,\beta(\varepsilon)})^* \|_{L^2(|m|)} \\ &= \| (c(\varepsilon)-1) J_{m,0,\alpha(\varepsilon)} J_{m,0,\alpha(\varepsilon)}^* \|_{L^2(|m|)} + \| (\alpha - \alpha(\varepsilon)) J_m G_{0,\alpha} (J_m G_{0,\alpha(\varepsilon)})^* \|_{L^2(|m|)} \\ &\qquad + \| c(\varepsilon) J_{m,0,\beta(\varepsilon)} J_{m,0,\beta(\varepsilon)}^* \|_{L^2(|m|)}, \qquad \varepsilon > 0. \end{aligned}$$

According to (25), the mapping $\| J_{m,0,\alpha} J_{m,0,\alpha}^* \|_{L^2(|m|)}$ is locally bounded for $\alpha \in (0,\infty)$ and tends to zero as α tends to infinity. Since $\alpha(\varepsilon) \longrightarrow \alpha$, $c(\varepsilon) \longrightarrow 1$, and $\beta(\varepsilon) \longrightarrow \infty$ as $\varepsilon \downarrow 0$, this implies, in conjunction with (32), that (33) holds. □

Remark 7. By the proof above, $\| G_{\varepsilon,\alpha}^m - G_{0,\alpha}^m \|$ is upper bounded by an expression of the form $c \cdot (\varepsilon^2 + \eta(m,\varepsilon))$, where the finite constant c can be extracted from the proof and $\eta(m,\varepsilon)$ has to be chosen (and can be chosen) such that (25) holds with η and β replaced by $\eta(m,\varepsilon)$ and $\beta(\varepsilon)$, respectively.

4. Eigenvalues and eigenspaces of the approximating operators. In this section let $d \le 3$, and let m be a finite real-valued Radon measure satisfying (25) (e.g., let m be from the Kato class). By the two preceding convergence results, we can approximate the operator $-\Delta + m$ in $L^2(\mathbb{R}^d, dx)$ by operators of the form $-\Delta + \varepsilon^2 \Delta^2 + \mu$, where $\varepsilon > 0$ and μ is a point measure with mass at only finitely many points. Since the convergence is in the norm-resolvent sense, we can thus approximate the negative eigenvalues and corresponding eigenspaces of the former operator by the corresponding eigenvalues and eigenfunctions of the latter one. Note that we know from (23) and [5, Theorem 3.1] that the essential spectra coincide.

The following theorem shows how to compute the eigenvalues and corresponding eigenspaces of the approximating operators.

THEOREM 8. *Let $d \le 3$ and $\varepsilon > 0$. Let $\mu = \sum_{j=1}^N c_j \delta_{x_j}$, where $N \in \mathbb{N}$, $x_1, \dots, x_N$ are N distinct points in $\mathbb{R}^d$, and $c_1, \dots, c_N$ are real numbers different from zero. Then the following holds:*

(a) *The real number $-\alpha < 0$ is an eigenvalue of $-\Delta + \varepsilon^2 \Delta^2 + \mu$ if and only if*

$$\det \left(\frac{\delta_{jk}}{c_k} + g_{\varepsilon,\alpha}(x_j - x_k) \right)_{1 \le j,k \le N} = 0.$$

(b) *For every eigenvalue $-\alpha < 0$ the corresponding eigenfunctions have the following form:*

$$\sum_{k=1}^N h_k g_{\varepsilon,\alpha}(\cdot - x_k), \qquad (h_k)_{1 \le k \le N}^T \in \ker \left(\frac{\delta_{jk}}{c_k} + g_{\varepsilon,\alpha}(x_j - x_k) \right)_{1 \le j,k \le N}.$$

Proof. Since $D(\mathcal{E}_\varepsilon) = H^2(\mathbb{R}^d)$, the mapping J_μ can be understood as

$$J_\mu f := \tilde{f} \qquad |\mu| \text{ a.e.}, \quad f \in H^2(\mathbb{R}^d).$$

By (6), $\int g_{\varepsilon,\alpha}(\cdot - y)f(y)dy$ is the unique continuous representative of $G_{\varepsilon,\alpha}f$. Hence $J_\mu G_{\varepsilon,\alpha}$ is the integral operator from $L^2(dx)$ to $L^2(|\mu|)$ with kernel $g_{\varepsilon,\alpha}(x-y)$, and its inverse operator $(J_\mu G_{\varepsilon,\alpha})^*$ is the integral operator from $L^2(|\mu|)$ to $L^2(dx)$ with the same kernel. Thus we get

$$J_\mu(J_\mu G_{\varepsilon,\alpha})^* A_\mu h(x_j) = \sum_{k=1}^{N} c_k g_{\varepsilon,\alpha}(x_j - x_k)h(x_k), \qquad 1 \le j \le N, \tag{34}$$

for every $h \in L^2(|\mu|)$.

Due to Krein's formula (29), $-\alpha < 0$ belongs to the resolvent set of $(-\Delta+\varepsilon^2\Delta^2+\mu)$ provided $1 + J_\mu(J_\mu G_{\varepsilon,\alpha})^* A_\mu$ is bijective. Since $L^2(|\mu|)$ is finite-dimensional and we have expression (34), that is true if and only if

$$\lambda(\alpha) := \det(\delta_{jk} + c_k g_{\varepsilon,\alpha}(x_j - x_k))_{1\le j,k\le N} \neq 0,$$

with $\delta_{j,k}$ being the Kronecker delta. As $g_{\varepsilon,\alpha}(x)$ is a real analytic function of $\alpha \in (0,\infty)$ for every $x \in \mathbb{R}^d$, the function $\lambda(\alpha)$ is also real analytic on $(0,\infty)$. By (5), it is different from zero for all sufficiently large α. Thus the set of zeros on $(0,\infty)$ of this function is discrete.

Since $J_\mu G_{\varepsilon,\alpha}$ is surjective and $(J_\mu G_{\varepsilon,\alpha})^* A_\mu$ injective, the resolvent formula (29) implies that any $\alpha_0 > 0$ satisfying $\lambda(\alpha_0) = 0$ is a pole of $R^\mu_{\varepsilon,\alpha}$. Thus we have proved that $-\alpha_0$ is an eigenvalue of $-\Delta + \varepsilon^2\Delta^2 + \mu$ if and only if $\lambda(\alpha_0) = 0$. Finally, the expression

$$\det(\delta_{jk} + c_k g_{\varepsilon,\alpha}(x_j - x_k))_{1\le j,k\le N} = \Pi_{k=1}^N c_k \cdot \det(\delta_{jk}/c_k + g_{\varepsilon,\alpha}(x_j - x_k))_{1\le j,k\le N}$$

implies assertion (a).

By the preceding considerations and [3, Lemma 1],

$$h \mapsto (J_\mu G_{\varepsilon,\alpha})^* A_\mu h$$

is a linear bijective mapping from $\ker(1+J_\mu(J_\mu G_{\varepsilon,\alpha})^* A_\mu)$ onto $\ker(-\Delta+\varepsilon^2\Delta^2+\mu+\alpha)$. Assertion (b) follows from a simple algebraic calculation. □

Remark 9. Since the Hilbert space $L^2(|\mu|)$ is N-dimensional with $N < \infty$, the resolvent formula (29) implies that the difference $G^\mu_{\varepsilon,\alpha} - G_{\varepsilon,\alpha}$ is a finite rank operator with rank less than or equal to N. Thus the number, counting multiplicity, of negative eigenvalues of $-\Delta + \varepsilon^2\Delta^2 + \mu$ is less than or equal to N.

Let us illustrate the approximation by point measures on a simple example in dimension two. Suppose that measure m is a minus length measure supported by a circle of radius R; i.e., m is a constant and negative measure. This makes the choice of approximating point measures very simple: We spread equidistantly N points along the circle, and all of the points have the same coupling constant c:

$$c = -\frac{\gamma 2\pi R}{N}.$$

Due to the symmetry, the spectrum of $-\Delta+m$ is known; it consists of the essential spectrum $[0,\infty)$ and a finite number of negative eigenvalues, which are all except the

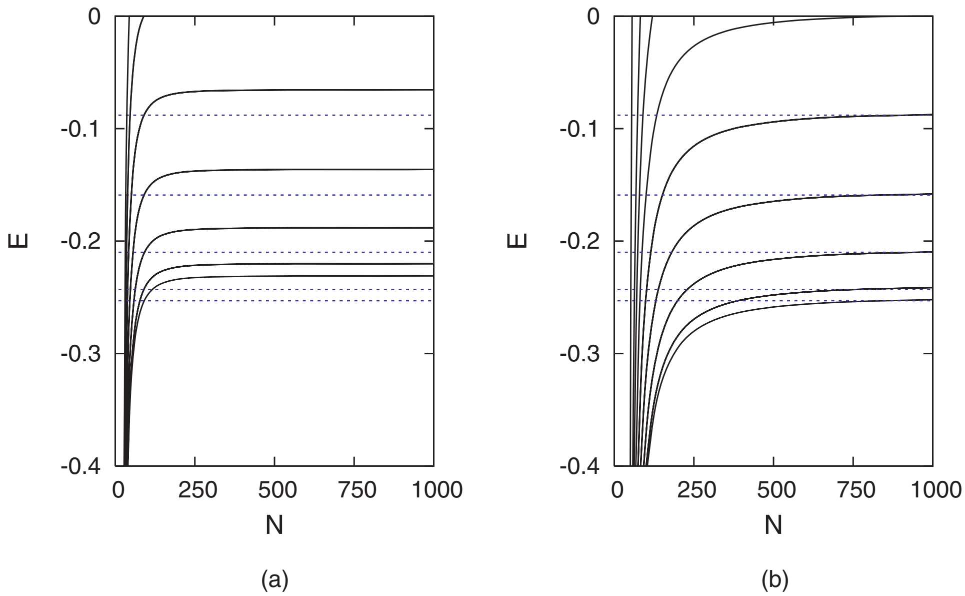


FIG. 1. *The dependence of the approximate eigenvalues on the number of point potentials for a circle with* $R = 10$ *and* $\varepsilon = 0.1$ (a), $\varepsilon = 0.01$ (b). *The dashed lines represent the exact eigenvalues of* $-\Delta + m$.

lowest one twice degenerate; see [2]. To find the eigenvalues, one has to decompose $L^2(\mathbb{R}^2)$ into angular momentum subspaces and then to look for solutions of an implicit equation in each of the subspaces. Therefore we can compute and compare both exact and approximate eigenvalues.

Each approximation is characterized by a pair of numbers, $\varepsilon > 0$ and $N \in \mathbb{N}$. In numerical calculations we fix ε, and we let N grow. The results for one chosen radius and two different parameters ε are depicted in Figure 1; cases (a) and (b) correspond to $\varepsilon = 0.1$ and $\varepsilon = 0.01$, respectively. We observe that below some threshold number of points the approximate discrete spectrum has no resemblance to the exact spectrum. The approximate eigenvalues may be very large negative, and their number may be much higher than the number of exact eigenvalues (in Figure 1, we have not even plotted all of the eigenvalues which exist only for small N).

It appears that, for larger ε, we get a fast convergence of eigenvalues; however, they are all shifted from the exact ones. The reason is that, since we work with fixed ε, the limit operator is, in fact, $-\Delta + \varepsilon^2\Delta^2 + m$ instead of $-\Delta + m$. On the contrary, small ε means that one needs more points to obtain a qualitatively correct spectrum, but then for a large number of points one gets much closer to the exact spectrum.

We can also compare this approximation to [8], where approximating operators were Laplacians with point potentials. Those point potentials are, of course, different: They are not defined via a quadratic form and cannot be understood as a special case $\varepsilon = 0$ of section 2; instead boundary conditons on wave functions are used; see [1]. Figure 2 presents the eigenvalues of Laplacians perturbed by point potentials which converge to $-\Delta + m$ with the same measure m as above. We have already mentioned in the introduction that, here, we obtain a stronger convergence result than the one in [8]. Moreover, comparing both Figures 1 and 2, we see that employing fourth-order differential operators in the approximation may improve significantly the spectral convergence.

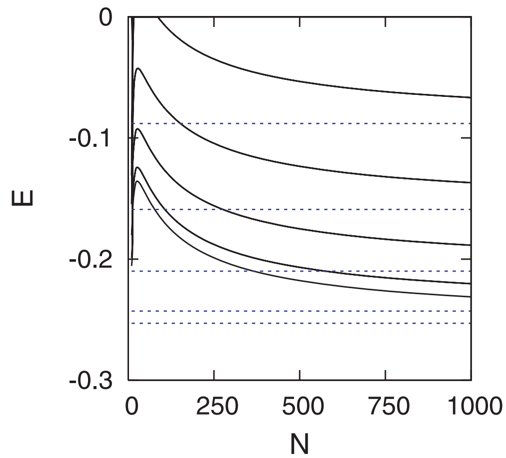


FIG. 2. *The dependence of the approximate eigenvalues on the number of point potentials for $R = 10$, using the standard two-dimensional point potentials. The dashed lines represent the exact eigenvalues of $-\Delta + m$.*

Appendix. In section 2 we have employed Krein's formula (9). Various forms of this formula can be found in the literature. Let us prove here the one we have used.

Let $f \in L^2(dx)$. Since $\mathcal{E}_\varepsilon$ and $\mathcal{E}_\varepsilon^\mu$ are associated to $-\Delta+\varepsilon^2\Delta^2$ and $-\Delta+\varepsilon^2\Delta^2+\mu$, respectively, it follows from Kato's representation theorem that

$$\mathcal{E}_{\varepsilon,\alpha}(G_{\varepsilon,\alpha}f, h) = (f, h) = \mathcal{E}^\mu_{\varepsilon,\alpha}((-\Delta + \varepsilon^2\Delta^2 + \mu + \alpha)^{-1}f, h) \tag{35}$$

for any $h \in H^2(\mathbb{R}^d)$ and $f \in L^2(\mathbb{R}^d)$. Moreover we have

$$\begin{aligned} \mathcal{E}_{\varepsilon,\alpha}(G^\mu_{\varepsilon,\alpha}\psi, h) &= (G^\mu_{\varepsilon,\alpha}\psi, (-\Delta + \varepsilon^2\Delta^2 + \alpha)h) \\ &= \int\int g_{\varepsilon,\alpha}(x-y)\bar{\tilde{\psi}}(y)\mu(dy)(-\Delta + \varepsilon^2\Delta^2 + \alpha)h(x)\,dx \\ &= \int\int g_{\varepsilon,\alpha}(x-y)(-\Delta + \varepsilon^2\Delta^2 + \alpha)h(x)\,dx\,\bar{\tilde{\psi}}(y)\mu(dy) \\ &= \int \tilde{h}\bar{\tilde{\psi}}\,\mu(dy), \qquad \psi \in H^2(\mathbb{R}^d), \quad h \in D(-\Delta + \varepsilon^2\Delta^2). \end{aligned} \tag{36}$$

We could change the order of integration in the second step. In fact, as $\mu^\pm$ are finite Radon measures and $g_{\varepsilon,\alpha}$ is square integrable w.r.t. the Lebesgue measure dx, the mappings $x \mapsto \int |g_{\varepsilon,\alpha}(x-y)|\mu^\pm(dy)$, $\mathbb{R}^d \longrightarrow \mathbb{R}$, are square integrable w.r.t. dx. Since $\tilde{\psi}$ is bounded and $(-\Delta + \varepsilon^2\Delta^2 + \alpha)h \in L^2(dx)$, it follows that

$$\int\int |g_{\varepsilon,\alpha}(x-y)\bar{\tilde{\psi}}(y)|\mu^\pm(dy)|(-\Delta + \varepsilon^2\Delta^2 + \alpha)h(x)|dx < \infty,$$

and, by Fubini's theorem, we could change the order of integration in the second step. In the last step we have used (6). Employing Sobolev's inequality and the fact that $D(-\Delta + \varepsilon^2\Delta^2)$ is dense in $(D(\mathcal{E}_\varepsilon), \mathcal{E}_{\varepsilon,\alpha})$, we can extend (36) to all functions ψ, $h \in D(\mathcal{E}_\varepsilon)$.

Put

$$\phi := G_{\varepsilon,\alpha}f - G^\mu_{\varepsilon,\alpha}(I + G^\mu_{\varepsilon,\alpha})^{-1}G_{\varepsilon,\alpha}f.$$

Then $\phi \in H^2(\mathbb{R}^d) = D(\mathcal{E}^\mu_\varepsilon)$, and (35) and extended (36) yield

$$\begin{aligned}\mathcal{E}^\mu_{\varepsilon,\alpha}(\phi,h) &= \mathcal{E}_{\varepsilon,\alpha}(G_{\varepsilon,\alpha}f,h) - \mathcal{E}_{\varepsilon,\alpha}(G^\mu_{\varepsilon,\alpha}(I+G^\mu_{\varepsilon,\alpha})^{-1}G_{\varepsilon,\alpha}f,h) \\ &\quad + \int [G_{\varepsilon,\alpha}f - G^\mu_{\varepsilon,\alpha}(I+G^\mu_{\varepsilon,\alpha})^{-1}G_{\varepsilon,\alpha}\tilde{\tilde{f}}]\tilde{h}\,d\mu \\ &= (f,h) - \int [(I+G^\mu_{\varepsilon,\alpha})^{-1}G_{\varepsilon,\alpha}\tilde{\tilde{f}}]\tilde{h}\,d\mu \\ &\quad + \int [(I+G^\mu_{\varepsilon,\alpha})(I+G^\mu_{\varepsilon,\alpha})^{-1}G_{\varepsilon,\alpha}f - G^\mu_{\varepsilon,\alpha}(I+G^\mu_{\varepsilon,\alpha})^{-1}G_{\varepsilon,\alpha}\tilde{\tilde{f}}]\tilde{h}\,d\mu \\ &= (f,h), \qquad h \in H^2(\mathbb{R}^d).\end{aligned}$$

Due to (10), $\mathcal{E}^\mu_{\varepsilon,\alpha}$ is a scalar product on $D(\mathcal{E}^\mu_{\varepsilon,\alpha}) = H^2(\mathbb{R}^d)$. Thus (35) and the calculation above imply that $\phi = (-\Delta + \varepsilon^2\Delta^2 + \mu + \alpha)^{-1}f$.

REFERENCES

[1] S. Albeverio, F. Gesztesy, R. Høegh–Krohn, and H. Holden, *Solvable Models in Quantum Mechanics*, 2nd ed., AMS Chelsea, Providence, RI, 2005.

[2] J.-P. Antoine, F. Gesztesy, and J. Shabani, *Exactly solvable models of sphere interactions in quantum mechanics*, J. Phys. A, 20 (1987), pp. 3687–3712.

[3] J. F. Brasche, *On the spectral properties of singular perturbed operators*, in Stochastic Processes and Dirichlet Forms, Z. Ma, M. Röckner, and A. Yan, eds., de Gruyter, Berlin, 1995, pp. 65–72.

[4] J. F. Brasche, *Upper bounds for Neumann – Schatten norms*, Potential Anal., 14 (2001), pp. 175–205.

[5] J. F. Brasche, P. Exner, Y. Kuperin, and P. Šeba, *Schrödinger operators with singular interactions*, J. Math. Anal. Appl., 183 (1994), pp. 112–139.

[6] J. F. Brasche, R. Figari, and A. Teta, *Singular Schrödinger operators as limits of point interaction Hamiltonians*, Potential Anal., 8 (1998), pp. 163–178.

[7] H. L. Cycon, R. G. Froese, W. Kirsch, and B. Simon, *Schrödinger Operators*, Springer, Berlin, 1987.

[8] P. Exner and K. Němcová, *Leaky quantum graphs: Approximations by point interaction Hamiltonians*, J. Phys. A, 36 (2003), pp. 10173–10193.

[9] K. S. Murigi, *On Eigenvalues of Schrödinger Operators with δ-Potentials*, Master's thesis, Chalmers University of Technology, Göteborg, 2004.

[10] A. Posilicano, *Convergence of distorted Brownian motions and singular Hamiltonians*, Potential Anal., 5 (1996), pp. 241–271.

[11] P. Stollmann and J. Voigt, *Perturbation of Dirichlet forms by measures*, Potential Anal., 5 (1996), pp. 109–138.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 298–318

OPTIMALLY SPARSE MULTIDIMENSIONAL REPRESENTATION USING SHEARLETS*

KANGHUI GUO† AND DEMETRIO LABATE‡

Abstract. In this paper we show that shearlets, an affine-like system of functions recently introduced by the authors and their collaborators, are essentially optimal in representing 2-dimensional functions f which are C^2 except for discontinuities along C^2 curves. More specifically, if f_N^S is the N-term reconstruction of f obtained by using the N largest coefficients in the shearlet representation, then the asymptotic approximation error decays as $\|f - f_N^S\|_2^2 \asymp N^{-2} (\log N)^3$, $N \to \infty$, which is essentially optimal, and greatly outperforms the corresponding asymptotic approximation rate N^{-1} associated with wavelet approximations. Unlike curvelets, which have similar sparsity properties, shearlets form an affine-like system and have a simpler mathematical structure. In fact, the elements of this system form a Parseval frame and are generated by applying dilations, shear transformations, and translations to a single well-localized window function.

Key words. affine systems, curvelets, geometric image processing, shearlets, sparse representation, wavelets

AMS subject classifications. 42C15, 42C40

DOI. 10.1137/060649781

1. Introduction.

The notion of efficient representation of data plays an increasingly important role in areas across applied mathematics, science, and engineering. Over the past few years, there has been a rapidly increasing pressure to handle ever larger and higher-dimensional data sets, with the challenge of providing representations of these data that are sparse (that is, "very" few terms of the representation are sufficient to accurately approximate the data) and computationally fast. Sparse representations have implications reaching beyond data compression. Understanding the compression problem for a given data type entails a precise knowledge of the modeling and approximation of that data type. This in turn is useful for many other important tasks, including classification, denoising, interpolation, and segmentation [13].

Multiscale techniques based on wavelets have emerged over the last two decades as the most successful methods for the efficient representation of data, as attested, for example, by their use in the new FBI fingerprint database [3] and in JPEG2000, the new standard for image compression [4, 19]. Indeed, wavelets are optimally efficient in representing functions with pointwise singularities [27, Chap. 9].

More specifically, consider the wavelet representation (using a "nice" wavelet basis) of a function f of a *single variable* that is smooth apart from a point discontinuity. Because the elements of the wavelet basis are well localized (i.e., they have very fast decay both in the spatial and in the frequency domain), very few of them interact significantly with the singularity, and thus very few elements of the wavelet expansion are sufficient to provide an accurate approximation. This contrasts sharply with the

*Received by the editors January 12, 2006; accepted for publication (in revised form) January 2, 2007; published electronically May 29, 2007.
http://www.siam.org/journals/sima/39-1/64978.html
†Department of Mathematics, Missouri State University, Springfield, MO 65804 (kag026f@smsu.edu).
‡Department of Mathematics, North Carolina State University, Campus Box 8205, Raleigh, NC 27695 (dlabate@math.ncsu.edu). This author acknowledges support from an NCSU FR&PD grant and from National Science Foundation grant DMS 0604561.

Fourier representation, for which the discontinuity interacts extensively with the elements of the Fourier basis. Denoting by f_N the approximation obtained by using the largest N coefficients in the wavelet expansion, the asymptotic approximation error satisfies

$$\|f - f_N\|_2^2 \asymp N^{-2}, \qquad N \to \infty.$$

This is the optimal approximation rate for this type of function [10], and outperforms the corresponding Fourier approximation error rate N^{-1} [13, 27]. In addition, the multiresolution analysis (MRA) associated with wavelets provides very fast numerical algorithms for computing the wavelet coefficients [27].

However, despite their remarkable success in applications, wavelets are far from optimal in dimensions larger than one. Indeed wavelets are very efficient in dealing with pointwise singularities *only*. In higher dimensions other types of singularities are usually present or even dominant, and wavelets are unable to handle them very efficiently. Consider, for example, the wavelet representation of a 2-dimensional (2-D) function that is C^2 away from a discontinuity along a curve of finite length (a reasonable model for an image containing an edge). Because the discontinuity is spatially distributed, it interacts extensively with the elements of the wavelet basis. As a consequence, the wavelet coefficients have a slow decay, and the approximation error $\|f - f_N\|_2^2$ decays at most as fast as $O(N^{-1})$ [27]. This is better than the rate of the Fourier approximation error $N^{-1/2}$, but far from the optimal theoretical approximation rate (cf. [12])

$$\|f - f_N\|_2^2 \asymp N^{-2}, \qquad N \to \infty. \tag{1.1}$$

There is, therefore, large room for improvements, and several attempts have been made in this direction both in the mathematical and engineering communities in recent years. Those include *contourlets*, *complex wavelets* and other "directional wavelets" in the filter bank literature [1, 2, 11, 22, 26, 28], as well as *brushlets* [8], *ridgelets* [5], *curvelets* [7], and *bandelets* [24].

The most successful approach so far are the curvelets of Candès and Donoho. This is the first and so far the only construction providing an essentially optimal approximation property for 2-D piecewise smooth functions with discontinuities along C^2 curves [7]. The main idea in the curvelet approach is that, in order to approximate functions with edges accurately, one has to exploit their geometric regularity much more efficiently than traditional wavelets. This is achieved by constructing an appropriate tight frame of well-localized functions at various scales, positions, and directions. We refer to [6, 7] for more details about this construction.

The main goal of this paper is to show that shearlets, a construction based on the theory of composite wavelets, also provides an essentially optimal approximation property for 2-D piecewise C^2 functions with discontinuities along C^2 curves. We will show that the approximation error associated with the N-term reconstruction f_N^S obtained by taking the N largest coefficients in the shearlet expansion satisfies

$$\|f - f_N^S\|_2^2 \asymp N^{-2}(\log N)^3, \qquad N \to \infty. \tag{1.2}$$

This is exactly the approximation rate obtained using curvelets. The proof of our result adapts several ideas from the corresponding sparsity result of the curvelets [7] and follows the general architecture of that proof, but does not follow directly from the curvelets construction. Indeed, as we will argue in the following, our alternative approach based on shearlets has some mathematical advantages with respect to curvelets, including a simplified construction that provides the framework for a simpler mathematical analysis and fast algorithmic implementation (see also [9, 14]).

The theory of composite wavelets, recently proposed by the authors and their collaborators [16, 17, 18], provides a most general setting for the construction of truly multidimensional, efficient, multiscale representations. Unlike the curvelets, this approach takes full advantage of the theory of affine systems on $\mathbb{R}^n$. Specifically, the *affine systems with composite dilations* are the systems

$$\mathcal{A}_{AB}(\psi) = \{\psi_{j,\ell,k}(x) = |\det A|^{j/2}\,\psi(B^\ell A^j x - k) : \, j,\ell \in \mathbb{Z}, k \in \mathbb{Z}^n\}, \tag{1.3}$$

where A, B are $n \times n$ invertible matrices and $|\det B| = 1$. The elements of this system are called *composite wavelets* if $\mathcal{A}_{AB}(\psi)$ forms a *Parseval frame* (also called a tight frame) for $L^2(\mathbb{R}^n)$; that is,

$$\sum_{j,\ell,k} |\langle f, \psi_{j,\ell,k}\rangle|^2 = \|f\|^2$$

for all $f \in L^2(\mathbb{R}^n)$. The shearlets, which will be considered in this paper, are a special class of composite wavelets where A is an anisotropic dilation and B is a shear matrix. Details for this construction will be given in section 1.2. These representations have fully controllable geometrical features, such as orientations, scales, and shapes, which set them apart from traditional wavelets as well as complex and directional wavelets. In addition, thanks to their mathematical structure, there is a multiresolution analysis naturally associated with composite wavelets. This is particularly useful for the development of fast algorithmic implementations of these transformations [23, 25].

Observe that curvelets are not of the form (1.3), and, unlike the shearlets, are not generated from the action of a family of operators on a single or finite family of functions.

1.1. Notation. Throughout this paper, we shall consider the points $x \in \mathbb{R}^n$ to be column vectors, i.e.,

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

and the points $\xi \in \widehat{\mathbb{R}}^n$ (the frequency domain) to be row vectors, i.e., $\xi = (\xi_1, \dots, \xi_n)$. A vector x multiplying a matrix $a \in GL_n(\mathbb{R})$ on the right is understood to be a column vector, while a vector ξ multiplying a on the left is a row vector. Thus, $ax \in \mathbb{R}^n$ and $\xi a \in \widehat{\mathbb{R}}^n$. The Fourier transform is defined as

$$\hat{f}(\xi) = \int_{\mathbb{R}^n} f(x)\, e^{-2\pi i \xi x}\, dx,$$

where $\xi \in \widehat{\mathbb{R}}^n$, and the inverse Fourier transform is

$$\check{f}(x) = \int_{\widehat{\mathbb{R}}^n} f(\xi)\, e^{2\pi i \xi x}\, d\xi.$$

1.2. Shearlets. The collection of shearlets that we are going to define in this section is a special example of composite wavelets in $L^2(\mathbb{R}^2)$, of the form (1.3), where

$$A = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \tag{1.4}$$

and ψ will be defined in the following. It is useful to observe that, by applying the Fourier transform to the elements $\psi_{j,\ell,k}$ in (1.3), we obtain

$$\hat{\psi}_{j,\ell,k}(\xi) = |\det A|^{-j/2}\,\psi(\xi A^{-j} B^{-\ell})\, e^{2\pi i \xi A^{-j} B^{-\ell} k}.$$

For any $\xi = (\xi_1, \xi_2) \in \widehat{\mathbb{R}}^2$, $\xi_1 \neq 0$, let ψ be given by

$$\hat{\psi}(\xi) = \hat{\psi}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_1)\,\hat{\psi}_2\left(\frac{\xi_2}{\xi_1}\right), \tag{1.5}$$

where $\hat{\psi}_1, \hat{\psi}_2 \in C^\infty(\widehat{\mathbb{R}})$, $\operatorname{supp}\hat{\psi}_1 \subset [-\frac{1}{2}, -\frac{1}{16}] \cup [\frac{1}{16}, \frac{1}{2}]$, and $\operatorname{supp}\hat{\psi}_2 \subset [-1, 1]$. We assume that

$$\sum_{j \geq 0} |\hat{\psi}_1(2^{-2j}\omega)|^2 = 1 \quad \text{for } |\omega| \geq \frac{1}{8} \tag{1.6}$$

and

$$|\hat{\psi}_2(\omega - 1)|^2 + |\hat{\psi}_2(\omega)|^2 + |\hat{\psi}_2(\omega + 1)|^2 = 1 \quad \text{for } |\omega| \leq 1. \tag{1.7}$$

It follows from the last equation that, for any $j \geq 0$,

$$\sum_{\ell=-2^j}^{2^j} |\hat{\psi}_2(2^j\,\omega + \ell)|^2 = 1 \quad \text{for } |\omega| \leq 1. \tag{1.8}$$

It also follows from our assumptions that $\hat{\psi} \in C_0^\infty(\widehat{\mathbb{R}}^2)$, with $\operatorname{supp}\hat{\psi} \subset [-\frac{1}{2}, \frac{1}{2}]^2$. There are several examples of functions ψ_1, ψ_2 satisfying the properties described above (see the appendix).

Observe that $(\xi_1, \xi_2)\, A^{-j} B^{-\ell} = (2^{-2j}\xi_1, -\ell 2^{-2j}\xi_1 + 2^{-j}\xi_2)$. Using (1.6) and (1.8), it is easy to see that

$$\begin{aligned}\sum_{j \geq 0} \sum_{\ell=-2^j}^{2^j} |\hat{\psi}(\xi\, A^{-j} B^{-\ell})|^2 &= \sum_{j \geq 0} \sum_{\ell=-2^j}^{2^j} |\hat{\psi}_1(2^{-2j}\,\xi_1)|^2 \left|\hat{\psi}_2\left(2^j\,\frac{\xi_2}{\xi_1} - \ell\right)\right|^2 \\ &= \sum_{j \geq 0} |\hat{\psi}_1(2^{-2j}\,\xi_1)|^2 \sum_{\ell=-2^j}^{2^j} \left|\hat{\psi}_2\left(2^j\,\frac{\xi_2}{\xi_1} - \ell\right)\right|^2 = 1\end{aligned}$$

for $(\xi_1, \xi_2) \in \mathcal{D}_C$, where $\mathcal{D}_C = \{(\xi_1, \xi_2) \in \widehat{\mathbb{R}}^2 : |\xi_1| \geq \frac{1}{8}, |\frac{\xi_2}{\xi_1}| \leq 1\}$. This equation, together with the fact that $\hat{\psi}$ is supported inside $[-\frac{1}{2}, \frac{1}{2}]^2$, implies that the collection of *shearlets*,

$$\mathcal{SH}(\psi) = \left\{\psi_{j,\ell,k}(x) = 2^{\frac{3j}{2}}\,\psi(B^\ell A^j x - k) : j \geq 0, -2^j \leq \ell \leq 2^j, k \in \mathbb{Z}^2\right\}, \tag{1.9}$$

is a Parseval frame for $L^2(\mathcal{D}_C)^\vee = \{f \in L^2(\mathbb{R}^2) : \operatorname{supp}\hat{f} \subset \mathcal{D}_C\}$. Details about the argument that this system is a Parseval frame can be found in [18, sect. 5.2.1].

To obtain a Parseval frame for $L^2(\mathbb{R}^2)$, one can construct a second system of shearlets which form a Parseval frame for the functions with frequency support in the vertical cone $\mathcal{D}_{\widetilde{C}} = \{(\xi_1, \xi_2) \in \widehat{\mathbb{R}}^2 : |\xi_2| \geq \frac{1}{8}, |\frac{\xi_1}{\xi_2}| \leq 1\}$. Finally, one can easily

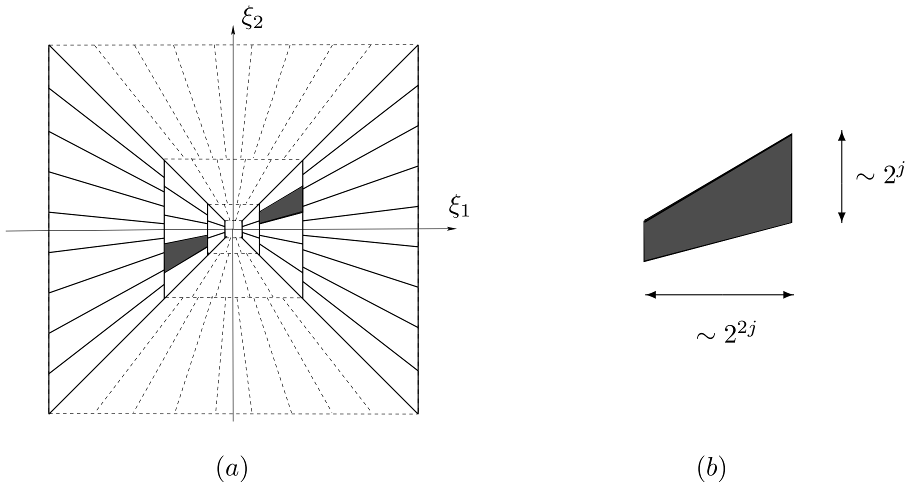


FIG. 1.1. (a) *The tiling of the frequency plane* $\widehat{\mathbb{R}}^2$ *induced by the shearlets.* (b) *Frequency support of the shearlet* $\psi_{j,\ell,k}$*, for* $\xi_1 > 0$*. The other half of the support, for* $\xi_1 < 0$*, is symmetrical.*

construct a Parseval frame (or an orthonormal basis) for $L^2([-\frac{1}{8}, \frac{1}{8}]^2)^\vee$. Then any function in $L^2(\mathbb{R}^2)$ can be expressed as a sum $f = P_C f + P_{\widetilde{C}} f + P_0 f$, where each component corresponds to the orthogonal projection of f into one of the three subspaces of $L^2(\mathbb{R}^2)$ described above. The tiling of the frequency plane $\widehat{\mathbb{R}}^2$ induced by this system is illustrated in Figure 1.1(a). The above construction was first introduced in [15].

The conditions on the support of $\hat{\psi}_1$ and $\hat{\psi}_2$ imply that the functions $\hat{\psi}_{j,\ell,k}$ have frequency support:

$$\operatorname{supp} \hat{\psi}_{j,\ell,k} \subset \left\{ (\xi_1, \xi_2) : \xi_1 \in [-2^{2j-1}, -2^{2j-4}] \cup [2^{2j-4}, 2^{2j-1}], \left| \frac{\xi_2}{\xi_1} - \ell\, 2^{-j} \right| \leq 2^{-j} \right\}.$$

Thus, the system $\mathcal{SH}(\psi)$, given by (1.9), is a Parseval frame exhibiting the following properties:

(i) *Time-frequency localization.* Since $\hat{\psi} \in C_0^\infty(\widehat{\mathbb{R}}^2)$, then $|\psi(x)| \leq C_N\,(1+|x|)^{-N}$ for any $N \in \mathbb{N}$, and thus the elements $\psi_{j,\ell,k}$ are well localized.

(ii) *Parabolic scaling.* Each element $\hat{\psi}_{j,\ell,k}$ has support on a pair of trapezoids, each one contained in a box of size approximately $2^{2j} \times 2^j$ (see Figure 1.1(b)). Because the shearlets are well localized, each $\psi_{j,\ell,k}$ is essentially supported on a box of size $2^{-2j} \times 2^{-j}$. Thus, their supports become increasingly thin as $j \to \infty$.

(iii) *Directional sensitivity.* The elements $\hat{\psi}_{j,\ell,k}$ are oriented along lines with slope given by $\ell\, 2^{-j}$. As a consequence, the corresponding elements $\psi_{j,\ell,k}$ are oriented along lines with slope $-\ell\, 2^{-j}$. The number of orientations (approximately) doubles at each finer scale.

(iv) *Spatial localization.* For any fixed scale and orientation, the shearlets are obtained by translations on the lattice $\mathbb{Z}^2$.

(v) *Oscillatory behavior.* The shearlets $\psi_{j,\ell,k}$ are nonoscillatory along the orientation axis of slope $-\ell\, 2^{-j}$, and oscillatory across this axis.

Observe that the curvelets of Candès and Donoho also satisfy similar properties, with the following main differences. Concerning property (iii), the number of orientations in the curvelet constructions doubles at every other scale. Concerning property (iv), the curvelets are not associated with a *fixed* translation lattice. However, for a given scale parameter j and orientation θ, they are obtained by translations on a grid that depends on j and θ. In fact, as we mentioned before, unlike the shearlets, the curvelets are not generated from the action of a family of operators on a single or finite family of functions.

1.3. Main results. One major feature of shearlets is that, if f is a compactly supported function which is C^2 away from a C^2 curve, then the sequence of shearlet coefficients $\{\langle f, \psi_{j,\ell,k}\rangle\}$ has (essentially) optimally fast decay. As a consequence, if f_N^S is the N-term approximation of f obtained from the N largest coefficients of its shearlet expansion, then the approximation error is essentially optimal.

Before stating the main theorems, let us define more precisely the class of functions we are interested in. We follow [7] and introduce $STAR^2(A)$, a class of indicator functions of sets B with C^2 boundaries ∂B. In polar coordinates, let $\rho(\theta) : [0, 2\pi) \to [0, 1]^2$ be a radius function, and define B by $x \in B$ if and only if $|x| \leq \rho(\theta)$. In particular, the boundary ∂B is given by the curve in $\mathbb{R}^2$:

$$\beta(\theta) = \begin{pmatrix} \rho(\theta)\,\cos(\theta) \\ \rho(\theta)\,\sin(\theta) \end{pmatrix}. \tag{1.10}$$

The class of boundaries of interest to us is defined by

$$\sup|\rho''(\theta)| \leq A, \quad \rho \leq \rho_0 < 1. \tag{1.11}$$

We say that a set $B \in STAR^2(A)$ if $B \subset [0, 1]^2$ and B is a translate of a set obeying (1.10) and (1.11). In addition, we set $C_0^2([0, 1]^2)$ to be the collection of twice differentiable functions supported inside $[0, 1]^2$.

Finally, we define the set $\mathcal{E}^2(A)$ of *functions which are* C^2 *away from a* C^2 *edge* as the collection of functions of the form

$$f = f_0 + f_1\,\chi_B,$$

where $f_0, f_1 \in C_0^2([0, 1]^2)$, $B \in STAR^2(A)$, and $\|f\|_{C^2} = \sum_{|\alpha|\leq 2}\|D^\alpha f\|_\infty \leq 1$.

Let M be the set of indices $\{(j, \ell, k) : j \geq 0, -2^j \leq \ell \leq 2^j, k \in \mathbb{Z}^2\}$, and let $\{\psi_\mu\}_{\mu\in M}$ be the Parseval frame of shearlets given by (1.9). The *shearlet coefficients* of a given function f are the elements of the sequence $\{s_\mu(f) = \langle f, \psi_\mu\rangle : \mu \in M\}$. We denote by $|s(f)|_{(N)}$ the Nth largest entry in this sequence. We can now state the following results.

THEOREM 1.1. *Let $f \in \mathcal{E}^2(A)$, and let $\{s_\mu(f) = \langle f, \psi_\mu\rangle : \mu \in M\}$ be the sequence of shearlet coefficients associated with f. Then*

$$\sup_{f\in\mathcal{E}^2(A)} |s(f)|_{(N)} \leq C\,N^{-3/2}\,(\log N)^{3/2}. \tag{1.12}$$

Let f_N^S be the N-term approximation of f obtained from the N largest coefficients of its shearlet expansion, namely,

$$f_N^S = \sum_{\mu\in I_N} \langle f, \psi_\mu\rangle\,\psi_\mu,$$

where $I_N \subset M$ is the set of indices corresponding to the N largest entries of the sequence $\{|\langle f, \psi_\mu\rangle|^2 : \mu \in M\}$. Then the approximation error satisfies

$$\|f - f_N^S\|_2^2 \leq \sum_{m>N} |s(f)|_{(m)}^2.$$

Therefore, from (1.12) we immediately have the following.

THEOREM 1.2. *Let $f \in \mathcal{E}^2(A)$ and f_N^S be the approximation to f defined above. Then*

$$\|f - f_N^S\|_2^2 \le C\, N^{-2}\, (\log N)^3.$$

1.4. Analysis of the shearlet coefficients. The argument that will be used to prove Theorem 1.1 follows essentially the architecture of the proofs in [7]. In order to measure the sparsity of the shearlet coefficients $\{\langle f, \psi_\mu \rangle : \mu \in M\}$, we will use the *weak-$\ell^p$ quasi-norm* $\|\cdot\|_{w\ell^p}$ defined as follows. Let $|s_\mu|_{(N)}$ be the Nth largest entry in the sequence $\{s_\mu\}$. Then

$$\|s_\mu\|_{w\ell^p} = \sup_{N>0} N^{\frac{1}{p}}\, |s_\mu|_{(N)}.$$

One can show (cf. [29, sect. 5.3]) that this definition is equivalent to

$$\|s_\mu\|_{w\ell^p} = \left(\sup_{\epsilon>0} \#\{\mu : |s_\mu| > \epsilon\}\, \epsilon^p \right)^{\frac{1}{p}}.$$

To analyze the decay properties of the shearlet coefficients $\{\langle f, \psi_\mu \rangle\}$ at a given scale 2^{-j}, we will smoothly localize the function f near dyadic squares. Fix the scale parameter $j \ge 0$. For this j fixed, let $M_j = \{(j,\ell,k) : -2^j \le \ell \le 2^j, k \in \mathbb{Z}^2\}$ and $\mathcal{Q}_j$ be the collection of dyadic cubes of the form $Q = [\frac{k_1}{2^j}, \frac{k_1+1}{2^j}] \times [\frac{k_2}{2^j}, \frac{k_2+1}{2^j}]$, with $k_1, k_2 \in \mathbb{Z}$. For w a nonnegative C^∞ function with support in $[-1,1]^2$, we define a smooth partition of unity

$$\sum_{Q \in \mathcal{Q}_j} w_Q(x) = 1, \quad x \in \mathbb{R}^2,$$

where, for each dyadic square $Q \in \mathcal{Q}_j$, $w_Q(x) = w(2^j x_1 - k_1, 2^j x_2 - k_2)$. We will then examine the shearlet coefficients of the localized function $f_Q = f\, w_Q$, i.e., $\{\langle f_Q, \psi_\mu \rangle : \mu \in M_j\}$.

For $f \in \mathcal{E}^2(A)$, the decay properties of the coefficients $\{\langle f_Q, \psi_\mu \rangle : \mu \in M_j\}$ will exhibit a very different behavior depending on whether the edge curve intersects the support of w_Q or not. Let $\mathcal{Q}_j = \mathcal{Q}_j^0 \cup \mathcal{Q}_j^1$, where the union is disjoint and $\mathcal{Q}_j^0$ is the collection of those dyadic cubes $Q \in \mathcal{Q}_j$ such that the edge curve intersects the support of w_Q. Since each Q has sidelength $2 \cdot 2^{-j}$, then $\mathcal{Q}_j^0$ has cardinality $|\mathcal{Q}_j^0| \le C_0\, 2^j$, where C_0 is independent of j. Similarly, since f is compactly supported in $[0,1]^2$, $|\mathcal{Q}_j^1| \le 2^{2j} + 4 \cdot 2^j$.

We have the following results, which will be proved in section 2.

THEOREM 1.3. *Let $f \in \mathcal{E}^2(A)$. For $Q \in \mathcal{Q}_j^0$, with $j \ge 0$ fixed, the sequence of shearlet coefficients $\{\langle f_Q, \psi_\mu \rangle : \mu \in M_j\}$ obeys*

$$\|\langle f_Q, \psi_\mu \rangle\|_{w\ell^{2/3}} \le C\, 2^{-\frac{3j}{2}}$$

for some constant C independent of Q and j.

THEOREM 1.4. *Let $f \in \mathcal{E}^2(A)$. For $Q \in \mathcal{Q}_j^1$, with $j \ge 0$ fixed, the sequence of shearlet coefficients $\{\langle f_Q, \psi_\mu \rangle : \mu \in M_j\}$ obeys*

$$\|\langle f_Q, \psi_\mu \rangle\|_{w\ell^{2/3}} \le C\, 2^{-3j}$$

for some constant C independent of Q and j.

As a consequence of these two theorems, we have the following.

COROLLARY 1.5. *Let $f \in \mathcal{E}^2(A)$ and, for $j \geq 0$, let $s_j(f)$ be the sequence $s_j(f) = \{\langle f, \psi_\mu \rangle : \mu \in M_j\}$. Then*

$$\|s_j(f)\|_{w\ell^{2/3}} \leq C$$

for some C independent of j.

Proof. Using Theorems 1.3 and 1.4, by the p-triangle inequality for weak ℓ^p spaces, $p \leq 1$, we have

$$\begin{aligned} \|s_j(f)\|_{w\ell^{2/3}}^{2/3} &\leq \sum_{Q \in \mathcal{Q}_j} \|\langle f_Q, \psi_\mu \rangle\|_{w\ell^{2/3}}^{2/3} \\ &= \sum_{Q \in \mathcal{Q}_j^0} \|\langle f_Q, \psi_\mu \rangle\|_{w\ell^{2/3}}^{2/3} + \sum_{Q \in \mathcal{Q}_j^1} \|\langle f_Q, \psi_\mu \rangle\|_{w\ell^{2/3}}^{2/3} \\ &\leq C\,|\mathcal{Q}_j^0|\,2^{-j} + C\,|\mathcal{Q}_j^1|\,2^{-2j}. \end{aligned}$$

The proof is completed by observing that $|\mathcal{Q}_j^0| \leq C_0\, 2^j$, where C_0 is independent of j, and $|\mathcal{Q}_j^1| \leq 2^{2j} + 4 \cdot 2^j$. ☐

We can now prove Theorem 1.1.

Proof of Theorem 1.1. By Corollary 1.5, we have that

$$R(j, \epsilon) = \#\{\mu \in M_j : |\langle f, \psi_\mu \rangle| > \epsilon\} \leq C\, \epsilon^{-2/3}. \tag{1.13}$$

Also, observe that, since $\hat{\psi} \in C_0^\infty(\mathbb{R}^2)$, then

$$\begin{aligned} |\langle f, \psi_\mu \rangle| &= \left| \int_{\mathbb{R}^2} f(x)\, 2^{3j/2}\, \psi(B^\ell A^j x - k)\, dx \right| \\ &\leq 2^{3j/2}\, \|f\|_\infty \int_{\mathbb{R}^2} |\psi(B^\ell A^j x - k)|\, dx \\ &= 2^{-3j/2}\, \|f\|_\infty \int_{\mathbb{R}^2} |\psi(y)|\, dy < C'\, 2^{-3j/2}. \end{aligned} \tag{1.14}$$

As a consequence, there is a scale j_ϵ such that $|\langle f, \psi_\mu \rangle| < \epsilon$ for each $j \geq j_\epsilon$. Specifically, it follows from (1.14) that $R(j, \epsilon) = 0$ for $j > \frac{2}{3}\,(\log_2(\epsilon^{-1}) + \log_2(C')) > \frac{2}{3}\, \log_2(\epsilon^{-1})$. Thus, using (1.13), we have that

$$\#\{\mu \in M : |\langle f, \psi_\mu \rangle| > \epsilon\} \leq \sum_{j \geq 0} R(j, \epsilon) \leq C\, \epsilon^{-2/3}\, \log_2(\epsilon^{-1}),$$

and this implies (1.12). ☐

2. Proofs. This section contains the constructions and proofs needed for the theorems in section 1.4.

2.1. Proof of Theorem 1.3. Suppose that a function in $\mathcal{E}^2(A)$ contains a C^2 edge. Following the approach in [7], we suppose that, for $j > j_0$, the scale 2^{-j} is small enough so that over the square $-2^{-j} \leq x_1, x_2 \leq 2^{-j}$ the edge curve may be parametrized as $\binom{E(x_2)}{x_2}$ or $\binom{x_1}{E(x_1)}$. (The case where $j \leq j_0$ is small requires a much simpler analysis and will be discussed in section 2.3.) Without loss of generality, let

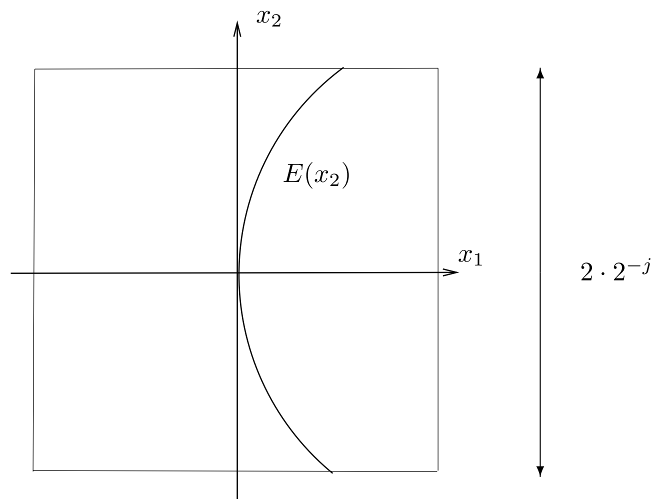


Fig. 2.1. *Representation of an edge fragment.*

us assume that the first parametrization holds. Then an *edge fragment* is a function of the form

$$f(x_1, x_2) = w(2^j x_1, 2^j x_2)\, g(x_1, x_2)\, \chi_{\{x_1 \geq E(x_2)\}},$$

where $g \in C_0^2((0,1)^2)$. For simplicity, let us assume that the edge goes through the origin and, at this point, its tangent is vertical (see Figure 2.1). Then, using the regularity of the edge curve, we have that

(i) $E(0) = 0$, $E'(0) = 0$;

(ii) $\sup_{|x_2| \leq 2^{-j}} |E(x_2)| \leq \frac{1}{2} \sup_{|x_2| \leq 2^{-j}} 2^{-2j} |E''(x_2)|$.

That means that, for $|x_2| \leq 2^{-j}$, the edge curve is almost straight. Observe that any arbitrary edge fragment is obtained by rotating and translating the one above. Since the analysis of the edge fragment that will be presented in the following is not affected by these transformations, there is no loss of generality in considering this case only.

In order to quantify the decay properties of the shearlet coefficients, we first need to analyze the decay of the Fourier transform of the edge fragment along radial lines in the region $\mathcal{D}_C \subset \widehat{\mathbb{R}}^2$, defined in section 1.2. It will be convenient to introduce polar coordinates. Let $\xi = (\xi_1, \xi_2) \in \mathcal{D}_C$. Using polar coordinates, we have

$$\xi_1 = \lambda \cos\theta,\ \xi_2 = \lambda \sin\theta, \quad \text{with } |\theta| \leq \frac{\pi}{4},\ \lambda \in \mathbb{R},\ |\xi_1| \geq \frac{1}{8}.$$

Using this notation, the radial lines have the form $(\lambda\cos\theta, \lambda\sin\theta)$, $\lambda \in \mathbb{R}$, $|\theta| \leq \frac{\pi}{4}$.

For $\xi = (\xi_1, \xi_2) \in \mathcal{D}_C$, $j \geq 0$, $-2^j \leq \ell \leq 2^j$, we introduce the notation

$$\Gamma_{j,\ell}(\xi) = \hat{\psi}_1\big(2^{-2j}\, \xi_1\big)\ \hat{\psi}_2\left(2^j\, \frac{\xi_2}{\xi_1} - \ell\right). \tag{2.1}$$

We have the following claim.

Proposition 2.1. *Let f be an edge fragment and $\Gamma_{j,\ell}$ be given by* (2.1). *Then*

$$\int_{\mathbb{R}^2} |\hat{f}(\xi)|^2\, |\Gamma_{j,\ell}(\xi)|^2\, d\xi \leq C\, 2^{-3j}\, (1 + |\ell|)^{-5}.$$

In order to prove this proposition, we need to recall the following result [7, Thm. 6.1].

THEOREM 2.2. *Let f be an edge fragment and I_j a dyadic interval* $[2^{2j-\alpha}, 2^{2j+\beta}]$ *with* $\alpha \in \{0,1,2,3,4\}$, $\beta \in \{0,1,2\}$. *Then, for all* θ,

$$\int_{|\lambda|\in I_j} |\hat{f}(\lambda\cos\theta, \lambda\sin\theta)|^2 \, d\lambda \le C\, 2^{-3j} \left(1 + 2^j |\sin\theta|\right)^{-5}.$$

Proof of Proposition 2.1. The assumptions on the support of $\hat{\psi}_1$ and $\hat{\psi}_2$ imply that

$$\operatorname{supp} \hat{\psi}_1(2^{-2j}\xi_1) \subset \left\{\xi_1 \in [-2^{2j-1}, -2^{2j-4}] \cup [2^{2j-4}, 2^{2j-1}]\right\} \tag{2.2}$$

and

$$\operatorname{supp} \hat{\psi}_2\left(2^j \frac{\xi_2}{\xi_1} - \ell\right) \subset \left\{(\xi_1, \xi_2) : \left|2^j \frac{\xi_2}{\xi_1} - \ell\right| \le 1\right\}.$$

Since $\tan\theta = \frac{\xi_2}{\xi_1}$, the last expression can be written as

$$\operatorname{supp} \hat{\psi}_2\left(2^j \frac{\xi_2}{\xi_1} - \ell\right) \subset \left\{(\lambda, \theta) :\, 2^{-j}(\ell - 1) \le \tan\theta \le 2^{-j}(\ell + 1)\right\}. \tag{2.3}$$

Since $\lambda^2 = \xi_1^2 + \xi_2^2 = \xi_1^2\,(1 + (\tan\theta)^2)$ and $|\ell| \le 2^j$, then, using (2.2) and (2.3), we have

$$|\lambda| \le 2^{2j-1}\left(1 + 2^{-2j}(1 + |\ell|)^2\right)^{\frac{1}{2}} \le 2^{2j-1}\left(1 + 2^{-2j}(1 + 2^j)^2\right)^{\frac{1}{2}} \le 2^{2j+1}$$

and

$$|\lambda| \ge 2^{2j-4}\left(1 + 2^{-2j}(|\ell| - 1)^2\right)^{\frac{1}{2}} \ge 2^{2j-4}.$$

Thus, the support of $\Gamma_{j,\ell}$ is contained in

$$W_{j,\ell} = \{(\lambda, \theta) :\, 2^{2j-4} \le |\lambda| \le 2^{2j+1},\ \arctan(2^{-j}(\ell - 1)) \le \theta \le \arctan(2^{-j}(\ell + 1))\}.$$

Observe that, in particular, $|\theta| \le \arctan 2$. Since, for $|\theta| \le 2$, we have that[1] $\tan\theta \approx \sin\theta$, it follows from (2.3) that, on $W_{j,\ell}$,

$$2^j\,|\sin\theta| \approx 2^j\,(2^{-j}\,|\ell|) = |\ell|. \tag{2.4}$$

Thus, using (2.4) and Theorem 2.2, we have that

$$\begin{aligned}
\int_{\widehat{\mathbb{R}}^2} |\hat{f}(\xi)|^2\,|\Gamma_{j,\ell}(\xi)|^2\,d\xi &\le C \int_{W_{j,\ell}} |\hat{f}(\lambda\cos\theta, \lambda\sin\theta)|^2\,\lambda\,d\lambda\,d\theta \\
&\le C \int_{\arctan(2^{-j}(\ell-1))}^{\arctan(2^{-j}(\ell+1))} \int_{2^{2j-4}}^{2^{2j+1}} |\hat{f}(\lambda\cos\theta, \lambda\sin\theta)|^2\,|\lambda|\,d\lambda\,d\theta \\
&\le C\,2^{2j+1} \int_{\arctan(2^{-j}(\ell-1))}^{\arctan(2^{-j}(\ell+1))} 2^{-4j}\left(1 + 2^j|\sin\theta|\right)^{-5} d\theta \\
&\le C\,2^{-2j}\,(1 + |\ell|)^{-5}\left(\arctan(2^{-j}(\ell - 1)) - \arctan(2^{-j}(\ell + 1))\right) \\
&= C\,2^{-3j}\,(1 + |\ell|)^{-5}. \qquad \square
\end{aligned}$$

[1] We use the notation $f(x) \approx g(x)$, $x \in D$, to mean that there are constants $C_1, C_2 > 0$ such that $C_1\,g(x) \le f(x) \le C_2\,g(x)$ for all $x \in D$.

The following proposition provides a similar estimate for the partial derivatives of the Fourier transform of the edge fragment.

PROPOSITION 2.3. *Let f be an edge fragment, $\Gamma_{j,\ell}$ be given by* (2.1), *and L be the differential operator*

$$L = \left(I - \left(\frac{2^{2j}}{2\pi\,(1+|\ell|)}\right)^2 \frac{\partial^2}{\partial \xi_1^2}\right) \left(1 - \left(\frac{2^j}{2\pi}\right)^2 \frac{\partial^2}{\partial \xi_2^2}\right).$$

Then

$$\int_{\widehat{\mathbb{R}}^2} \left| L\left(\hat{f}(\xi)\,\Gamma_{j,\ell}(\xi)\right)\right|^2 d\xi \le C\, 2^{-3j}\,(1+|\ell|)^{-5}.$$

In order to prove this proposition, we need to recall the following result [7, Cor. 6.6].

COROLLARY 2.4. *Let f be an edge fragment and I_j a dyadic interval $[2^{2j-\alpha}, 2^{2j+\beta}]$ with $\alpha \in \{0,1,2,3,4\}$, $\beta \in \{0,1,2\}$. Then, for each $m = (m_1, m_2) \in \overline{\mathbb{N}} \times \overline{\mathbb{N}}$ and for each θ,*

$$\begin{aligned}\int_{|\lambda|\in I_j} &\left| \frac{\partial^{m_1}}{\partial \xi_1^{m_1}} \frac{\partial^{m_2}}{\partial \xi_2^{m_2}} \hat{f}(\lambda\cos\theta, \lambda\sin\theta)\right|^2 d\lambda \\ &\le C_m\, 2^{-2j|m|}\left(2^{-(4+2m_1)j}\,(1+2^j|\sin\theta|)^{-5} + 2^{-10j}\right),\end{aligned}$$

where C_m is independent of j and ℓ and $\overline{\mathbb{N}} = \mathbb{N}\cup\{0\}$.

We also need the following lemma, which follows from a direct computation.

LEMMA 2.5. *Let $\Gamma_{j,\ell}$ be given by* (2.1). *Then, for each $m = (m_1, m_2) \in \overline{\mathbb{N}} \times \overline{\mathbb{N}}$, $m_1, m_2 \in \{0,1,2\}$,*

$$\left| \frac{\partial^{m_1}}{\partial \xi_1^{m_1}} \frac{\partial^{m_2}}{\partial \xi_2^{m_2}} \Gamma_{j,\ell}(\xi_1,\xi_2)\right| \le C_m\, 2^{-(2m_1+m_2)j}\,(1+|\ell|)^{m_1},$$

where $|m| = m_1 + m_2$ and C_m is independent of j and ℓ.

We can now prove Proposition 2.3.

Proof of Proposition 2.3. From Corollary 2.4, using (2.4), we have

$$\int_{2^{2j-4}}^{2^{2j+1}} \left| \frac{\partial^2}{\partial \xi_1^2}\hat{f}(\lambda\cos\theta,\lambda\sin\theta)\right|^2 d\lambda \le C\,2^{-4j}\left(2^{-8j}\,(1+|\ell|)^{-5} + 2^{-10j}\right).$$

Thus, using the same idea as in the proof of Proposition 2.1,

$$\begin{aligned}\int_{\widehat{\mathbb{R}}^2} &\left| \left(\frac{\partial^2}{\partial \xi_1^2}\hat{f}(\xi)\right)\,\Gamma_{j,\ell}(\xi)\right|^2 d\xi \\ &\le C \int_{\arctan(2^{-j}(\ell-1))}^{\arctan(2^{-j}(\ell+1))} \int_{2^{2j-4}}^{2^{2j+1}} \left| \frac{\partial^2}{\partial \xi_1^2}\hat{f}(\lambda\cos\theta,\lambda\sin\theta)\right|^2 |\lambda|\,d\lambda\,d\theta \\ &\le C\,2^{-3j}\left(2^{-8j}\,(1+|\ell|)^{-5} + 2^{-10j}\right). \qquad (2.5)\end{aligned}$$

Similarly, using Corollary 2.4 and Lemma 2.5, we have

$$\begin{aligned}
&\int_{\widehat{\mathbb{R}}^2} \left| \left(\frac{\partial}{\partial \xi_1} \hat{f}(\xi) \right) \left(\frac{\partial}{\partial \xi_1} \Gamma_{j,\ell}(\xi) \right) \right|^2 d\xi \\
&\quad \le C\, 2^{-4j} (1+|\ell|)^2 \int_{\arctan(2^{-j}(\ell-1))}^{\arctan(2^{-j}(\ell+1))} \int_{2^{2j-4}}^{2^{2j+1}} \left| \frac{\partial}{\partial \xi_1} \hat{f}(\lambda\cos\theta, \lambda\sin\theta) \right|^2 |\lambda| \, d\lambda \, d\theta \\
&\quad \le C\, 2^{-4j} (1+|\ell|)^2\, 2^{-j} \Big(2^{-6j} (1+|\ell|)^{-5} + 2^{-10j} \Big) \\
&\quad = C\ 2^{-5j} (1+|\ell|)^2 \Big(2^{-6j} (1+|\ell|)^{-5} + 2^{-10j} \Big) \qquad (2.6)
\end{aligned}$$

and

$$\begin{aligned}
&\int_{\widehat{\mathbb{R}}^2} \left| \hat{f}(\xi) \left(\frac{\partial^2}{\partial \xi_1^2} \Gamma_{j,\ell}(\xi) \right) \right|^2 d\xi \\
&\quad \le C\, 2^{-8j} (1+|\ell|)^4 \int_{\arctan(2^{-j}(\ell-1))}^{\arctan(2^{-j}(\ell+1))} \int_{2^{2j-4}}^{2^{2j+1}} \left| \hat{f}(\lambda\cos\theta, \lambda\sin\theta) \right|^2 |\lambda| \, d\lambda \, d\theta \\
&\quad \le C\, 2^{-8j} (1+|\ell|)^4\, 2^{-3j} (1+|\ell|)^{-5} = C\, 2^{-11j} (1+|\ell|)^{-1}. \qquad (2.7)
\end{aligned}$$

Finally, combining (2.5), (2.6), (2.7), and using the fact that $|\ell| \le 2^j$, we have that

$$\int_{\widehat{\mathbb{R}}^2} \left| \left(\frac{2^{2j}}{2\pi(1+|\ell|)} \right)^2 \frac{\partial^2}{\partial \xi_1^2} \Big(\hat{f}(\xi)\, \Gamma_{j,\ell}(\xi) \Big) \right|^2 d\xi \le C\, 2^{-3j} (1+|\ell|)^{-5}. \qquad (2.8)$$

Similarly for the derivatives with respect to ξ_2 we have

$$\int_{\widehat{\mathbb{R}}^2} \left| \left(\frac{\partial^2}{\partial \xi_2^2} \hat{f}(\xi) \right) \Gamma_{j,\ell}(\xi) \right|^2 d\xi \le C\, 2^{-3j} \Big(2^{-4j} (1+|\ell|)^{-5} + 2^{-10j} \Big), \qquad (2.9)$$

$$\int_{\widehat{\mathbb{R}}^2} \left| \left(\frac{\partial}{\partial \xi_2} \hat{f}(\xi) \right) \left(\frac{\partial}{\partial \xi_2} \Gamma_{j,\ell}(\xi) \right) \right|^2 d\xi \le C\, 2^{-3j} \Big(2^{-4j} (1+|\ell|)^{-5} + 2^{-10j} \Big), \qquad (2.10)$$

$$\int_{\widehat{\mathbb{R}}^2} \left| \hat{f}(\xi) \left(\frac{\partial^2}{\partial \xi_2^2} \Gamma_{j,\ell}(\xi) \right) \right|^2 d\xi \le C\, 2^{-7j} (1+|\ell|)^{-5}. \qquad (2.11)$$

Combining (2.9), (2.10), (2.11), and using again the fact that $|\ell| \le 2^j$, we have that

$$\int_{\widehat{\mathbb{R}}^2} \left| \left(\frac{2^j}{2\pi} \right)^2 \frac{\partial^2}{\partial \xi_2^2} \Big(\hat{f}(\xi)\, \Gamma_{j,\ell}(\xi) \Big) \right|^2 d\xi \le C\, 2^{-3j} (1+|\ell|)^{-5}. \qquad (2.12)$$

Similarly, one can show that

$$\int_{\widehat{\mathbb{R}}^2} \left| \frac{2^{3j}}{(1+|\ell|)(2\pi)^2} \frac{\partial^2}{\partial \xi_2^2} \frac{\partial^2}{\partial \xi_1^2} \Big(\hat{f}(\xi)\, \Gamma_{j,\ell}(\xi) \Big) \right|^2 d\xi \le C\, 2^{-3j} (1+|\ell|)^{-5}. \qquad (2.13)$$

The proof is completed using (2.8), (2.12), (2.13), and Lemma 2.5. ∎

We can now prove Theorem 1.3. The following proof adapts some ideas from [7].

Proof of Theorem 1.3. Fix $j \geq 0$ and, for simplicity of notation, let $f = f_Q$. For $\mu \in M_j$, the shearlet coefficient of f is

$$\langle f, \psi_\mu \rangle = \langle f, \psi_{j,\ell,k} \rangle = |\det A|^{-j/2} \int_{\widehat{\mathbb{R}}^2} \hat{f}(\xi)\, \Gamma_{j,\ell}(\xi)\, e^{2\pi i \xi A^{-j} B^{-\ell} k}\, d\xi,$$

where $\Gamma_{j,\ell}(\xi)$ is given by (2.1) and A, B are given by (1.4). Observe that

$$\begin{aligned} 2\pi i \xi A^{-j} B^{-\ell} k &= 2\pi i \begin{pmatrix} \xi_1 & \xi_2 \end{pmatrix} \begin{pmatrix} 2^{-2j} & 0 \\ 0 & 2^{-j} \end{pmatrix} \begin{pmatrix} 1 & -\ell \\ 0 & 1 \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} \\ &= 2\pi i \left((k_1 - k_2 \ell) 2^{-2j} \xi_1 + k_2 2^{-j} \xi_2 \right). \end{aligned} \tag{2.14}$$

Using (2.14), a direct computation shows that

$$\frac{\partial^2}{\partial \xi_1^2} \left(2\pi \xi A^{-j} B^{-\ell} k \right) = -(2\pi)^2 2^{-4j} (k_1 - k_2 \ell)^2 = \begin{cases} -(2\pi)^2\, \ell^2\, 2^{-4j} (\frac{k_1}{\ell} - k_2)^2 & \text{if } \ell \neq 0, \\ -(2\pi)^2\, 2^{-4j} k_1^2 & \text{if } \ell = 0, \end{cases}$$

$$\frac{\partial^2}{\partial \xi_2^2} \left(2\pi \xi A^{-j} B^{-\ell} k \right) = -(2\pi)^2\, 2^{-2j} k_2^2. \tag{2.15}$$

By the equivalent definition of weak ℓ^p norm, the theorem is proved, provided we show that

$$\#\{\mu \in M_j : |\langle f, \psi_\mu \rangle| > \epsilon\} \leq C\, 2^{-j}\, \epsilon^{-\frac{2}{3}}. \tag{2.16}$$

Let L be the second order differential operator defined in Proposition 2.3. Using (2.14) and (2.15), it follows that

$$L\big(e^{2\pi i \xi A^{-j} B^{-\ell} k}\big) = \begin{cases} \left(1 + \left(\frac{\ell}{(1+|\ell|)} \right)^2 \left(\frac{k_1}{\ell} - k_2 \right)^2 \right) (1 + k_2^2)\, e^{2\pi i \xi A^{-j} B^{-\ell} k} & \text{if } \ell \neq 0, \\ (1 + k_1^2)(1 + k_2^2)\, e^{2\pi i \xi A^{-j} B^{-\ell} k} & \text{if } \ell = 0. \end{cases} \tag{2.17}$$

Integration by parts gives

$$\langle f, \psi_\mu \rangle = |\det A|^{-j/2} \int_{\widehat{\mathbb{R}}^2} L \left(\hat{f}(\xi)\, \Gamma_{j,\ell}(\xi) \right)\, L^{-1} \left(e^{2\pi i \xi A^{-j} B^{-\ell} k} \right)\, d\xi.$$

Let us consider first the case $\ell \neq 0$. In this case, from (2.17) we have that

$$L^{-1}\big(e^{2\pi i \xi A^{-j} B^{-\ell} k}\big) = G(k, \ell)^{-1}\, e^{2\pi i \xi A^{-j} B^{-\ell} k}, \tag{2.18}$$

where $G(k, \ell) = \big(1 + \big(\frac{\ell}{(1+|\ell|)}\big)^2 \big(\frac{k_1}{\ell} - k_2\big)^2\big)(1 + k_2^2)$. Thus

$$\langle f, \psi_\mu \rangle = |\det A|^{-j/2}\, G(k, \ell)^{-1} \int_{\widehat{\mathbb{R}}^2} L \left(\hat{f}(\xi)\, \Gamma_{j,\ell}(\xi) \right)\, e^{2\pi i \xi A^{-j} B^{-\ell} k}\, d\xi,$$

or, equivalently,

$$G(k, \ell)\, \langle f, \psi_\mu \rangle = |\det A|^{-j/2} \int_{\widehat{\mathbb{R}}^2} L \left(\hat{f}(\xi)\, \Gamma_{j,\ell}(\xi) \right)\, e^{2\pi i \xi A^{-j} B^{-\ell} k}\, d\xi.$$

Let $K = (K_1, K_2) \in \mathbb{Z}^2$, and define $R_K = \{k = (k_1, k_2) \in \mathbb{Z}^2 : \frac{k_1}{\ell} \in [K_1, K_1 + 1],\ k_2 = K_2\}$. Since, for j, ℓ fixed, the set $\{|\det A|^{-j/2}\, e^{2\pi i \xi A^{-j} B^{-\ell} k} : k \in \mathbb{Z}^2\}$ is an orthonormal basis for the L^2 functions on $[-\frac{1}{2}, \frac{1}{2}]A^j B^\ell$, and the function $\Gamma_{j,\ell}(\xi)$ is supported on this set, then

$$\sum_{k \in R_K} |\langle G(k,\ell)\, f, \psi_\mu \rangle|^2 \le \int_{\widehat{\mathbb{R}}^2} \left| L\left(\hat{f}(\xi)\, \Gamma_{j,\ell}(\xi)\right)\right|^2 \, d\xi.$$

From the definition of R_K it follows that

$$\sum_{k \in R_K} |\langle f, \psi_\mu \rangle|^2 \le C\, \left(1 + (K_1 - K_2)^2\right)^{-2} (1 + K_2)^{-2} \int_{\widehat{\mathbb{R}}^2} \left| L\left(\hat{f}(\xi)\, \Gamma_{j,\ell}(\xi)\right)\right|^2 \, d\xi.$$

By Proposition 2.3,

$$\sum_{k \in R_K} |\langle f, \psi_\mu \rangle|^2 \le C\, L_K^{-2}\, 2^{-3j} (1 + |\ell|)^{-5}, \tag{2.19}$$

where $L_K = (1 + (K_1 - K_2)^2)(1 + K_2^2)$. For j, ℓ fixed, let $N_{j,\ell,K}(\epsilon) = \#\{k \in R_K : |\psi_{j,\ell,k}| > \epsilon\}$. Then it is clear that $N_{j,\ell,K}(\epsilon) \le C\,(|\ell| + 1)$, and (2.19) implies that

$$N_{j,\ell,K}(\epsilon) \le C\, L_K^{-2}\, 2^{-3j}\, \epsilon^{-2} (1 + |\ell|)^{-5}.$$

Thus

$$N_{j,\ell,K}(\epsilon) \le C\, \min\left(|\ell| + 1, L_K^{-2}\, 2^{-3j}\, \epsilon^{-2} (1 + |\ell|)^{-5}\right). \tag{2.20}$$

Using (2.20), we will now show that

$$\sum_{\ell = -2^j}^{2^j} N_{j,\ell,K}(\epsilon) \le C\, L_K^{-\frac{2}{3}}\, 2^{-j}\, \epsilon^{-\frac{2}{3}}. \tag{2.21}$$

In fact, let ℓ^* be defined by $(\ell^* + 1) = L_K^{-2}\, 2^{-3j}\, \epsilon^{-2} (1 + \ell^*)^{-5}$. That is, $\ell^* + 1 = L_K^{-1/3}\, 2^{-j/2}\, \epsilon^{-1/3}$. Then

$$\begin{aligned}
\sum_{\ell = -2^j}^{2^j} N_{j,\ell,K}(\epsilon) &\le \sum_{|\ell| \le (\ell^* + 1)} N_{j,\ell,K}(\epsilon) + \sum_{|\ell| > (\ell^* + 1)} N_{j,\ell,K}(\epsilon) \\
&\le \sum_{|\ell| \le (\ell^* + 1)} (|\ell| + 1) + \sum_{|\ell| > (\ell^* + 1)} L_K^{-2}\, 2^{-3j}\, \epsilon^{-2} (1 + |\ell|)^{-5} \\
&\le (\ell^* + 1)^2 + C\, L_K^{-2}\, 2^{-3j}\, \epsilon^{-2} (1 + \ell^*)^{-4} \le C\, (\ell^* + 1)^2,
\end{aligned}$$

which gives (2.21).

Since $\sum_{K \in \mathbb{Z}^2} L_K^{-2/3} < \infty$, using (2.21) we then have that

$$\#\{\mu \in M_j : |\langle f, \psi_\mu \rangle| > \epsilon\} \le \sum_{K \in \mathbb{Z}^2} \sum_{\ell = -2^j}^{2^j} N_{j,\ell,K}(\epsilon) \le C\, 2^{-j}\, \epsilon^{-\frac{2}{3}} \sum_{K \in \mathbb{Z}^2} L_K^{-\frac{2}{3}} \le C\, 2^{-j}\, \epsilon^{-\frac{2}{3}},$$

and, thus (2.16) holds.

The case $\ell = 0$ is similar. Indeed, in this case

$$L^{-1}\left(e^{2\pi i \xi A^{-j} B^{-\ell} k}\right) = (1 + k_1^2)^{-1} (1 + k_2^2)^{-1}\, e^{2\pi i \xi A^{-j} B^{-\ell} k},$$

and we can proceed as in the case $\ell \neq 0$, with $L_K = (1 + K_1^2)\,(1 + K_2^2)$. It is clear that also in this case $\sum_{K \in \mathbb{Z}^2} L_K^{-2/3} < \infty$. This completes the proof of the theorem. □

2.2. Proof of Theorem 1.4. In order to prove Theorem 1.4, the following lemmata will be useful.

LEMMA 2.6. *Let $f = g\, w_Q$, where $g \in \mathcal{E}^2(A)$ and $Q \in \mathcal{Q}_j^1$. Then*

$$\int_{W_{j,\ell}} |\hat{f}(\xi)|^2 \, d\xi \le C\, 2^{-10j}. \tag{2.22}$$

Proof. The proof follows [7, Lem. 8.1] and is reported here for completeness.

The function f belongs to $C_0^2(\mathbb{R}^2)$, and its second partial derivative with respect to x_1 is

$$\frac{\partial^2 f}{\partial x_1^2} = \frac{\partial^2 g}{\partial x_1^2}\, w_Q + 2\, \frac{\partial\, g}{\partial x_1}\, \frac{\partial\, w_Q}{\partial x_1} + f\, \frac{\partial^2 w_Q}{\partial x_1^2} = h_1 + h_2 + h_3.$$

Using the fact that w_Q is supported in a square of sidelength $2 \cdot 2^{-j}$, we have

$$\int_{\widehat{\mathbb{R}}^2} |\hat{h}_1(\xi)|^2 \, d\xi = \int_{\mathbb{R}^2} |h_1(x)|^2 \; dx \le C\, 2^{-2j}.$$

Next, observe that $\|\frac{\partial}{\partial x_1} h_2\|_\infty \le C\, 2^{2j}$. Using again the condition on the support of w_Q, it follows that

$$\int_{\widehat{\mathbb{R}}^2} |2\pi\xi_1\, \hat{h}_2(\xi)|^2 \, d\xi = \int_{\mathbb{R}^2} \left| \frac{\partial}{\partial x_1} h_2(x) \right|^2 \; dx \le C\, 2^{2j},$$

and thus, for $\xi \in W_{j,\ell}$ (hence $\xi_1 \approx 2^{2j}$),

$$\int_{W_{j,\ell}} |\hat{h}_2(\xi)|^2 \, d\xi \le C\, 2^{-2j}.$$

Finally, observing that $\|\frac{\partial^2}{\partial x_1^2} h_3\|_\infty \le C\, 2^{4j}$, a computation similar to the one above shows that

$$\int_{W_{j,\ell}} |\hat{h}_3(\xi)|^2 \, d\xi \le C\, 2^{-2j}.$$

Since $-(2\pi)^2\, \xi_1^2\, \hat{f}(\xi) = \hat{h}_1(\xi) + \hat{h}_2(\xi) + \hat{h}_3(\xi)$, it follows from the estimates above that

$$\int_{W_{j,\ell}} |\hat{f}(\xi)|^2 \, d\xi \le C\, 2^{-10j}.$$

This completes the proof. □

LEMMA 2.7. *Let $m = (m_1, m_2) \in \overline{\mathbb{N}} \times \overline{\mathbb{N}}$, $\xi = (\xi_1, \xi_2) \in \widehat{\mathbb{R}}^2$, and $\Gamma_{j,\ell}$ be given by (2.1). Then*

$$\sum_{\ell=-2^j}^{2^j} \left| \frac{\partial^{m_1}}{\partial \xi_1^{m_1}}\, \frac{\partial^{m_2}}{\partial \xi_2^{m_2}}\, \Gamma_{j,\ell}(\xi) \right|^2 \le C_m\, 2^{-2|m|j},$$

where C_m is independent of j and ξ and $|m| = m_1 + m_2$.

Proof. Observe that $W_{j,\ell} \cap W_{j,\ell+\ell'} = \emptyset$ whenever $|\ell'| \ge 3$. Since $|\ell| \le 2^j$, the lemma then follows from Lemma 2.5. □

LEMMA 2.8. *Let $f = g\, w_Q$, where $g \in \mathcal{E}^2(A)$ and $Q \in \mathcal{Q}_j^1$. Define*

$$T = \left(I - \frac{2^j}{(2\pi)^2}\,\Delta\right), \tag{2.23}$$

where $\Delta = \frac{\partial^2}{\partial \xi_1^2} + \frac{\partial^2}{\partial \xi_2^2}$. Then

$$\int_{\widehat{\mathbb{R}}^2} \sum_{\ell=-2^j}^{2^j} \left| T^2 \left(\hat f\, \Gamma_{j,\ell}\right)(\xi)\right|^2 d\xi \le C\, 2^{-10j}.$$

Proof. Observe that, for $N \in \overline{\mathbb{N}}$,

$$\Delta^N \left(\hat f\, \Gamma_{j,\ell}\right) = \sum_{|\alpha|+|\beta|=2N} C_{\alpha,\beta} \left(\frac{\partial^{\alpha_1}}{\partial \xi_1^{\alpha_1}}\, \frac{\partial^{\alpha_2}}{\partial \xi_2^{\alpha_2}}\, \hat f\right) \left(\frac{\partial^{\beta_1}}{\partial \xi_1^{\beta_1}}\, \frac{\partial^{\beta_2}}{\partial \xi_2^{\beta_2}}\, \Gamma_{j,\ell}\right).$$

Then, using Lemma 2.7, we have that

$$\begin{aligned}\int_{\widehat{\mathbb{R}}^2} \sum_{\ell=-2^j}^{2^j} &\left|\frac{\partial^{\alpha_1}}{\partial \xi_1^{\alpha_1}}\, \frac{\partial^{\alpha_2}}{\partial \xi_2^{\alpha_2}}\, \hat f(\xi)\right|^2 \left|\frac{\partial^{\beta_1}}{\partial \xi_1^{\beta_1}}\, \frac{\partial^{\beta_2}}{\partial \xi_2^{\beta_2}}\, \Gamma_{j,\ell}(\xi)\right|^2 d\xi \\ &\le C_\beta\, 2^{-2|\beta| j} \int_{W_{j,\ell}} \left|\frac{\partial^{\alpha_1}}{\partial \xi_1^{\alpha_1}}\, \frac{\partial^{\alpha_2}}{\partial \xi_2^{\alpha_2}}\, \hat f(\xi)\right|^2 d\xi.\end{aligned}$$

Recall that $f(x)$ is of the form $g(x)\, w(2^j x)$. It follows that $x^\alpha f(x) = 2^{-j|\alpha|}\, g(x)\, w_\alpha(2^j x)$, where $w_\alpha(x) = x^\alpha w(x)$. By Lemma 2.6, $g(x)\, w_\alpha(2^j x)$ obeys the estimate (2.22). Thus, observing that $\frac{\partial^{\alpha_1}}{\partial \xi_1^{\alpha_1}}\, \frac{\partial^{\alpha_2}}{\partial \xi_2^{\alpha_2}}\, \hat f(\xi)$ is the Fourier transform of $(-2\pi i x)^\alpha f(x)$, we have that

$$\int_{W_{j,\ell}} \left|\frac{\partial^{\alpha_1}}{\partial \xi_1^{\alpha_1}}\, \frac{\partial^{\alpha_2}}{\partial \xi_2^{\alpha_2}}\, \hat f(\xi)\right|^2 d\xi \le C_\alpha\, 2^{-2j|\alpha|}\, 2^{-10j}.$$

Combining the estimates above, we have that, for each α, β with $|\alpha| + |\beta| = 2N$,

$$\int_{\widehat{\mathbb{R}}^2} \sum_{\ell=-2^j}^{2^j} \left|\frac{\partial^{\alpha_1}}{\partial \xi_1^{\alpha_1}}\, \frac{\partial^{\alpha_2}}{\partial \xi_2^{\alpha_2}}\, \hat f(\xi)\right|^2 \left|\frac{\partial^{\beta_1}}{\partial \xi_1^{\beta_1}}\, \frac{\partial^{\beta_2}}{\partial \xi_2^{\beta_2}}\, \Gamma_{j,\ell}(\xi)\right|^2 d\xi \le C_{\alpha,\beta}\, 2^{-10j}\, 2^{-4jN}. \tag{2.24}$$

Since $T^2 = 1 - 2\,\frac{2^j}{(2\pi)^2}\,\Delta + \frac{2^{2j}}{(2\pi)^4}\,\Delta^2$, the lemma now follows from (2.24) and Lemma 2.7. $\square$

We can now prove Theorem 1.4.

Proof of Theorem 1.4. Using (2.15), for T given by (2.23), we have that

$$T\left(e^{2\pi i \xi A^{-j} B^{-\ell} k}\right) = \left(1 + 2^{-2j}(k_1 - k_2\,\ell)^2 + k_2^2\right)\, e^{2\pi i \xi A^{-j} B^{-\ell} k}. \tag{2.25}$$

Fix $j \ge 0$ and let $f = f_Q$. Then, using integration by parts as in the proof of Theorem 1.3, from (2.25) it follows that

$$\langle f, \psi_\mu\rangle = |\det A|^{-j} \left(1 + 2^{-2j}(k_1 - k_2\,\ell)^2 + k_2^2\right)^{-2} \int_{\widehat{\mathbb{R}}^2} T^2 \left(\hat f\, \Gamma_{j,\ell}\right)(\xi)\, e^{2\pi i \xi A^{-j} B^{-\ell} k}\, d\xi.$$

Let $K = (K_1, K_2) \in \mathbb{Z}^2$ and R_K be the set $\{(k_1, k_2) \in \mathbb{Z}^2 : k_2 = K_2, 2^{-j}(k_1 - K_2\ell) \in [K_1, K_1 + 1]\}$. Observing that, for each K, there are only $1 + 2^j$ choices for k_1 in R_K, it follows that the number of terms in R_K is bounded by $1 + 2^j$. Thus, arguing again as in the proof of Theorem 1.3, we have that

$$\sum_{k \in R_K} |\langle f, \psi_\mu \rangle|^2 \le C \, \left(1 + K_1^2 + K_2^2\right)^{-4} \int_{\widehat{\mathbb{R}}^2} \left| T^2 \left(\hat{f} \, \Gamma_{j,\ell} \right)(\xi) \right|^2 \, d\xi.$$

From this inequality, using Lemma 2.8, we have that

$$\begin{aligned} \sum_{\ell=-2^j}^{2^j} \sum_{k \in R_K} |\langle f, \psi_\mu \rangle|^2 &\le C \, (1 + K^2)^{-4} \int_{\widehat{\mathbb{R}}^2} \sum_{\ell=-2^j}^{2^j} \left| T^2 \left(\hat{f} \, \Gamma_{j,\ell} \right)(\xi) \right|^2 \, d\xi \\ &\le C \, (1 + K^2)^{-4} \, 2^{-10j}. \end{aligned} \tag{2.26}$$

For any $N \in \mathbb{N}$, provided $\frac{1}{2} < p < 2$, the Hölder inequality yields

$$\sum_{m=1}^{N} |a_m|^p \le \left(\sum_{m=1}^{N} |a_m|^2 \right)^{\frac{p}{2}} \, N^{\left(1 - \frac{p}{2}\right)}. \tag{2.27}$$

Since the cardinality of R_K is bounded by $1 + 2^j$, it follows from (2.26) and (2.27) that, for $\frac{1}{2} < p < 2$,

$$\sum_{\ell=-2^j}^{2^j} \sum_{k \in R_K} |\langle f, \psi_\mu \rangle|^p \le C \, \left(2^{2j}\right)^{(1 - \frac{p}{2})} \, (1 + K^2)^{-2p} \, 2^{-5pj}.$$

Thus, since $p > \frac{1}{2}$,

$$\sum_{\mu \in M_j} |\langle f, \psi_\mu \rangle|^p \le C \, 2^{\left(2j(1 - \frac{p}{2}) - 5pj\right)} \sum_{K \in \mathbb{Z}^2} (1 + K^2)^{-2p} \le C \, 2^{(2-3p)j},$$

and, in particular,

$$\|\langle f, \psi_\mu \rangle\|_{\ell^{2/3}} \le C \, 2^{-3j}. \qquad \square$$

2.3. Coarse scale analysis. In section 2.1, we assumed that the scale parameter j was large enough. The situation where j is small can be treated in a much simpler way. In fact, if f_Q is an edge fragment, then a trivial estimate shows that

$$\|f_Q\|_2 = \left(\int_{\operatorname{supp} w_Q} |f_Q(x)|^2 \, dx \right)^{1/2} \le C \, |\operatorname{supp} w_Q| = C \, 2^{-j}.$$

It follows that $\|\langle f_Q, \psi_\mu \rangle\|_{\ell^2} \le C \, 2^{-j}$, and thus, by the Hölder inequality,

$$\|\langle f_Q, \psi_\mu \rangle\|_{\ell^{2/3}} \le C \, 2^{j}.$$

This satisfies Theorem 1.3 for j small.

2.4. Additional remarks.

- In order to define the collection of shearlets, in section 1.2 we have constructed a function $\hat{\psi} \in C_0^\infty$. This property allows us to obtain a collection of elements that are well localized. Observe, however, that we need only $\hat{\psi} \in C_0^2$ in order to prove all the results presented in this paper.
- In this paper, we have considered the representation of functions containing a discontinuity along a C^2 curve. More generally, we can consider the situation where a function f contains many edge curves of this type, exhibiting finitely many junctions or corners between them. In this setting, the discontinuity curve is not globally C^2 but only piecewise C^2. The results reported in this paper, namely Theorems 1.1 and 1.2, extend to this setting as well. We refer to [7] for a similar discussion in the case of curvelets.
- The assumption we made about the regularity of the discontinuity curve plays a critical role in our construction. If the discontinuity curve is in C^α, with $\alpha > 2$, then our argument still works and we can still prove Theorem 1.2. This result, however, is not (essentially) optimal as in the case $\alpha = 2$. On the other hand, if the discontinuity curve is in C^α, with $\alpha < 2$, then the estimate given by Theorem 1.2 does not hold, and the estimate could be worse in general. We refer to [24] for additional observations about this fact, and for an alternative approach, based on an adaptive construction, to the sparse representation of functions with edges.
- There are natural ways of extending the shearlets to dimensions larger than 2. We refer to [18] for a discussion of these extensions, as well as the extensions of the shear transformations to the general multidimensional setting. For example, in dimension 3, let $A = \left(\begin{smallmatrix} 4 & 0 \\ 0 & 2\,I_2 \end{smallmatrix}\right)$; define the shear matrices $\{S_k = \left(\begin{smallmatrix} 1 & k \\ 0 & I_2 \end{smallmatrix}\right) : k \in \mathbb{Z}^2\}$, where I_2 is the 2×2 identity matrix, $0 = \left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$; and, for $\xi = (\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3$, define ψ by

$$\hat{\psi}(\xi) = \hat{\psi}_1(\xi_1)\, \hat{\psi}_2\left(\frac{\xi_2}{\xi_1}\right)\, \hat{\psi}_2\left(\frac{\xi_3}{\xi_1}\right),$$

where ψ_1 and ψ_2 are given as in the 2-D case. Then, similarly to their 2-D counterpart, one can construct a Parseval frame of well-localized 3-D shearlets

$$\{\psi_{j,\ell,k} = |\det A|^{-j/2}\, \psi(S_\ell\, A^{-j} x - k) : \, j \in \mathbb{Z}, \ell \in \mathbb{Z}^2, k \in \mathbb{Z}^2\},$$

with frequency support on a parallelepiped of approximate size $2^{2j} \times 2^j \times 2^j$, at various scales j, with orientations controlled by the 2-D index ℓ and spatial location k. Then, using a heuristic argument, one can argue that these systems provide sparse representations for 3-D functions f that are smooth away from "nice" surface discontinuities of finite area. In fact, thanks to their frequency support and their localization properties, the elements $\psi_{j,\ell,k}$, at scale j, are essentially supported on a parallelepiped of size $2^{-2j} \times 2^{-j} \times 2^{-j}$, with location controlled by k and orientation controlled by ℓ. Thus, there are at most $O(2^{2j})$ significant shearlet coefficients $\mathcal{SH}_{j,\ell,k}(f) = \langle f, \psi_{j,\ell,k} \rangle$, and they are bounded by $C\, 2^{-2j}$. This implies that the Nth largest 3-D shearlet coefficient $|\mathcal{SH}_N(f)|$ is bounded by $O(N^{-1})$, and thus, if f is approximated by taking the N largest coefficients in the 3-D shearlets expansion, the L^2-error would approximately obey

$$\|f - f_N^S\|_{L^2}^2 \le \sum_{\ell > N} |\mathcal{SH}_\ell(f)|^2 \le C\, N^{-1},$$

up to lower order factors. A rigorous proof of this fact will be presented elsewhere.

Appendix. Construction of $\boldsymbol{\psi_1, \psi_2}$. In this section we show how to construct examples of functions ψ_1, ψ_2 satisfying the properties described in section 1.2.

In order to construct ψ_1, let $h(t)$ be an even C_0^∞ function, with support in $(-\frac{1}{6}, \frac{1}{6})$, satisfying $\int_{\mathbb{R}} h(t)\,dt = \frac{\pi}{2}$, and define $\theta(\omega) = \int_{-\infty}^{\omega} h(t)\,dt$. Then one can construct a smooth bell function as

$$b(\omega) = \begin{cases} \sin\left(\theta\left(|\omega| - \frac{1}{2}\right)\right) & \text{if } \frac{1}{3} \le |\omega| \le \frac{2}{3}, \\ \sin\left(\frac{\pi}{2} - \theta\left(\frac{|\omega|}{2} - \frac{1}{2}\right)\right) & \text{if } \frac{2}{3} < |\omega| \le \frac{4}{3}, \\ 0 & \text{otherwise.} \end{cases}$$

It follows from our assumptions (cf. [20, sect. 1.4]) that

$$\sum_{j=-1}^{\infty} b^2(2^{-j}\omega) = 1 \quad \text{for } |\omega| \ge \frac{1}{3}.$$

Now letting $u^2(\omega) = b^2(2\omega) + b^2(\omega)$, it follows that

$$\sum_{j\ge 0}^{\infty} u^2(2^{-2j}\omega) = \sum_{j=-1}^{\infty} b^2(2^{-j}\omega) = 1 \quad \text{for } |\omega| \ge \frac{1}{3}.$$

Finally, let ψ_1 be defined by $\hat{\psi}_1(\omega) = u(\frac{8}{3}\omega)$. Then $\operatorname{supp}\hat{\psi}_1 \subset [-\frac{1}{2}, -\frac{1}{16}] \cup [\frac{1}{16}, \frac{1}{2}]$, and (1.6) is satisfied. This construction is illustrated in Figure A.1(a).

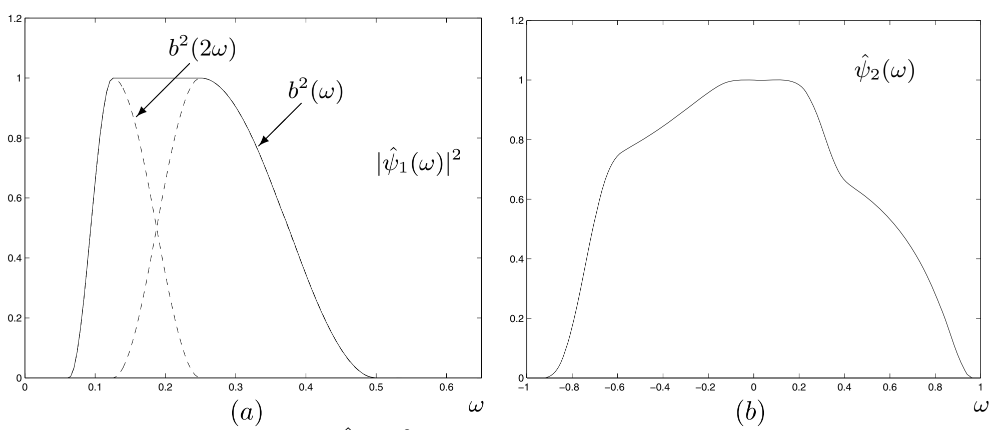


FIG. A.1. (a) *The function* $|\hat{\psi}_1(\omega)|^2$ *(solid line), for* $\omega > 0$*; the negative side is symmetrical. This function is obtained, after rescaling, from the sum of the window functions* $b^2(\omega) + b^2(2\omega)$ *(dashed lines).* (b) *The function* $\hat{\psi}_2(\omega)$.

For the construction of ψ_2, we start by considering a smooth bump function $f_1 \in C_0^\infty(-1,\ 1)$ such that $0 \le f_1 \le 1$ on $(-1, 1)$ and $f_1 = 1$ on $[-\frac{1}{2}, \frac{1}{2}]$ (cf. [21, sect. 1.4]). Next, let $f_2(t) = \sqrt{1 - \exp(1/t)}$. Then (in the left limit sense) $f_2(0) = 1$, $f_2^{(k)}(0) = 0$ for $k \ge 1$ and $0 < f_2 < 1$ on $(-1, 0)$. Define $f(t) = f_1(t) f_2(t)$ for $t \in [-1, 0]$. It is then easy to see that $f^{(k)}(-1) = 0$ for $k \ge 0$, and $f(0) = 1$, $f^{(k)}(0) = 0$ for $k \ge 1$.

Since $g(t) = \exp\left(\frac{1}{2(t-1)}\right)$ for $t \in (\frac{1}{2}, 1)$, it follows that $\lim_{t\to 1-} g^{(k)}(t) = 0$ for $k \geq 0$. Finally, we define

$$\hat{\psi}_2(\omega) = \begin{cases} f(\omega) & \text{if } \omega \in [-1, 0), \\ g(\omega) & \text{if } \omega \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

Then $\hat{\psi}_2 \in C_0^\infty(\mathbb{R})$, with $\operatorname{supp} \hat{\psi}_2 \subset [-1, 1]$, and

$$\hat{\psi_2}^2(\omega) + \hat{\psi_2}^2(\omega - 1) = 1, \quad \omega \in [0, 1].$$

The last equality implies (1.7). The function $\hat{\psi}_2$ is illustrated in Figure A.1(b).

Acknowledgments. The authors thank E. Candès, P. Gremaud, G. Kutyniok, W. Lim, G. Weiss, and E. Wilson for useful discussions.

REFERENCES

[1] J. P. Antoine, R. Murenzi, and P. Vandergheynst, *Directional wavelets revisited: Cauchy wavelets and symmetry detection in patterns*, Appl. Comput. Harmon. Anal., 6 (1999), pp. 314–345.

[2] R. H. Bamberger and M. J. T. Smith, *A filter bank for directional decomposition of images: Theory and design*, IEEE Trans. Signal Process., 40 (1992), pp. 882–893.

[3] C. M. Brislawn, *Fingerprints go digital*, Notices Amer. Math. Soc., 42 (1995), pp. 1278–1283.

[4] C. M. Brislawn and M. D. Quirk, *Image compression with the JPEG-*2000 *standard*, in Encyclopedia of Optical Engineering, R. G. Driggers, ed., Marcel Dekker, New York, 2003, pp. 780–785.

[5] E. J. Candès and D. L. Donoho, *Ridgelets: A key to higher-dimensional intermittency?*, Phil. Trans. Royal Soc. London A, 357 (1999), pp. 2495–2509.

[6] E. J. Candès and D. L. Donoho, *Curvelets—A surprising effective nonadaptive representation for objects with edges*, in Curves and Surfaces, C. Rabut, A. Cohen, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 2000.

[7] E. J. Candès and D. L. Donoho, *New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities*, Comm. Pure Appl. Math., 56 (2004), pp. 216–266.

[8] R. R. Coifman and F. G. Meyer, *Brushlets: A tool for directional image analysis and image compression*, Appl. Comput. Harmonic Anal., 5 (1997), pp. 147–187.

[9] S. Dahlke, G. Kutyniok, P. Maass, C. Sagiv, and H.-G. Stark, *The Uncertainty Principle Associated with the Continuous Shearlet Transform*, preprint, 2006.

[10] R. A. DeVore, *Nonlinear approximation*, in Acta Numerica, A. Iserles, ed., Cambridge University Press, Cambridge, UK, 1998, pp. 51–150.

[11] M. N. Do and M. Vetterli, *The contourlet transform: An efficient directional multiresolution image representation*, IEEE Trans. Image Process., 14 (2005), pp. 2091–2106.

[12] D. L. Donoho, *Sparse components of images and optimal atomic decomposition*, Constr. Approx., 17 (2001), pp. 353–382.

[13] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, *Data compression and harmonic analysis*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2435–2476.

[14] G. Easley, D. Labate, and W.-Q. Lim, *Sparse Directional Image Representations Using the Discrete Shearlet Transform*, preprint, 2006.

[15] K. Guo, G. Kutyniok, and D. Labate, *Sparse multidimensional representations using anisotropic dilation and shear operators,* in Wavelets and Splines, G. Chen and M. Lai, eds., Nashboro Press, Nashville, TN, 2006, pp. 189–201.

[16] K. Guo, W.-Q. Lim, D. Labate, G. Weiss, and E. Wilson, *Wavelets with composite dilations*, Electron. Res. Announc. Amer. Math. Soc., 10 (2004), pp. 78–87.

[17] K. Guo, W.-Q. Lim, D. Labate, G. Weiss, and E. Wilson, *The theory of wavelets with composite dilations*, in Harmonic Analysis and Applications, C. Heil, ed., Birkhäuser Boston, Cambridge, MA, 2006, pp. 231–249.

[18] K. Guo, W.-Q. Lim, D. Labate, G. Weiss, and E. Wilson, *Wavelets with composite dilations and their MRA properties*, Appl. Computat. Harmon. Anal., 20 (2006), pp. 231–249.

[19] *JPEG* 2000 *Image Coding System, Part* 1, Ser. ISO/IEC Intl. Standard 15444-1, ITU-T Rec. T.800, International Organization for Standardization, 2000.
[20] E. Hernández and G. Weiss, *A First Course on Wavelets*, CRC Press, Boca Raton, FL, 1996.
[21] L. Hörmander, *The Analysis of Linear Partial Differential Operators.* I. *Distribution Theory and Fourier Analysis,* Springer-Verlag, Berlin, 2003.
[22] N. Kingsbury, *Complex wavelets for shift invariant analysis and filtering of signals*, Appl. Computat. Harmon. Anal., 10 (2001), pp. 234–253.
[23] D. Labate, W.-Q. Lim, G. Kutyniok, and G. Weiss, *Sparse multidimensional representation using shearlets*, in Proceedings of Wavelets XI (San Diego, CA, 2005), SPIE Proc. 5914, SPIE, Bellingham, WA, 2005, pp. 254–262.
[24] E. Le Pennec and S. Mallat, *Sparse geometric image representations with bandelets*, IEEE Trans. Image Process., 14 (2005), pp. 423–438.
[25] W.-Q. Lim, *Wavelets with Composite Dilations*, Ph.D. Thesis, Department of Mathematics, Washington University in St. Louis, St. Louis, MO, 2006.
[26] Y. Lu and M. N. Do, *Multidimensional directional filter banks and surfacelets*, IEEE Trans. Image Process., (2007), to appear.
[27] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, CA, 1998.
[28] D. D. Po and M. N. Do, *Directional multiscale modeling of images using the contourlet transform*, IEEE Trans. Image Process., 15 (2006), pp. 1610–1620.
[29] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. MATH. ANAL.
Vol. 39, No. 1, pp. 319–344

TRAVELING FRONTS OF PYRAMIDAL SHAPES IN THE ALLEN–CAHN EQUATIONS*

MASAHARU TANIGUCHI†

Abstract. This paper studies pyramidal traveling fronts in the Allen–Cahn equation or in the Nagumo equation. For the nonlinearity we are concerned mainly with the bistable reaction term with unbalanced energy density. Two-dimensional V-form waves and cylindrically symmetric waves in higher dimensions have been recently studied. Our aim in this paper is to construct truly three-dimensional traveling waves. For a pyramid that satisfies a condition, we construct a traveling front for which the contour line has a pyramidal shape. We also construct generalized pyramidal fronts and traveling waves of a hybrid type between pyramidal waves and planar V-form waves. We use the comparison principles and construct traveling fronts between supersolutions and subsolutions.

Key words. pyramidal traveling wave, Allen–Cahn equation, bistable

AMS subject classification. 35K57

DOI. 10.1137/060661788

1. Introduction.

In this paper we consider the following equation:

$$\frac{\partial u}{\partial t} = \Delta u + f(u) \quad \text{in } \mathbb{R}^3, \quad t > 0,$$
$$u\big|_{t=0} = u_0 \quad \text{in } \mathbb{R}^3.$$

Here a given function u_0 is bounded and of class C^1. The Laplacian Δ stands for $\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$. If the nonlinearity f is cubic, it is called the Allen–Cahn equation or the Nagumo equation. We study general nonlinear terms of a bistable type including cubic ones.

In the one-dimensional space, let $\Phi(x - kt)$ be a traveling wave that connects two stable equilibrium states ± 1 with speed k. By putting $\mu = x - kt$, Φ satisfies

$$(1) \qquad \begin{gathered} -\Phi''(\mu) - k\Phi'(\mu) - f(\Phi(\mu)) = 0 \quad -\infty < \mu < \infty, \\ \Phi(-\infty) = 1, \quad \Phi(\infty) = -1. \end{gathered}$$

To fix the phase we set $\Phi(0) = 0$. Such one-dimensional traveling waves have been studied in many works. See Fife and McLeod [5], Aronson and Weinberger [1], Kanel' [10, 11], Chen [2], and Terman [18], for instance. We state equations for the unbalanced nonlinearity and the balanced one.

The unbalanced case is as follows. The following are the assumptions on f in this paper:

(A1) f is of class $C^1[-1, 1]$, with $f(\pm 1) = 0$ and $f'(\pm 1) < 0$.
(A2) $\int_{-1}^{1} f > 0$ holds true.
(A3) There exists $\Phi(\mu)$ that satisfies (1) for some $k \in \mathbb{R}$.

The assumption (A1) implies that f is bistable and that (A2) means that it is unbalanced. Note that (A2) implies $k > 0$. Under (A1), $(k, \Phi(\mu))$ is uniquely

*Received by the editors June 2, 2006; accepted for publication (in revised form) January 4, 2007; published electronically May 29, 2007. This research was partially supported by the Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (C), 18540208, 2006.
http://www.siam.org/journals/sima/39-1/66178.html

†Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, O-okayama 2-12-1-W8-38, Tokyo 152-8552, Japan (masaharu.taniguchi@is.titech.ac.jp).

determined if it exists. For the proof of this uniqueness, see [5] or [2]. We show simple examples for f here.

Example 1. If f satisfies $f'(\beta) > 0$ and

$$\begin{aligned} f(s) > 0 \quad &\text{for } \beta < s < 1, \\ f(s) < 0 \quad &\text{for } -1 < s < \beta, \end{aligned}$$

with some $\beta \in (-1, 1)$ in addition to (A1), then it is well known that (A3) is valid. See [5] or [2]. Especially, $f(u) = -(u+1)(u+a)(u-1)$ has a one-dimensional traveling wave $\Phi(\mu) = -\tanh(\mu/\sqrt{2})$ with speed $k = \sqrt{2}a$ for every $a \in [0, 1)$. This traveling wave is sometimes called the Huxley solution. See Nagumo, Yoshizawa, and Arimoto [14].

Example 2 (Fife and McLeod [5, Theorem 2.7]). Assume f satisfies (A1) and (A2). For $-1 < \lambda < 1$ assume that there exists (c_L, Φ_L) to

$$\begin{gathered} -\Phi_L''(\mu) - k\Phi_L'(\mu) - f(\Phi_L(\mu)) = 0 \quad -\infty < \mu < \infty, \\ \Phi_L(-\infty) = 1, \quad \Phi_L(\infty) = \lambda, \end{gathered}$$

and there exists (c_R, Φ_R) to

$$\begin{gathered} -\Phi_R''(\mu) - k\Phi_R'(\mu) - f(\Phi_R(\mu)) = 0 \quad -\infty < \mu < \infty, \\ \Phi_R(-\infty) = \lambda, \quad \Phi_R(\infty) = -1. \end{gathered}$$

If $c_L > c_R$, then (A3) holds true. If $c_L \le c_R$, there exists no solution to (1).

Example 3. For G with (B1) and (B2) below, we define

$$f(u) = -G'(u) + k\sqrt{2G(u)}$$

for $k > 0$. Then $\Phi_0(\mu)$ given by (3) is a solution to (1) for $k > 0$. If k is small enough, $f(u)$ satisfies (A1), (A2), and (A3).

We note that (A1) and (A2) do not always imply (A3), because we can construct such f with $c_L \le c_R$ in Example 2. If it exists, it is always monotone in μ as in Lemma 1. See [5, Lemma 2.1] for the proof of the monotony of one-dimensional fronts. We use this monotony and the comparison principles in this paper.

The balanced case is as follows:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \Delta u - G'(u) \quad \text{in } \mathbb{R}^3, \quad t > 0, \\ u|_{t=0} &= u_0 \quad \text{in } \mathbb{R}^3. \end{aligned} \tag{2}$$

The assumptions on G are as follows:

(B1) G is of class $C^2[-1, 1]$, with $G'(\pm 1) = 0$, $G''(\pm 1) > 0$.

(B2) $G(1) = 0$ and $G(s) > 0$ for $-1 < s < 1$.

Under (B1) and (B2), (2) has a standing wave solution $\Phi_0(x)$ to

$$\begin{gathered} -\Phi_0''(\mu) + G'(\Phi_0(x)) = 0 \quad -\infty < \mu < \infty, \\ \Phi_0(-\infty) = 1, \quad \Phi_0(\infty) = -1. \end{gathered}$$

Φ_0 is given by

$$x = -\int_0^{\Phi_0} \frac{dv}{\sqrt{2G(v)}}. \tag{3}$$

The condition $G(1) = 0$ means that a potential density term G has minimizers with an equal depth. If G takes a negative value or zero in $(-1, 1)$, there exists no standing wave. Thus (B2) is the condition for the existence of a standing wave solution Φ_0. A typical balanced nonlinearity term is $-G'(u) = u - u^3$, with $G(u) = (1/4)(1 - u^2)^2$.

First we study traveling waves for the unbalanced nonlinearity. We adopt the moving coordinate of speed c toward the z-axis without loss of generality. We put $s = z - ct$ and $u(x, y, z, t) = w(x, y, s, t)$. We denote $w(x, y, s, t)$ by $w(x, y, z, t)$ for simplicity. Then we obtain

$$\begin{aligned} w_t - w_{xx} - w_{yy} - w_{zz} - cw_z - f(w) = 0 \quad &\text{in } \mathbb{R}^3, \quad t > 0, \\ w|_{t=0} = u_0 \quad &\text{in } \mathbb{R}^3. \end{aligned} \tag{4}$$

Here w_t stands for $\partial w/\partial t$ and so on. We write the solution as $w(x, y, z, t; u_0)$. If v is a traveling wave with speed c, it satisfies

$$\mathcal{L}[v] \overset{\text{def}}{=} -v_{xx} - v_{yy} - v_{zz} - cv_z - f(v) = 0 \quad \text{in } \mathbb{R}^3. \tag{5}$$

We assume

$$c > k$$

throughout this paper. There exist many traveling waves in this situation, because k is the speed of a planar traveling wave, and the curvature effect often accelerates the speed.

In the two-dimensional plane there exists the following V-form wave.

THEOREM 1 (see [15]). *Under the assumptions* $c > k$, (A1), (A2), *and* (A3), *there exists* $v_*(x, y)$, *with*

$$-(v_*)_{xx} - (v_*)_{yy} - c(v_*)_y - f(v_*) = 0 \qquad \text{for } (x, y) \in \mathbb{R}^2,$$
$$\lim_{R\to\infty} \sup_{x^2+y^2>R^2} \left| v_*(x, y) - \Phi\left(\frac{k}{c}\left(y - \frac{\sqrt{c^2 - k^2}}{k}|x|\right)\right)\right| = 0.$$

Under these two equalities $v_*(x, y)$ *is uniquely determined.*

See also Hamel, Monneau, and Roquejoffre [8, 9] for V-form waves in the Allen–Cahn equation. Recently Haragus and Scheel [13] studied V-form waves in reaction-diffusion systems including the Allen–Cahn equation by using the bifurcation theory when the angle $\arctan(\sqrt{c^2 - k^2}/k)$ is small enough. Such a bifurcation technique is applicable to the cases where a one-dimensional traveling front loses its monotony.

For three- or higher-dimensional cases with cylindrical symmetry, Hamel, Monneau, and Roquejoffre [8, 9] studied conical traveling waves for unbalanced bistable nonlinearity. The proof is based on the results for bounded cylinders, and a passage to the limit gives the existence of a conical front in the whole domain.

Now we study three-dimensional traveling waves, and our aim is to search truly three-dimensional traveling waves that have pyramidal structures and are neither cylindrically symmetric nor reducible to two-dimensional traveling waves. For this purpose, we construct pyramidal traveling waves to (5). We apply the method of Ninomiya and Taniguchi [15, 16]. A supersolution for the V-form wave is constructed in [15] as follows. In the moving coordinate we put an almost flat planar front above the shape "V." This curve is almost flat, and then the real solution goes downwards with speed $c - k > 0$, since we are using a moving coordinate. This means that an almost flat stationary planar front is a supersolution. This method is based on the monotony of a one-dimensional traveling front and the comparison methods. The application of this method is restricted to equations for which the comparison principle holds true. In this paper, we put a mollified pyramid above a pyramid in $\mathbb{R}^3$ and construct a supersolution carefully, because a pyramidal wave is everywhere apart from a pyramid near the edges.

Let $n \geq 3$ be a given integer. We put

$$\tau \stackrel{\text{def}}{=} \frac{\sqrt{c^2 - k^2}}{k} > 0. \tag{6}$$

Assume $(A_j, B_j) \in \mathbb{R}^2$ satisfies

$$A_j^2 + B_j^2 = 1 \qquad \text{for all } j = 1, \ldots, n \tag{7}$$

and

$$\begin{gathered} A_j B_{j+1} - A_{j+1} B_j > 0, \qquad 1 \leq j \leq n-1, \\ A_n B_1 - A_1 B_n > 0. \end{gathered} \tag{8}$$

We assume $(A_{j_1}, B_{j_1}) \neq (A_{j_2}, B_{j_2})$ if $j_1 \neq j_2$. Now $(-\tau A_j, -\tau B_j, 1)$ is the normal vector of a surface $\{z = \tau(A_j x + B_j y)\}$. We put

$$\begin{aligned} h_j(x,y) &\stackrel{\text{def}}{=} \tau \left(A_j x + B_j y\right), \\ h(x,y) &\stackrel{\text{def}}{=} \max_{1 \leq j \leq n} h_j(x,y) = \tau \max_{1 \leq j \leq n} \left(A_j x + B_j y\right). \end{aligned} \tag{9}$$

Then $z = h(x,y)$ represents a pyramid in $\mathbb{R}^3$. We set

$$\Omega_j = \{(x,y) \,|\, h(x,y) = h_j(x,y)\}$$

and obtain

$$\mathbb{R}^2 = \cup_{j=1}^n \Omega_j.$$

We locate $\Omega_1, \Omega_2, \ldots, \Omega_n$ counterclockwise as in Figure 1. To ensure this location we assumed (8). We set

$$E \stackrel{\text{def}}{=} \cup_{j=1}^n \partial\Omega_j \subset \mathbb{R}^2.$$

Now the lateral surfaces of a pyramid are given by

$$S_j = \{(x,y,z) \in \mathbb{R}^3 \,|\, z = h_j(x,y), \quad (x,y) \in \Omega_j\}$$

for $j = 1, \ldots, n$. We put

$$\Gamma_j \stackrel{\text{def}}{=} \begin{cases} S_j \cap S_{j+1} & \text{if } 1 \leq j \leq n-1, \\ S_n \cap S_1 & \text{if } j = n. \end{cases}$$

Then Γ_j represents an edge of a pyramid. Also

$$\Gamma \stackrel{\text{def}}{=} \cup_{j=1}^n \Gamma_j$$

represents the set of all edges.

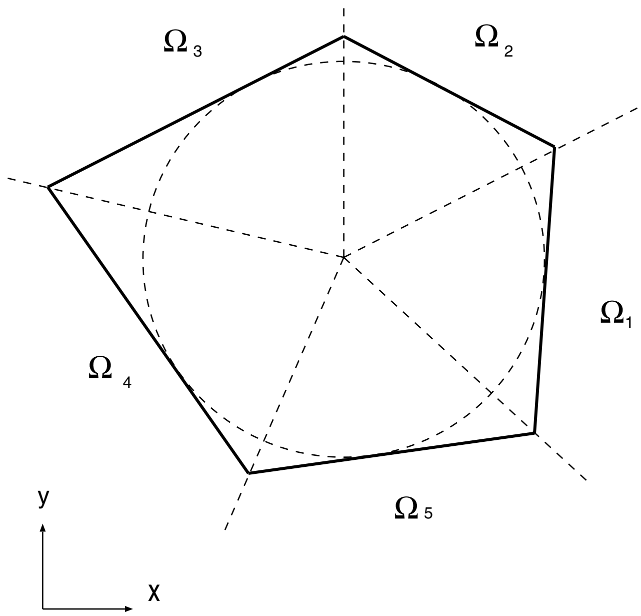


FIG. 1. *The decomposition of the x-y plane by Ω_j for $n = 5$.*

For every (A_j, B_j) with (7), (5) has a solution $\Phi\left((k/c)(z - h_j(x,y))\right)$, which is called a planar wave. Now we have

$$\Phi\left(\frac{k}{c}(z - h(x,y))\right) = \max_{1\le j\le n} \Phi\left(\frac{k}{c}(z - h_j(x,y))\right) = \max_{1\le j\le n} \Phi\left(\frac{k}{c}(z - a_j x - b_j y)\right).$$

This becomes a subsolution to (5). We define

$$D(\gamma) \stackrel{\text{def}}{=} \{(x,y,z) \in \mathbb{R}^3 \mid \operatorname{dist}((x,y,z), \mathit{\Gamma}) > \gamma\} \tag{10}$$

for $\gamma > 0$. We will construct a supersolution that is larger than this subsolution and obtain a traveling wave between them.

The following theorem is the main assertion in this paper.

THEOREM 2. *Let $c > k$, and let $h(x,y)$ be given by* (9). *Under the assumptions* (A1), (A2), *and* (A3), *there exists a solution $V(x,y,z)$ to* (5) *with*

$$\Phi\left(\frac{k}{c}(z - h(x,y))\right) < V(x,y,z) < 1 \qquad \text{in } \mathbb{R}^3$$

and

$$\lim_{\gamma\to+\infty} \sup_{(x,y,z)\in D(\gamma)} \left| V(x,y,z) - \Phi\left(\frac{k}{c}(z - h(x,y))\right)\right| = 0, \tag{11}$$

$$V_z(x,y,z) < 0 \qquad \text{for all } (x,y,z) \in \mathbb{R}^3. \tag{12}$$

We state the proof of this theorem in section 3. A domain $D(\gamma)$ is a complement of a neighborhood of the edges. The property (11) implies that the geometric shape of V can be approximated by a combination of n planar waves except on a neighborhood of the edges (see Figure 2). We conjecture that the geometric shape of V

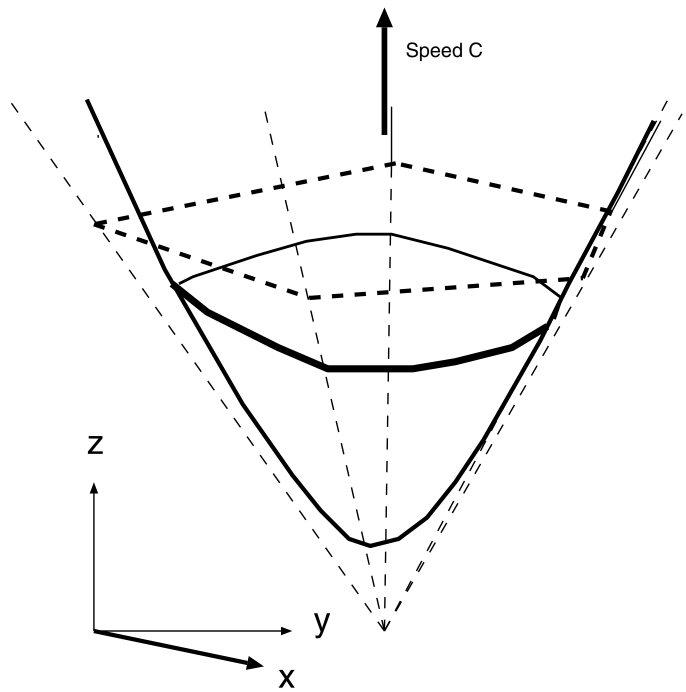


FIG. 2. *A pyramidal traveling wave.*

can be approximated by two-dimensional V-form waves on the edges and that V is a combination of planar waves and two-dimensional V-form waves. The uniqueness and the stability of V is yet to be proved.

Section 4 is devoted to applications of Theorem 2. A two-dimensional V-form wave in Theorem 1 immediately gives a three-dimensional wave $v_*(x,z)$. We call this wave a planar V-form wave. It is natural to search for a combination of a pyramidal wave and a planar V-form wave. In section 4 we study a traveling wave of a hybrid type between pyramidal waves and planar V-form waves as a special case of Theorem 2.

We studied pyramids whose lateral surfaces contain the origin in $\mathbb{R}^3$ in Theorem 2. We consider the case where the surfaces do not have a common point in section 5. Even in that case a combination of n planar waves gives three-dimensional traveling waves, and we construct generalized pyramidal traveling waves when the zero level sets of planar waves have no common point.

We study traveling waves for the balanced nonlinearity in section 6. For any given $c > 0$ we study

$$\mathcal{L}_0[v] \stackrel{\text{def}}{=} -v_{xx} - v_{yy} - v_{zz} - cv_z + G'(v) = 0 \quad \text{in } \mathbb{R}^3. \tag{13}$$

We call $-G'(u)$ in Example 3 a balanced nonlinearity. Cylindrically symmetric traveling waves for balanced nonlinearity have been studied in Chen et al [3] for two or higher dimensions. The limit of traveling waves for unbalanced nonlinearity terms when the difference of energy density goes to zero gives a traveling wave in (13). Pyramidal traveling waves for unbalanced nonlinear terms converge to traveling waves for a balanced nonlinearity term as the difference of the energy density goes to zero. Up to now the profile of the limit traveling wave is unknown and is yet to be studied.

The characterization and classification of all traveling waves for unbalanced and balanced nonlinearities will give interesting problems and are left for further studies.

2. Pyramids and mollified pyramids. In this section we make preparations. We state known results for one-dimensional traveling waves and construct mollified pyramids

LEMMA 1 (Fife and McLeod [5]). *Under the assumptions* (A1) *and* (A3), $\Phi(\mu)$ *as in* (1) *satisfies*

$$\Phi'(\mu) < 0 \qquad \text{for all } \mu \in \mathbb{R},$$
$$\max\{|\Phi'(\mu)|, |\Phi''(\mu)|, |\mu\Phi'(\mu)|\} \leq K_0 \exp(-\kappa_0|\mu|).$$

Here K_0, κ_0 are positive constants.

There exists a positive constant δ_* $(0 < \delta_* < 1/4)$, with

$$-f'(s) > \kappa_1 \qquad \text{if } |s+1| < 2\delta_* \text{ or } |s-1| < 2\delta_*,$$

where

$$\kappa_1 \stackrel{\text{def}}{=} \frac{1}{2}\min\{-f'(-1), -f'(1)\} > 0.$$

We construct mollified pyramids. Let $\widetilde{\rho}(r) \in C^\infty[0,\infty)$ be a function with the following properties:

$$\widetilde{\rho}(r) > 0, \quad \widetilde{\rho}_r(r) \leq 0 \qquad \text{for } r \geq 0,$$
$$\widetilde{\rho}(r) \equiv 1 \qquad \text{if } 0 \leq r \leq \tfrac{1}{2},$$
$$\widetilde{\rho}(r) = e^{-r} \qquad \text{if } r > 0 \text{ is large enough},$$
$$2\pi \int_0^\infty r\widetilde{\rho}(r)dr = 1.$$

Then $\rho(x,y) \stackrel{\text{def}}{=} \widetilde{\rho}(\sqrt{x^2+y^2})$ belongs to $C^\infty(\mathbb{R}^2)$ and satisfies $\int_{\mathbb{R}^2} \rho = 1$. For a pyramid $z = h(x,y)$ we define a mollified pyramid $z = \varphi(x,y)$ as $\varphi(x,y) \stackrel{\text{def}}{=} \rho * h$, which means

$$\varphi(x,y) = \int_{\mathbb{R}^2} \rho(x-x', y-y')h(x', y')dx'dy' = \int_{\mathbb{R}^2} \rho(x', y')h(x-x', y-y')dx'dy'. \tag{14}$$

We set $(a_j, b_j) \stackrel{\text{def}}{=} \tau(A_j, B_j)$. Then $(a_j, b_j) \in \mathbb{R}^2$ satisfies

$$\frac{c}{\sqrt{1+a_j^2+b_j^2}} = k \qquad \text{for all } j = 1, \ldots, n. \tag{15}$$

We put

$$S(x,y) \stackrel{\text{def}}{=} \frac{c}{\sqrt{1+\varphi_x(x,y)^2+\varphi_y(x,y)^2}} - k. \tag{16}$$

Then we have the following lemma.

LEMMA 2. *Let φ and S be as in* (14) *and* (16), *respectively. Then one has*

$$\sup_{(x,y)\in\mathbb{R}^2} |D_x^{i_1} D_y^{i_2} \varphi(x,y)| < +\infty$$

for all integers $i_1 \geq 0$, $i_2 \geq 0$, *and*

$$\begin{aligned} h(x,y) < \varphi(x,y) \leq h(x,y) + 2\pi\tau \int_0^\infty r^2 \widetilde{\rho}(r)\, dr \\ |(\nabla\varphi)(x,y)| < \tau, \quad 0 < S(x,y) < c \end{aligned} \tag{17}$$

for all $(x,y) \in \mathbb{R}^2$.

Proof. Now ρ satisfies $|D_x^{i_1} D_y^{i_2} \rho(x,y)| \leq (\text{const}) e^{-\sqrt{x^2+y^2}}$ for large $\sqrt{x^2+y^2} > 0$. We get the first estimate from $D_x^{i_1} D_y^{i_2}(\rho * g_j) = (D_x^{i_1} D_y^{i_2} \rho) * g_j$. Note that $\rho * h_j = h_j$. Using $\rho > 0$, $h_j(x,y) \leq h(x,y)$, and $h_j(x,y) \not\equiv h(x,y)$, we have a strict inequality $h_j(x,y) < \varphi(x,y)$. Thus we get $h(x,y) = \max_{1 \leq j \leq n}(h_j(x,y)) < \varphi(x,y)$. Now

$$|h_j(x',y') - h_j(x,y)| \leq \tau\sqrt{(x'-x)^2 + (y'-y)^2}$$

gives

$$|h(x',y') - h(x,y)| \leq \tau\sqrt{(x'-x)^2 + (y'-y)^2}.$$

Thus we obtain

$$\varphi - h \leq \int_{\mathbb{R}^2} |h(x-x', y-y') - h(x,y)|\, \rho(x',y') dx'dy' \leq \tau \int_{\mathbb{R}^2} \sqrt{x^2+y^2}\rho(x,y) dxdy$$

and prove the first inequality. We have

$$(\nabla\varphi)(x) = \int_{\mathbb{R}^2} \rho(x',y')(\nabla h)(x-x', y-y')\, dx'dy'.$$

Here ∇h is a constant vector in each Ω_j, and at least two of these vectors are linearly independent. Thus we get a strict inequality

$$|(\nabla\varphi)(x)| < \int_{\mathbb{R}^2} \rho(x',y')|(\nabla h)(x-x', y-y')|\, dx'dy'.$$

The right-hand side equals

$$\int_{\mathbb{R}^2} \sqrt{a_j^2 + b_j^2}\, \rho(x',y')\, dx'dy' = \tau.$$

Clearly $S < c$ is valid, and $S > 0$ follows from $|\nabla\varphi| < \tau$. This completes the proof. □

The following proposition plays a key role in this paper.

Proposition 1. *For every integer* $i_1 \geq 0$, $i_2 \geq 0$, *with* $2 \leq i_1 + i_2 \leq 3$,

$$\sup_{(x,y)\in\mathbb{R}^2} \frac{|(D_x^{i_1} D_y^{i_2}\varphi)(x,y)|}{S(x,y)} < +\infty$$

holds true.

The proof of this proposition is given at the end of this section.

We study the difference of a mollified pyramid and the original pyramid, that is, $\varphi(x,y) - h(x,y)$. We put

$$\widetilde{\varphi}_j(x,y) \stackrel{\text{def}}{=} \varphi(x,y) - h_j(x,y) = \varphi(x,y) - a_j x - b_j y. \tag{18}$$

Then we have $\varphi(x,y) - h(x,y) = \widetilde{\varphi}_j(x,y)$ in Ω_j. It suffices to study $\widetilde{\varphi}_j(x,y)$ in each Ω_j for studying $\varphi(x,y) - h(x,y)$ in $\mathbb{R}^2$. To do this we study here the simplest case. For

$$q(x,y) \stackrel{\text{def}}{=} \max\{x,0\} = \begin{cases} -x & x < 0, \\ 0 & x \geq 0, \end{cases}$$

we define

$$\begin{aligned} P(x) &\stackrel{\text{def}}{=} \int_{\mathbb{R}^2} \rho(x',y') q(x-x', y-y')\, dx'dy' \\ &= -\int_{-\infty}^{\infty} \left(\int_x^{\infty} \rho(x',y')(x-x')\, dx' \right) dy' > 0. \end{aligned} \tag{19}$$

This $P(x)$ is a mollified function for $q(x,y)$. We use it to estimate $\varphi(x,y) - h(x,y)$ because it stands for the influence of a lateral surface when we construct a mollified pyramid from the original pyramid. Then we have

$$\begin{aligned} P'(x) &= -\int_{-\infty}^{\infty} \left(\int_x^{\infty} \rho(x',y)\, dx' \right) dy < 0, \\ P''(x) &= \int_{-\infty}^{\infty} \rho(x,y)\, dy = \int_{-\infty}^{\infty} \widetilde{\rho}(\sqrt{x^2+y^2})\, dy > 0, \\ P^{(3)}(x) &= \int_{-\infty}^{\infty} \frac{x}{\sqrt{x^2+y^2}} \widetilde{\rho}_r(\sqrt{x^2+y^2})\, dy \leq 0. \end{aligned}$$

Especially we have

$$P''(x) = \int_{-\infty}^{\infty} e^{-\sqrt{x^2+y^2}}\, dy, \quad P^{(3)}(x) = -\int_{-\infty}^{\infty} \frac{x}{\sqrt{x^2+y^2}} e^{-\sqrt{x^2+y^2}}\, dy,$$

if $x > 0$ is large enough. Now we have the following lemma.

LEMMA 3. *Let $P(x)$ be as in* (19). *Then it satisfies*

$$\lim_{x\to\infty} \frac{P(x)}{\sqrt{2\pi x} e^{-x}} = 1$$

and

$$\lim_{x\to\infty} \frac{|P^{(i)}(x)|}{P(x)} = 1, \quad 0 < \inf_{x\geq 1} \frac{|P^{(i)}(x)|}{P(x)} \leq \sup_{x\geq 1} \frac{|P^{(i)}(x)|}{P(x)} < +\infty$$

for all i with $1 \leq i \leq 3$.

Proof. For $x > 0$ we use $y = \sqrt{s^2 + 2sx}$ and obtain

$$2\int_0^{\infty} e^{-\sqrt{x^2+y^2}}\, dy = 2e^{-x} \int_0^{\infty} e^{-s} \frac{s+x}{\sqrt{s^2+2sx}}\, ds = \sqrt{2x} e^{-x} Q(x).$$

Here

$$Q(x) \stackrel{\text{def}}{=} \int_0^{\infty} \frac{1}{\sqrt{s}} e^{-s} \left(1 + \frac{s}{x}\right) \left(1 + \frac{s}{2x}\right)^{-\frac{1}{2}}\, ds.$$

By Lebesgue's convergence theorem we have

$$\lim_{x\to\infty} Q(x) = \int_0^{\infty} \frac{1}{\sqrt{s}} e^{-s}\, ds = \sqrt{\pi}.$$

Thus we have

$$P''(x) = \sqrt{2\pi x} e^{-x}(1+o(1)) \qquad \text{as } x \to \infty. \tag{20}$$

Similarly we get

$$\begin{aligned} &2\int_0^\infty \frac{x}{\sqrt{x^2+y^2}} e^{-\sqrt{x^2+y^2}}\, dy = \sqrt{2x} e^{-x} \int_0^\infty e^{-s} \sqrt{\frac{2x}{s^2+2sx}}\, ds \\ &= \sqrt{2x} e^{-x} \int_0^\infty e^{-s} \frac{1}{\sqrt{s}} \left(1+\frac{s}{2x}\right)^{-\frac{1}{2}} ds = \sqrt{2\pi x} e^{-x} \left(1+o(1)\right) \qquad \text{as } x \to \infty. \end{aligned}$$

Thus we obtain

$$\lim_{x\to\infty} \frac{-P^{(3)}(x)}{P''(x)} = 1.$$

Now the Cauchy mean value theorem gives

$$\frac{P''(x)}{P'(x)} = \frac{P^{(3)}(x')}{P''(x')}$$

for some $x' > x$. This yields

$$\lim_{x\to\infty} \frac{P''(x)}{-P'(x)} = 1.$$

Similarly we have

$$\lim_{x\to\infty} \frac{-P'(x)}{P(x)} = \lim_{x\to\infty} \frac{P''(x)}{-P'(x)} = 1.$$

Thus we obtain

$$\lim_{x\to\infty} \frac{-P'(x)}{P(x)} = \lim_{x\to\infty} \frac{P''(x)}{-P'(x)} = \lim_{x\to\infty} \frac{-P^{(3)}(x)}{P''(x)} = 1.$$

This completes the proof. □

Now we come back to study

$$\widetilde{\varphi}_j(x,y) = \varphi(x,y) - h_j(x,y) = \left(\rho * (h-h_j)\right)(x,y)$$

in Ω_j. Hereafter we assume $(x,y) \in \Omega_j$. We write

$$\boldsymbol{a}_j = (a_j, b_j) \qquad (1 \le j \le n).$$

Then we get

$$0 < \tau^2 - |\nabla\varphi|^2 = -2\boldsymbol{a}_j \cdot \nabla\widetilde{\varphi}_j - |\nabla\widetilde{\varphi}_j|^2. \tag{21}$$

We have

$$h(x,y) - h_j(x,y) = \begin{cases} (a_{j+1}-a_j)x + (b_{j+1}-b_j)y & \text{in } \Omega_{j+1}, \\ (a_{j-1}-a_j)x + (b_{j-1}-b_j)y & \text{in } \Omega_{j-1}. \end{cases}$$

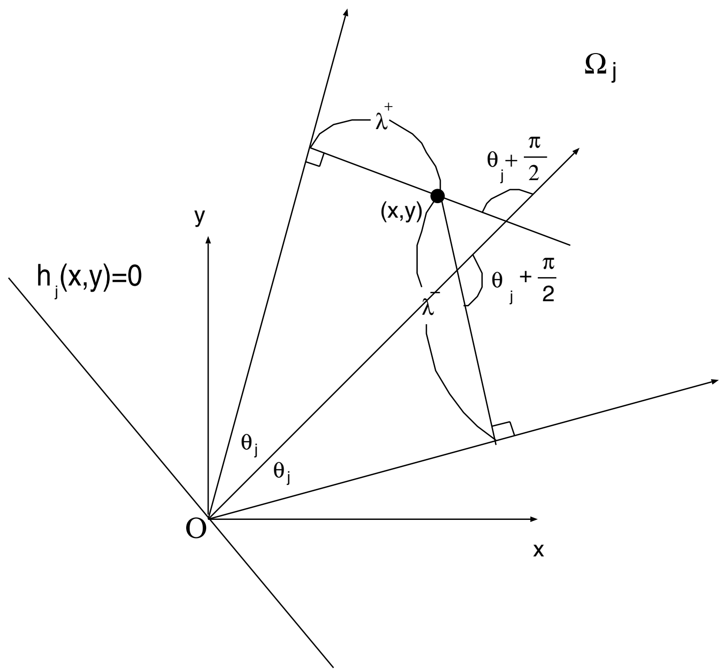


FIG. 3. *The definition of* $\lambda^{\pm}$.

Now

$$m_j^+ \stackrel{\text{def}}{=} \sqrt{(a_{j+1}-a_j)^2+(b_{j+1}-b_j)^2}, \quad m_j^- \stackrel{\text{def}}{=} \sqrt{(a_{j-1}-a_j)^2+(b_{j-1}-b_j)^2} \tag{22}$$

give the gradients of the adjacent surfaces S_{j+1} and S_{j-1} from a surface S_j, respectively. Let the angle of Ω_j be denoted by $2\theta_j$, with $\theta_j \in (0, \pi/2)$ for $j = 1, \ldots, n$ as in Figure 3. For $(x, y) \in \Omega_j$, let λ^+ and λ^- be the lengths of the perpendiculars onto $\partial\Omega_j$. We have

$$\lambda^+ = \frac{(a_j - a_{j+1})x + (b_j - b_{j+1})y}{m_j^+}, \quad \lambda^- = \frac{(a_j - a_{j-1})x + (b_j - b_{j-1})y}{m_j^-}. \tag{23}$$

We study $\widetilde{\varphi}_j(x, y)$ and its derivatives in Ω_j when $\sqrt{x^2+y^2}$ is large enough. The number of the nearest latent surfaces for $(x, y) \in \Omega_j$ is at most two. This fact suggests that $\widetilde{\varphi}_j(x, y)$ can be approximated by $m^+P(\lambda^+) + m^-P(\lambda^-)$ in Ω_j if $\sqrt{x^2+y^2} \to \infty$ up to the derivatives. We have

$$\widetilde{\varphi}_j = \rho * (\max\{h_{j+1} - h_j, 0\}) + \rho * (\max\{h_{j-1} - h_j, 0\}) + \rho * g_j,$$

where

$$g_j \stackrel{\text{def}}{=} h - h_j - \max\{h_{j+1} - h_j, 0\} - \max\{h_{j-1} - h_j, 0\}.$$

Using P, we write the first and the second terms as

$$(\rho * (\max\{h_{j+1} - h_j, 0\}))(x, y) = m_j^+ P(\lambda^+),$$
$$(\rho * (\max\{h_{j-1} - h_j, 0\}))(x, y) = m_j^- P(\lambda^-),$$

respectively. We estimate the third term. We have

$$g_j \equiv 0 \quad \text{on } \Omega_{j-1} \cup \Omega_j \cup \Omega_{j+1}$$

and

$$h_j(x,y) \geq 0 \qquad \text{for all } (x,y) \in \Omega_j.$$

The distance between (x,y) and a line $\{(x,y) \mid h_j(x,y)=0\}$ is $(1/\tau)h_j(x,y)$. The gradients of the planes h_j $(1 \leq j \leq n)$ are at most τ. We put $\Lambda^+ = \mathrm{dist}((x,y), \Omega_{j+2})$ and $\Lambda^- = \mathrm{dist}((x,y), \Omega_{j-2})$. We have $0 < \sin\theta_j < 1$ and $\min\{\lambda^+, \lambda^-\} \leq (h_j(x,y)/\tau)\sin\theta_j$. There exists $\gamma_0 > 1$ such that we have

$$\gamma_0 \min\{\lambda^+, \lambda^-\} \leq \min\left\{\Lambda^+, \Lambda^-, \frac{1}{\tau}h_j(x,y)\right\}.$$

The following lemma is useful to estimate $\rho * g_j$ and $\varphi(x,y) - h(x,y)$.

Lemma 4. *For every j $(1 \leq j \leq n)$, one has*

$$\varphi(x,y) - h(x,y) = m_j^+ P(\lambda^+) + m_j^- P(\lambda^-) + (\rho * g_j)(x,y) \qquad \text{for all } (x,y) \in \Omega_j,$$

where $m_j^{\pm}$, $\lambda^{\pm}$ are given by (22) *and* (23), *respectively. For all nonnegative integers i_1, i_2, with $0 \leq i_1 + i_2 \leq 3$, one has*

$$\left|D_x^{i_1} D_y^{i_2} (\rho * g_j)(x,y)\right| \leq K \left(\gamma_0 \min\{\lambda^+, \lambda^-\}\right)^{\frac{1}{2}} \exp\left(-\gamma_0 \min\{\lambda^+, \lambda^-\}\right)$$

for $(x,y) \in \Omega_j$ and $x^2 + y^2 \geq 1$. Here $K > 0$ and $\gamma_0 > 1$ are constants independent of j, i_1, and i_2. In particular one has

$$\lim_{\lambda\to\infty} \sup\left\{S(x,y) \mid (x,y) \in \mathbb{R}^2, \mathrm{dist}((x,y), E) \geq \lambda\right\} = 0,$$
$$\lim_{\lambda\to\infty} \sup\left\{\varphi(x,y) - h(x,y) \mid (x,y) \in \mathbb{R}^2, \mathrm{dist}((x,y), E) \geq \lambda\right\} = 0.$$

Proof. We already obtained the first equality. We decompose g_j as

$$g_j = g_j \chi_{\{h_{j+2}-h_{j+1}>0\}} + g_j \chi_{\{h_{j-2}-h_{j-1}>0\}} + g_j \chi_{\{h_{j+2}-h_{j+1}\leq 0, h_{j-2}-h_{j-1}\leq 0\}},$$

where $\chi_{\{h_{j+2}-h_{j+1}>0\}}$ is the characteristic function of $\{h_{j+2} - h_{j+1} > 0\}$ and so on. For all nonnegative integers i_1, i_2, with $0 \leq i_1 + i_2 \leq 3$, we take $\gamma_1 > 0$ so large to get $|D_x^{i_1} D_y^{i_2} \rho(x,y)| \leq \gamma_1 \rho(x,y)$ for all $(x,y) \in \mathbb{R}^2$. Then applying Lemma 3 we obtain

$$|(D_x^{i_1} D_y^{i_2} \rho) * g_j(x,y)| \leq 6\tau\gamma_1 P(\Lambda^+) + 6\tau\gamma_1 P(\Lambda^-) + 6\tau\gamma_1 P\left(-\frac{1}{\tau}h_j(x,y)\right)$$

for all $(x,y) \in \mathbb{R}^2$. Using $D_x^{i_1} D_y^{i_2}(\rho * g_j) = (D_x^{i_1} D_y^{i_2} \rho) * g_j$, we get the desired inequality. The last two equalities follow from this inequality. $\square$

We have

$$\widetilde{\varphi}_j(x,y) = m_j^+ P(\lambda^+) + m_j^- P(\lambda^-) + (\rho * g_j)(x,y).$$

Using Lemma 4 we obtain

$$\lim_{\sqrt{x^2+y^2}\to\infty} \frac{\widetilde{\varphi}_j(x,y)}{m_j^+ P(\lambda^+) + m_j^- P(\lambda^-)} = 1,$$
$$\lim_{\sqrt{x^2+y^2}\to\infty} \frac{\boldsymbol{a}_j \cdot (\nabla\widetilde{\varphi}_j)(x,y)}{-\tau m_j^+ P'(\lambda^+)\cos(\theta_j + \frac{\pi}{2}) - \tau m_j^- P'(\lambda^-)\cos(\theta_j + \frac{\pi}{2})} = 1.$$

For all integers $i_1 \geq 0$, $i_2 \geq 0$, with $2 \leq i_1 + i_2 \leq 3$, we can estimate $|D_x^{i_1} D_y^{i_2} \widetilde{\varphi}_j(x,y)|$ by

$$\left|P^{''}(\lambda^+)\right| + \left|P^{(3)}(\lambda^+)\right| + \left|P^{''}(\lambda^-)\right| + \left|P^{(3)}(\lambda^-)\right|.$$

From Lemma 4 there exists a constant $M > 0$, with

$$|D_x^{i_1} D_y^{i_2} \widetilde{\varphi}_j(x,y)| \leq M\left(P(\lambda^+) + P(\lambda^-)\right) \qquad \text{in } \Omega_j \tag{24}$$

for every j $(1 \leq j \leq n)$ and all integers $i_1 \geq 0$, $i_2 \geq 0$, with $0 \leq i_1 + i_2 \leq 3$.

The definition of $S(x,y)$ and (21) give

$$\frac{k^3}{2c^2}\left(-2\boldsymbol{a}_j \cdot \nabla\widetilde{\varphi}_j - |\nabla\widetilde{\varphi}_j|^2\right) < S(x,y) < \frac{k^2}{c+k}\left(-2\boldsymbol{a}_j \cdot \nabla\widetilde{\varphi}_j - |\nabla\widetilde{\varphi}_j|^2\right). \tag{25}$$

Lemma 5. *For any given $\omega > 0$*

$$0 < \inf\left\{S(x,y) \mid \operatorname{dist}((x,y), E) \leq \omega\right\}$$

holds true.

Proof. It suffices to prove the lemma assuming $(x,y) \in \Omega_j$ and $\operatorname{dist}((x,y), \partial\Omega_j) \leq \omega$. We have

$$\begin{aligned}
&-2(\boldsymbol{a}_j, \nabla(m_j^+ P(\lambda^+))) - \left|\nabla(m_j^+ P(\lambda^+))\right|^2 \\
&= -P'(\lambda^+)\left(2(\boldsymbol{a}_j, \boldsymbol{a}_j - \boldsymbol{a}_{j+1}) + P'(\lambda^+)|\boldsymbol{a}_{j+1} - \boldsymbol{a}_j|^2\right) \\
&\geq -P'(\lambda^+)\left(2(\boldsymbol{a}_j, \boldsymbol{a}_j - \boldsymbol{a}_{j+1}) - \frac{1}{2}|\boldsymbol{a}_{j+1} - \boldsymbol{a}_j|^2\right) = -P'(\lambda^+)\left(|\boldsymbol{a}_j|^2 - (\boldsymbol{a}_j, \boldsymbol{a}_{j+1})\right) > 0.
\end{aligned}$$

As $\sqrt{x^2+y^2} \to \infty$, we can assume λ^+ remains finite and $\lambda^- \to \infty$ without loss of generality. Then the inequality stated above implies

$$\lim_{r\to\infty} \inf\left\{S(x,y) \,|\, (x,y) \in \Omega_j, \operatorname{dist}((x,y), \partial\Omega_j) \leq \omega, x^2 + y^2 \geq r^2\right\} > 0.$$

This completes the proof. □

Now we prove the following lemma.

Lemma 6. *There exists positive constants ν_1, ν_2 so that*

$$0 < \nu_1 \leq \frac{\varphi(x,y) - h(x,y)}{S(x,y)} \leq \nu_2$$

holds true for $(x,y) \in \mathbb{R}^2$.

Proof. We note that $(\varphi(x,y) - h(x,y))/S(x,y)$ is a positive function in $\mathbb{R}^2$. Without loss of generality, we assume (x,y) lies in Ω_j. Due to Lemma 5 it suffices to prove that it remains no less than a positive constant as $\sqrt{x^2+y^2} \to \infty$ under the condition $|\nabla\widetilde{\varphi}_j| \to 0$. We have

$$\limsup_{\sqrt{x^2+y^2}\to\infty} \left|\frac{\widetilde{\varphi}_j(x,y)}{-\boldsymbol{a}_j \cdot (\nabla\widetilde{\varphi}_j)(x,y)}\right| = \frac{1}{\tau\sin\theta_j} \limsup_{\sqrt{x^2+y^2}\to\infty} \left|\frac{m_j^+ P(\lambda^+) + m_j^- P(\lambda^-)}{-m_j^+ P'(\lambda^+) - m_j^- P'(\lambda^-)}\right|.$$

The right-hand side takes a positive bounded value. Using Lemma 5, (25), and this fact, we complete the proof. □

Proof of Proposition 1. Without loss of generality we can assume $(x, y) \in \Omega_j$ for some j. By Lemma 5 it suffices to prove

$$\sup_{(x,y)\in\Omega_j} \frac{|(D_x^{i_1} D_y^{i_2} \widetilde{\varphi}_j)(x,y)|}{S(x,y)} < +\infty \tag{26}$$

for each $i_1 \geq 0$, $i_2 \geq 0$, with $2 \leq i_1 + i_2 \leq 3$, under the condition $|\nabla \widetilde{\varphi}_j| \to 0$. From (24) we obtain

$$\lim_{\sqrt{x^2+y^2}\to\infty} \left| \frac{(D_x^{i_1} D_y^{i_2} \widetilde{\varphi}_j)(x,y)}{-\boldsymbol{a}_j \cdot (\nabla \widetilde{\varphi}_j)(x,y)} \right| \leq \frac{M'}{\tau \sin\theta_j} \lim_{\sqrt{x^2+y^2}\to\infty} \left| \frac{m_j^+ P(\lambda^+) + m_j^- P(\lambda^-)}{-m_j^+ P'(\lambda^+) - m_j^- P'(\lambda^-)} \right|.$$

Here $M' > 0$ is a constant. The right-hand side is bounded. Using this estimate, (25), and (21), we obtain (26). This completes the proof. □

3. Proof of Theorem 2. In this section we prove Theorem 2 by constructing a supersolution and a subsolution and by finding a pyramidal traveling wave between them.

For $\alpha \in (0, 1)$ we consider the graph of

$$z = \frac{1}{\alpha}\varphi(\alpha x, \alpha y). \tag{27}$$

Later we will choose α to be small enough. We use this function as a mollified pyramid. We note that

$$\frac{1}{\alpha} h(\alpha x, \alpha y) = h(x, y).$$

We use a rescaled coordinate (ξ, η, ζ) as

$$\xi = \alpha x, \quad \eta = \alpha y, \quad \zeta = \alpha z$$

and write (27) as $\zeta = \varphi(\xi, \eta)$.

For $(x_0, y_0) \in \mathbb{R}^2$, the tangent plane of (27) at $(x_0, y_0, \alpha^{-1}\varphi(\alpha x_0, \alpha y_0))$ is expressed by

$$-\varphi_\xi(\xi_0, \eta_0)(x - x_0) - \varphi_\eta(\xi_0, \eta_0)(y - y_0) + z - \frac{1}{\alpha}\varphi(\xi_0, \eta_0) = 0,$$

where $\xi_0 = \alpha x_0$, $\eta_0 = \alpha y_0$. The length of the perpendicular from (x_0, y_0, z_0) onto the tangent plane is

$$\frac{\left| z_0 - \frac{1}{\alpha}\varphi(\xi_0, \eta_0) \right|}{\sqrt{1 + \varphi_\xi(\xi_0, \eta_0)^2 + \varphi_\eta(\xi_0, \eta_0)^2}}.$$

We define

$$\widehat{\mu} \stackrel{\text{def}}{=} \frac{z - \frac{1}{\alpha}\varphi(\alpha x, \alpha y)}{\sqrt{1 + \varphi_\xi(\alpha x, \alpha y)^2 + \varphi_\eta(\alpha x, \alpha y)^2}} = \frac{1}{\alpha} \frac{\zeta - \varphi(\xi, \eta)}{\sqrt{1 + \varphi_\xi(\xi, \eta)^2 + \varphi_\eta(\xi, \eta)^2}}. \tag{28}$$

Then we have

$$\widehat{\mu}_z = \frac{1}{\sqrt{1 + \varphi_\xi^2 + \varphi_\eta^2}}, \qquad \widehat{\mu}_{zz} = 0.$$

Also we get

$$\widehat{\mu}_x = -\frac{\varphi_\xi}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}} + \alpha\widehat{\mu}F_1(\xi,\eta), \qquad \widehat{\mu}_{xx} = \alpha G_{11}(\xi,\eta) + \alpha^2\widehat{\mu}H_{11}(\xi,\eta),$$

where

$$\begin{aligned}
F_1(\xi,\eta) &\stackrel{\text{def}}{=} \sqrt{1+\varphi_\xi^2+\varphi_\eta^2}\left(\frac{1}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\right)_\xi, \\
G_{11}(\xi,\eta) &\stackrel{\text{def}}{=} -\left(\frac{\varphi_\xi}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\right)_\xi - \varphi_\xi\left(\frac{1}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\right)_\xi \\
&= \frac{(-1+\varphi_\xi^2-\varphi_\eta^2)\varphi_{\xi\xi} + (2\varphi_\xi^2+2\varphi_\xi\varphi_\eta)\varphi_{\xi\eta}}{(1+\varphi_\xi^2+\varphi_\eta^2)^{\frac{3}{2}}}, \\
H_{11}(\xi,\eta) &\stackrel{\text{def}}{=} (F_1(\xi,\eta))_\xi + F_1(\xi,\eta)^2.
\end{aligned}$$

Similarly we obtain

$$\widehat{\mu}_{xy} = \alpha G_{12}(\xi,\eta) + \alpha^2\widehat{\mu}H_{12}(\xi,\eta),$$

where

$$\begin{aligned}
G_{12}(\xi,\eta) &\stackrel{\text{def}}{=} -\left(\frac{\varphi_\xi}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\right)_\eta - \varphi_\eta\left(\frac{1}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\right)_\xi, \\
H_{12}(\xi,\eta) &\stackrel{\text{def}}{=} (F_1(\xi,\eta))_\eta + F_1(\xi,\eta)F_2(\xi,\eta).
\end{aligned}$$

We get

$$\widehat{\mu}_y = -\frac{\varphi_\eta}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}} + \alpha\widehat{\mu}F_2(\xi,\eta), \qquad \widehat{\mu}_{yy} = \alpha G_{22}(\xi,\eta) + \alpha^2\widehat{\mu}H_{22}(\xi,\eta),$$

where

$$\begin{aligned}
F_2(\xi,\eta) &\stackrel{\text{def}}{=} \sqrt{1+\varphi_\xi^2+\varphi_\eta^2}\left(\frac{1}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\right)_\eta, \\
G_{22}(\xi,\eta) &\stackrel{\text{def}}{=} -\left(\frac{\varphi_\eta}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\right)_\eta - \varphi_\eta\left(\frac{1}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\right)_\eta, \\
H_{22}(\xi,\eta) &\stackrel{\text{def}}{=} (F_2(\xi,\eta))_\eta + F_2(\xi,\eta)^2.
\end{aligned}$$

We define

$$U(x,y,z) = \Phi(\widehat{\mu}) + \sigma(x,y), \tag{29}$$

where $\widehat{\mu}$ is as in (28) and

$$\sigma(x,y) \stackrel{\text{def}}{=} \varepsilon S(\alpha x, \alpha y).$$

Here we will fix $\varepsilon > 0$ later. We have

$$U_z = \frac{1}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\Phi'(\widehat{\mu}), \quad U_{zz} = \frac{1}{1+\varphi_\xi^2+\varphi_\eta^2}\Phi''(\widehat{\mu}),$$

and

$$U_{xx}+U_{yy} = \Phi^{'}(\widehat{\mu})(\widehat{\mu}_{xx}+\widehat{\mu}_{yy}) + \Phi^{''}(\widehat{\mu})(\widehat{\mu}_x^2+\widehat{\mu}_y^2) + \sigma_{xx}+\sigma_{yy}.$$

Thus we get

$$\begin{aligned} U_{xx}+U_{yy} = {}& \alpha\Phi^{'}(\widehat{\mu})(G_{11}(\xi,\eta)+G_{22}(\xi,\eta)) + \alpha^2\widehat{\mu}\Phi^{'}(\widehat{\mu})(H_{11}(\xi,\eta)+H_{22}(\xi,\eta)) \\ & + \Phi^{''}(\widehat{\mu})\frac{\varphi_\xi^2+\varphi_\eta^2}{1+\varphi_\xi^2+\varphi_\eta^2} - 2\alpha\widehat{\mu}\Phi^{''}(\widehat{\mu})\frac{\varphi_\xi(\xi,\eta)F_1(\xi,\eta)+\varphi_\eta(\xi,\eta)F_2(\xi,\eta)}{\sqrt{1+\varphi_\xi(\xi,\eta)^2+\varphi_\eta(\xi,\eta)^2}} \\ & + \alpha^2\widehat{\mu}^2\Phi^{''}(\widehat{\mu})(F_1(\xi,\eta)^2+F_2(\xi,\eta)^2) + \sigma_{xx}+\sigma_{yy}. \end{aligned}$$

We calculate $\mathcal{L}[U]$ as

$$\begin{aligned} \mathcal{L}[U] = {}& -\Phi^{''}(\widehat{\mu}) - \frac{c}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\Phi^{'}(\widehat{\mu}) - f(\Phi+\sigma) \\ & - \alpha\Phi^{'}(\widehat{\mu})(G_{11}(\xi,\eta)+G_{22}(\xi,\eta)) - \alpha^2\widehat{\mu}\Phi^{'}(\widehat{\mu})(H_{11}(\xi,\eta)+H_{22}(\xi,\eta)) \\ & + 2\alpha\widehat{\mu}\Phi^{''}(\widehat{\mu})\frac{\varphi_\xi(\xi,\eta)F_1(\xi,\eta)+\varphi_\eta(\xi,\eta)F_2(\xi,\eta)}{\sqrt{1+\varphi_\xi(\xi,\eta)^2+\varphi_\eta(\xi,\eta)^2}} \\ & - \alpha^2\widehat{\mu}^2\Phi^{''}(\widehat{\mu})(F_1(\xi,\eta)^2+F_2(\xi,\eta)^2) - \varepsilon\alpha^2(S_{\xi\xi}+S_{\eta\eta}). \end{aligned}$$

We have

$$S_{\xi\xi}(\xi,\eta)+S_{\eta\eta}(\xi,\eta) = \left(\frac{c}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\right)_{\xi\xi} + \left(\frac{c}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\right)_{\eta\eta}$$

and define

$$\begin{aligned} R(\xi,\eta,\mu;\varepsilon,\alpha) \overset{\text{def}}{=} {}& -\Phi^{'}(\mu)(G_{11}(\xi,\eta)+G_{22}(\xi,\eta)) - \alpha\mu\Phi^{'}(\mu)(H_{11}(\xi,\eta)+H_{22}(\xi,\eta)) \\ & + 2\mu\Phi^{''}(\mu)\frac{\varphi_\xi(\xi,\eta)F_1(\xi,\eta)+\varphi_\eta(\xi,\eta)F_2(\xi,\eta)}{\sqrt{1+\varphi_\xi(\xi,\eta)^2+\varphi_\eta(\xi,\eta)^2}} \\ & - \alpha\mu^2\Phi^{''}(\mu)(F_1(\xi,\eta)^2+F_2(\xi,\eta)^2) - \varepsilon\alpha\left(S_{\xi\xi}(\xi,\eta)+S_{\eta\eta}(\xi,\eta)\right). \end{aligned}$$

Thus we get

$$\mathcal{L}[U] = -\Phi^{''}(\widehat{\mu}) - \frac{c}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\Phi^{'}(\widehat{\mu}) - f(\Phi+\sigma) + \alpha R(\xi,\eta,\widehat{\mu};\varepsilon,\alpha).$$

Using $-\Phi^{''}(\mu) - k\Phi^{'}(\mu) - f(\Phi) = 0$, we obtain

$$\mathcal{L}[U] = -\Phi^{'}(\widehat{\mu})S(\xi,\eta) - \sigma\int_0^1 f^{'}(\Phi(\widehat{\mu})+s\sigma)ds + \alpha R(\xi,\eta,\widehat{\mu};\varepsilon,\alpha).$$

We estimate $|R(\xi,\eta,\mu;\varepsilon,\alpha)|$ using

$$\begin{aligned}&|R(\xi,\eta,\mu;\varepsilon,\alpha)| \leq \max\left\{|\Phi^{'}(\mu)|, |\mu\Phi^{'}(\mu)|, |\mu\Phi^{''}(\mu)|, |\mu^2\Phi^{''}(\mu)|\right\}\\&\times (|G_{11}(\xi,\eta)| + |G_{22}(\xi,\eta)| + |H_{11}(\xi,\eta)| + |H_{22}(\xi,\eta)| + 2|F_1(\xi,\eta) + F_2(\xi,\eta)|\\&\quad +|F_1(\xi,\eta)|^2 + |F_2(\xi,\eta)|^2 + |S_{\xi\xi}(\xi,\eta)| + |S_{\eta\eta}(\xi,\eta)|)\end{aligned}$$

if $0 < \alpha < 1$. The first term $|G_{11}(\xi,\eta)|$ includes the second derivatives of φ as in the definition of G_{11}. Other terms also include the second or third derivatives of φ. Using Lemmas 1 and 2, we estimate all terms and obtain

$$\begin{aligned}&|G_{11}(\xi,\eta)| + |G_{22}(\xi,\eta)| + |H_{11}(\xi,\eta)| + |H_{22}(\xi,\eta)| + 2|F_1(\xi,\eta) + F_2(\xi,\eta)|\\&+|F_1(\xi,\eta)|^2 + |F_2(\xi,\eta)|^2 + |S_{\xi\xi}(\xi,\eta)| + |S_{\eta\eta}(\xi,\eta)|\\&\leq A' \sum_{2\leq i_1+i_2\leq 3} |(D_\xi^{i_1} D_\eta^{i_2}\varphi)(\xi,\eta)| \qquad \text{for all } (\xi,\eta)\in\mathbb{R}^2\end{aligned}$$

with a constant A'. Using Proposition 1 we find a constant A so that

$$\frac{|R(\xi,\eta,\mu;\varepsilon,\alpha)|}{S(\xi,\eta)} < A$$

holds true for all $(\xi,\eta)\in\mathbb{R}^2$, $\mu\in\mathbb{R}$, $\varepsilon\in(0,1)$, and $\alpha\in(0,1)$. Constants A' and A depend only on f and c. We continue to calculate $\mathcal{L}[U]$ as

$$\mathcal{L}[U] = S(\xi,\eta)\left(-\Phi^{'}(\widehat{\mu}) - \varepsilon\int_0^1 f^{'}(\Phi(\widehat{\mu}) + s\sigma)ds + \alpha\frac{R(\xi,\eta,\widehat{\mu};\varepsilon,\alpha)}{S(\xi,\eta)}\right).$$

Thus we get

$$\mathcal{L}[U] \geq S(\xi,\eta)\left(-\Phi^{'}(\widehat{\mu}) - \varepsilon\int_0^1 f^{'}(\Phi(\widehat{\mu}) + s\sigma)ds - \alpha A\right). \tag{30}$$

Now we choose ε and α as was mentioned before. We take ε small enough to get

$$0 < \varepsilon < \min\left\{\frac{1}{2}, \frac{\delta_*}{c}, \frac{2K_0}{c}, \frac{\min_{-1+\delta_*\leq\Phi(p)\leq 1-\delta_*}(-\Phi^{'}(p))}{4\max_{|s|\leq 1+\delta_*}|f^{'}(s)|}\right\}. \tag{31}$$

Then we choose α small enough to get

$$0 < \alpha < \min\left\{\frac{1}{2}, \frac{\varepsilon\kappa_1}{2A}, \frac{\min_{-1+\delta_*\leq\Phi(p)\leq 1-\delta_*}(-\Phi^{'}(p))}{4A}, \frac{k\kappa_0\nu_1}{\log\left(\frac{2K_0}{c\varepsilon}\right)}\right\}. \tag{32}$$

Now we show that U is a supersolution and is larger than the maximum of planar solutions.

Lemma 7. *Assume ε and α satisfy* (31) *and* (32)*, respectively. Let U be as in* (29)*. Then*

$$\mathcal{L}[U] > 0 \qquad \text{in } \mathbb{R}^3$$

holds true. Moreover

$$\Phi\left(\frac{k}{c}(z - h(x,y))\right) < U(x,y,z) \qquad \text{in } \mathbb{R}^3$$

holds true.

Proof. If $\Phi(\widehat{\mu}) < -1+\delta_*$ or $\Phi(\widehat{\mu}) > 1-\delta_*$, we have $|s\varepsilon S| \leq s\varepsilon c \leq \delta_*$ for $0 \leq s \leq 1$ in view of Lemma 2. We get $\Phi(\widehat{\mu}) + s\varepsilon S < -1 + 2\delta_*$ or $\Phi(\widehat{\mu}) + s\varepsilon S > 1 - 2\delta_*$. Combining $-\Phi'(\widehat{\mu}) > 0$ and (30), we obtain

$$\mathcal{L}[U] \geq S(\xi,\eta)\,(\varepsilon\kappa_1 - \alpha A) > 0.$$

If $-1+\delta_* \leq \Phi(\widehat{\mu}) \leq 1-\delta_*$, then we have

$$\mathcal{L}[U] \geq S(\xi,\eta)\left(\min_{-1+\delta_*\leq\Phi(p)\leq 1-\delta_*}(-\Phi^{'}(p)) - \varepsilon\max_{|s|\leq 1+\delta_*}|f^{'}(s)| - \alpha A\right) > 0.$$

In both cases we proved that U is a supersolution.

We use a similar argument as in [15] to prove the latter statement. It suffices to prove

$$\Phi\left(\frac{k}{c}(z - a_j x - b_j y)\right) < U(x,y,z) \tag{33}$$

for fixed j. Temporarily we denote a_j, b_j simply by a, b to prove (33). If

$$\widehat{\mu} \leq \frac{k}{c}(z - ax - by),$$

we get

$$U(x,y,z) > \Phi(\widehat{\mu}) \geq \Phi\left(\frac{k}{c}(z - ax - by)\right).$$

Thus it suffices to prove (33) by assuming

$$\widehat{\mu} > \frac{k}{c}(z - ax - by).$$

Substituting the definition of $\widehat{\mu}$ into this inequality, we obtain

$$\frac{z - ax - by + \left(ax + by - \frac{1}{\alpha}\varphi(\xi,\eta)\right)}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}} > \frac{k}{c}(z - ax - by),$$

which is equivalent to

$$\left(\frac{c}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}} - k\right)(z - ax - by) \geq \frac{c}{\alpha}\;\frac{\varphi(\xi,\eta) - a\xi - b\eta}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}.$$

Combining this inequality with the definition of $S(\xi,\eta)$, we get

$$z - ax - by \geq \frac{c\nu_1}{\alpha\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}} \geq \frac{k\nu_1}{\alpha}. \tag{34}$$

Using $\alpha(ax+by) = a\xi + b\eta \le \varphi(\xi,\eta)$, we obtain

$$
\begin{aligned}
U(x,y,z) &- \Phi\left(\frac{k}{c}(z-ax-by)\right) \\
&\ge \Phi\left(\frac{z-ax-by}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}}\right) - \Phi\left(\frac{k}{c}(z-ax-by)\right) + \varepsilon S(\xi,\eta) \\
= \frac{(z-ax-by)S(\xi,\eta)}{c} &\int_0^1 \Phi'\left(\left(\frac{\theta}{\sqrt{1+\varphi_\xi^2+\varphi_\eta^2}} + \frac{k}{c}(1-\theta)\right)(z-ax-by)\right) d\theta \\
&\qquad + \varepsilon S(\xi,\eta) \\
&\qquad\qquad \ge S(\xi,\eta)\left(\varepsilon - \frac{1}{c}\sup_{|\mu|\ge\frac{k\nu_1}{\alpha}}\left|\mu\Phi'\left(\frac{k}{c}\mu\right)\right|\right).
\end{aligned}
$$

By virtue of Lemma 1 and (32) we have

$$
\frac{1}{c}\sup_{|\mu|\ge\frac{k\nu_1}{\alpha}}\left|\mu\Phi'\left(\frac{k}{c}\mu\right)\right| < \frac{\varepsilon}{2}
$$

and obtain

$$
U(x,y,z) - \Phi\left(\frac{k}{c}(z-ax-by)\right) > \frac{\varepsilon}{2}S(\xi,\eta) > 0,
$$

which yields (33). This completes the proof. ∎

Thus U is a supersolution to (5). Now we prove the main assertion.

Proof of Theorem 2. We put

$$
\underline{v}(x,y,z) = \Phi\left(\frac{k}{c}(z-h(x,y))\right) \tag{35}
$$

and consider solutions of (4) given by $w(x,y,z,t;\underline{v})$ and $w(x,y,z,t;U)$. Since U is a supersolution and $\underline{v}$ is a subsolution, we have

$$
\underline{v} \le w(x,y,z,t;\underline{v}) \le w(x,y,z,t;U) \le U
$$

for $(x,y,z)\in\mathbb{R}^3$ and $t\ge 0$ by using [17, Theorem 3.4]. Then

$$
V(x,y,z) \stackrel{\text{def}}{=} \lim_{t\to\infty} w(x,y,z,t;\underline{v}) \tag{36}
$$

exists in $L^\infty(\mathbb{R}^3)$, with

$$
\underline{v}(x,y,z) < V(x,y,z) < U(x,y,z) \qquad \text{in } \mathbb{R}^3.
$$

This $V(x,y,z)$ is a solution of (5). See Sattinger [17, Theorem 3.6] for detailed arguments. Now we have

$$
\underline{v}(x,y,z) < V(x,y,z) < \Phi(\widehat{\mu}) + \varepsilon S.
$$

Now we prove (11). Let ε be arbitrarily given. Let U be as in (29). It suffices to prove

$$\sup_{(x,y,z)\in D(\gamma)} \left(U(x,y,z) - \Phi\left(\frac{k}{c}(z-h(x,y))\right)\right) < 2\varepsilon \tag{37}$$

if $\gamma > 0$ is large enough. Assume the contrary. Then there exists (γ_n) such that we have

$$\lim_{n\to\infty} \gamma_n = \infty, \quad (x_n, y_n, z_n) \in D(\gamma_n), \tag{38}$$

and

$$\left| \Phi(\widehat{\mu}_n) - \Phi\left(\frac{k}{c}(z_n - h(x_n, y_n))\right)\right| \geq \varepsilon. \tag{39}$$

Here we put $\xi_n = \alpha x_n$, $\eta_n = \alpha y_n$, $\zeta_n = \alpha z_n$, and

$$\widehat{\mu}_n = \frac{1}{\alpha} \frac{\zeta_n - \varphi(\xi_n, \eta_n)}{\sqrt{1 + \varphi_\xi(\xi_n,\eta_n)^2 + \varphi_\eta(\xi_n,\eta_n)^2}} = \frac{z_n - h(x_n,y_n) - \frac{1}{\alpha}(\varphi(\xi_n,\eta_n) - h(\xi_n,\eta_n))}{\sqrt{1 + \varphi_\xi(\xi_n,\eta_n)^2 + \varphi_\eta(\xi_n,\eta_n)^2}}.$$

If we have $\lim_{n\to\infty} \operatorname{dist}((\xi_n,\eta_n), E) = \infty$, then we obtain $\lim_{n\to\infty} |\varphi(\xi_n,\eta_n) - h(\xi_n,\eta_n)| = 0$ and $\lim_{n\to\infty} S(\xi_n,\eta_n) = 0$ by applying Lemma 4. Recall $E \stackrel{\text{def}}{=} \cup_{j=1}^n \partial\Omega_j \subset \mathbb{R}^2$. Then we get

$$\lim_{n\to\infty} \left| \widehat{\mu}_n - \frac{k}{c}(z_n - h(x_n,y_n))\right| = 0.$$

This contradicts (39). If $\operatorname{dist}((\xi_n,\eta_n), E)$ remains finite uniformly in n, then (38) implies that $\lim_{n\to\infty}(z_n - h(x_n,y_n)) = \pm\infty$ and $\lim_{n\to\infty} \widehat{\mu}_n = \pm\infty$, respectively. This contradicts (39). This completes the proof of Theorem 2. ∎

4. Application of Theorem 2. In this section we state applications of Theorem 2. Traveling waves in Theorem 2 have a contour line of a pyramidal shape if the normal vectors of lateral surfaces are linearly independent. What is the shape of traveling waves in Theorem 2 if lateral surfaces are linearly dependent? In this section we show an example of such a traveling wave.

LEMMA 8. *Let $h(x,y)$ be given by (9) with (7) and (8). Assume that $h(-x,y) = h(x,y)$ and that at least one A_j is positive. For any fixed y, assume that $h(x,y)$ is nondecreasing for $x > 0$. Then V in Theorem 2 satisfies*

$$V(-x,y,z) = V(x,y,z) \qquad \text{in } \mathbb{R}^3,$$
$$V_x(x,y,z) > 0 \qquad \text{for } x > 0.$$

The same statement holds for y.

Proof. We have $\underline{v}(-x,y,z) = \underline{v}(x,y,z)$ and thus $w(-x,y,z,t;\underline{v}) = w(x,y,z,t;\underline{v})$. Then V given by (36) satisfies $V(-x,y,z) = V(x,y,z)$. We have $(\underline{v})_x(x,y,z) \geq 0$ for $x > 0$. Now $w_x(x,y,z,t;\underline{v})$ satisfies the derivative of (4) by x in $\{(x,y,z)\in\mathbb{R}^3 \mid x > 0\}$ with the Neumann boundary condition $w_x(x,y,z,t;\underline{v}) = 0$ on $\{(x,y,z)\in\mathbb{R}^3 \mid x = 0\}$. Then the comparison principle gives $w_x \geq 0$ and thus $V_x \geq 0$ for $x > 0$. From Theorem 2, $V_x \not\equiv 0$, and thus we get $V_x > 0$. ∎

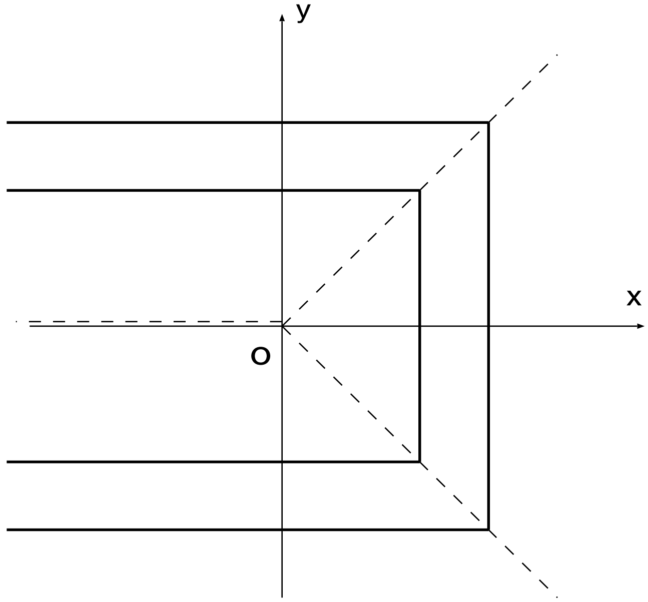


FIG. 4. *The contour lines of* $\tau \max\{x, |y|\}$.

We consider

$$h_1(x, y) = \tau y, \quad h_2(x, y) = -\tau y,$$

and thus $h(x, y) = \tau|y|$. Theorem 2 and its proof are applicable to this case. Then $V(x, y, z)$ as in Theorem 2 equals $v_*(y, z)$, where v_* is as in Theorem 1. The uniqueness follows from that of Theorem 1 in this case. We call this a planar V-form wave.

As an application of Theorem 2 we consider the following example:

$$h_1(x, y) = \tau x, \quad h_2(x, y) = \tau y, \quad h_3(x, y) = -\tau y,$$

and thus

$$h(x, y) = \max_{1 \leq j \leq 3} h_j(x, y) = \tau \max\{x, |y|\}. \tag{40}$$

See Figure 4. The edge lines are given by

$$\begin{aligned} \Gamma_1 &= \{(x, y, z) \,|\, x = y = z, z \geq 0\}, \\ \Gamma_2 &= \{(x, y, z) \,|\, x = -y = z, z \geq 0\}, \\ \Gamma_3 &= \{(x, 0, 0) \,|\, x \leq 0\}. \end{aligned}$$

We have $\Gamma = \cup_{j=1}^{3} \Gamma_j$ and $D(\gamma)$ as in (10).

PROPOSITION 2. *Assume* $c > k$, (A1), (A2), *and* (A3). *Let* $V_1(x, y, z)$ *be a solution of* (5) *as in Theorem* 2 *for* (40). *Then* $V_1(x, y, z)$ *satisfies* $V_1(x, -y, z) = V_1(x, y, z)$ *and*

$$\begin{gathered} 0 \leq V_1(x, 0, 0) \qquad \textit{for all } x \leq 0, \\ (V_1)_z(x, y, z) < 0, \quad (V_1)_x(x, y, z) > 0 \qquad \textit{in } \mathbb{R}^3, \\ (V_1)_y(x, y, z) > 0 \qquad \textit{if } (x, y, z) \in \mathbb{R}^3,\ y > 0. \end{gathered}$$

Proof. We put $v_1^-(x,y,z) \stackrel{\text{def}}{=} \Phi\left((k/c)(z-\tau\max\{x,|y|\})\right)$. It suffices to prove $(V_1)_x > 0$. We have $(v_1^-)_x \geq 0$ in $\mathbb{R}^3$. The comparison principle yields

$$w_x(x,y,z,t;v_1^-) \geq 0, \quad (V_1)_x \geq 0 \qquad \text{in } \mathbb{R}^3.$$

The maximum principle gives $(V_1)_x > 0$. $(V_1)_y > 0$ follows from Lemma 8 for $y > 0$. This completes the proof. ∎

From Theorem 2, $V_1(x,y,z)$ satisfies

$$\lim_{\gamma\to\infty} \sup_{(x,y,z)\in D(\gamma)} \left| V_1(x,y,z) - \Phi\left(\frac{k}{c}(z-\tau\max\{x,|y|\})\right)\right| = 0.$$

If $x < 0$ and $|x|$ is large enough, V_1 has a profile of the planar V-form wave. If $x > 0$ is large, V_1 has a profile of a pyramidal wave. Thus V_1 is a hybrid of them.

5. Generalized pyramidal traveling waves. The lateral surfaces of a pyramid have a common point. As a combination of planar traveling waves associated with the surfaces, we construct a pyramidal traveling wave in Theorem 2. How about if the surfaces have no common point? In this section we treat planes that have no common point and construct a generalized pyramidal traveling wave from a combination of planar traveling waves.

We introduce the following example:

$$\begin{aligned} h_1(x,y) &= \tau x, & h_2(x,y) &= \tau y, \\ h_3(x,y) &= -\tau x, & h_4(x,y) &= -\tau y. \end{aligned}$$

Then we have

$$h(x,y) = \tau\max\{|x|,|y|\}. \tag{41}$$

Let V_2 be a solution as in Theorem 2 for (41). Then Lemma 8 gives

$$\begin{aligned} (V_2)_x(x,y,z) &> 0 \qquad \text{for } x > 0, \\ (V_2)_y(x,y,z) &> 0 \qquad \text{for } y > 0. \end{aligned}$$

Let $U_2(x,y,z)$ be a supersolution as in Lemma 7 for (41). For any given $a \geq 0$, we define

$$\begin{aligned} \widetilde{h}_1(x,y) &= \tau(x-a), & \widetilde{h}_2(x,y) &= \tau y, \\ \widetilde{h}_3(x,y) &= -\tau(x+a), & \widetilde{h}_4(x,y) &= -\tau y, \end{aligned}$$

and

$$\widetilde{h}(x,y;a) \stackrel{\text{def}}{=} \max_{1\leq j\leq 4} \widetilde{h}_j(x,y;a) = \tau\max\{|y|,|x|-a\}. \tag{42}$$

The edges of a pyramid $z = \widetilde{h}(x,y;a)$ are given by

$$\begin{aligned} \widetilde{\Gamma}_1 &= \{(x,y,z) \mid z = \tau(x-a), x-a = y, z \geq 0\}, \\ \widetilde{\Gamma}_2 &= \{(x,y,z) \mid z = \tau y, y = -x-a, z \geq 0\}, \\ \widetilde{\Gamma}_3 &= \{(x,y,z) \mid z = -\tau(x+a), x+a = y, z \geq 0\}, \\ \widetilde{\Gamma}_4 &= \{(x,y,z) \mid z = -\tau y, -y = x-a, z \geq 0\}. \end{aligned}$$

We put $\widetilde{\Gamma} = \cup_{j=1}^4 \widetilde{\Gamma}_j$ and

$$\widetilde{D}(\gamma) \stackrel{\text{def}}{=} \left\{ (x,y,z) \in \mathbb{R}^3 \mid \operatorname{dist}((x,y,z), \widetilde{\Gamma}) > \gamma \right\}.$$

We set

$$v_2^-(x,y,z) \stackrel{\text{def}}{=} \Phi\left(\frac{k}{c}(z - \widetilde{h}(x,y;a))\right) = \max_{1\le j\le 4} \Phi\left(\frac{k}{c}(z - \widetilde{h}_j(x,y;a))\right).$$

Let $w(x,y,z,t;v_2^-)$ be the solution of (4) with an initial condition $w|_{t=0} = v_2^-$. From the comparison principle we obtain

$$v_2^-(x,y,z) < w(x,y,z,t;v_2^-) < U_2(x-x_0,y,z) \tag{43}$$

for any x_0 with $|x_0| \le a$. Thus we get

$$v_2^-(x,y,z) < w(x,y,z,t;v_2^-) \le \inf_{-a\le x_0\le a} U_2(x-x_0,y,z).$$

Then we get the limit function

$$\widetilde{V}(x,y,z) \stackrel{\text{def}}{=} \lim_{t\to\infty} w(x,y,z,t;v_2^-) \qquad \text{in } C^2_{\text{loc}}(\mathbb{R}^3).$$

This satisfies (5). See Sattinger [17] for the general arguments. For every $x_0 \in [-a,a]$ we have

$$\widetilde{h}(x,y;a) \le \tau \max\{|y|, |x-x_0|\}$$

and thus

$$v_2^-(x,y,z) \le \Phi\left(\frac{k}{c}(z - \tau\max\{|y|,|x-x_0|\})\right).$$

We consider each side as an initial function of (4) and send $t \to \infty$. Then we get

$$\Phi\left(\frac{k}{c}(z - \tau\max\{|y|,|x|-a\})\right) < \widetilde{V}(x,y,z) < V_2(x-x_0,y,z).$$

The strict inequality follows from the strong maximum principle. See Figure 5.

THEOREM 3. *Assume $c > k$, (A1), (A2), and (A3). Let V_2 be the solution of (5) in Theorem 2 for $h(x,y) = \tau\max\{|x|,|y|\}$. There exists a solution $\widetilde{V}(x,y,z)$ to (5) with*

$$\Phi\left(\frac{k}{c}(z - \tau\max\{|y|,|x|-a\})\right) < \widetilde{V}(x,y,z) < \inf_{-a\le x_0\le a} V_2(x-x_0,y,z)$$

and

$$(\widetilde{V})_z(x,y,z) < 0 \qquad \text{in } \mathbb{R}^3.$$

$\widetilde{V}$ satisfies $\widetilde{V}(-x,y,z) = \widetilde{V}(x,y,z)$, $\widetilde{V}(x,-y,z) = \widetilde{V}(x,y,z)$, and

$$(\widetilde{V})_x(x,y,z) > 0 \qquad \text{for } x > 0,$$
$$(\widetilde{V})_y(x,y,z) > 0 \qquad \text{for } y > 0.$$

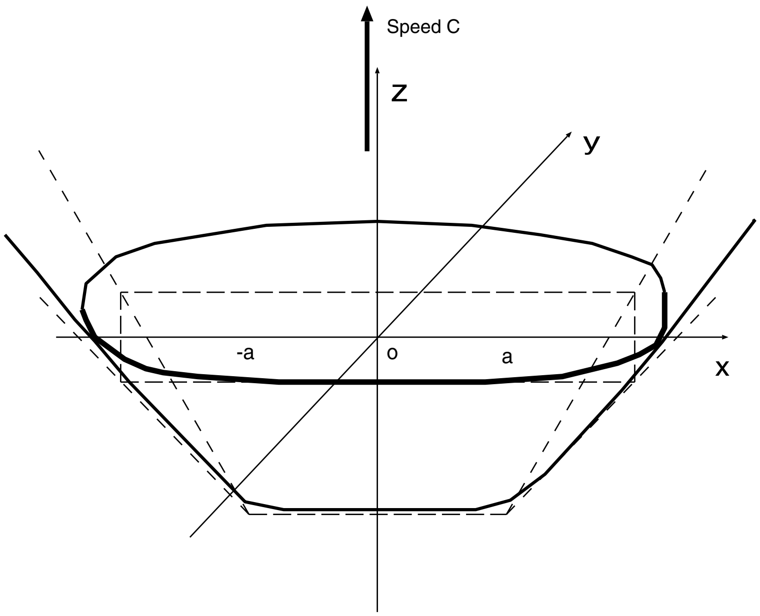


FIG. 5. *A generalized pyramidal traveling wave.*

Moreover

$$\lim_{\gamma\to\infty} \sup_{(x,y,z)\in D(\gamma)} \left| \widetilde{V}(x,y,z) - \Phi\left(\frac{k}{c}(z - \widetilde{h}(x,y;a))\right) \right| = 0$$

holds true.

Proof. Since $(v_2^-)_z \le 0$, we get $w_z(x,y,z,t;v_2^-) \le 0$ and also get $(\widetilde{V})_z < 0$. Lemma 8 and the proof are applicable to $\widetilde{h}(x,y;a)$. Thus we get $(\widetilde{V})_x > 0$ for $x > 0$ and $(\widetilde{V})_y > 0$ for $y > 0$. The asymptotic property of $\widetilde{V}(x,y,z)$ follows from that of V_2 in Theorem 2. ◻

This $\widetilde{V}(x,y,z)$ is a generalized pyramidal traveling wave. The method of this section might be applicable to a general case. The classification of all generalized pyramidal waves will give interesting problems.

6. Traveling fronts for balanced bistable nonlinearity. In this section we study traveling waves for balanced nonlinearity. Recently Chen et al. [3] constructed two-dimensional traveling waves and n-dimensional cylindrically symmetric traveling waves for balanced nonlinearity. They constructed such traveling waves as the limit of traveling waves for an unbalanced nonlinearity term when the difference of the energy density goes to zero.

Now we construct traveling waves for balanced nonlinearity by taking the limit of pyramidal traveling waves for unbalanced nonlinearity terms when the difference of the energy density goes to zero.

We consider (2) with a balanced nonlinear term $-G'(u)$. Let $c > 0$ be arbitrarily fixed. We study (13) in section 1. We define

(44) $$\mathcal{L}_\delta[v] \stackrel{\text{def}}{=} -v_{xx} - v_{yy} - v_{zz} - cv_z - f_\delta(v) = 0 \quad \text{in } \mathbb{R}^3$$

for any δ with $0 < \delta < 1$, where

$$f_\delta(v) \stackrel{\text{def}}{=} -G'(v) + \delta c\sqrt{2G(v)}.$$

Putting $k = \delta c$, we see that $\Phi_0(\mu)$ given by (3) satisfies (1). Let $V_\delta(x,y,z)$ be a solution of (44) as in Theorem 2 for

$$h_\delta(x,y) = \frac{\sqrt{1-\delta^2}}{\delta}\max\{|x|,|y|\}.$$

We fix $\lambda_1 \in (-1,1)$, with $G'(\lambda_1) < 0$. Let $z_1(\delta)$ be defined by

$$V_\delta(0,0,z_1(\delta)) = \lambda_1. \tag{45}$$

We construct a solution of (13) as the limit of $V_\delta(x,y,z+z_1(\delta))$.

PROPOSITION 3. *Assume* (B1) *and* (B2). *Let $c > 0$ be arbitrarily fixed. Let $V_\delta(x,y,z)$ be a solution of* (44) *as in Theorem* 2 *for $h_\delta(x,y) = (\sqrt{1-\delta^2}/\delta)\max\{|x|,|y|\}$. There exists $1 > \delta_1 > \delta_2 > \cdots > \delta_i > \cdots \to 0$ so that one has*

$$\lim_{i\to\infty} V_{\delta_i}(x,y,z+z_1(\delta_i)) = V_*(x,y,z) \quad \text{in } C^2_{\text{loc}}(\mathbb{R}^3).$$

This solution V_* satisfies $V_*(0,0,0) = \lambda_1$ and

$$\mathcal{L}_0[V_*] = 0, \quad (V_*)_z < 0 \qquad \text{in } \mathbb{R}^3.$$

Proof. We denote $V_{\delta_i}(x,y,z+z_1(\delta_i))$ simply by $v_i(x,y,z)$. Let $B(N)$ be a closed ball defined by

$$B(N) \stackrel{\text{def}}{=} \left\{(x,y,z) \mid \sqrt{x^2+y^2+z^2} \le N\right\}$$

for $N \in \mathbb{N}$. For any fixed N, $v_i(x,y,z)$ satisfies

$$\mathcal{L}[v_i] = 0, \quad -1 < v_i < 1 \qquad \text{in } B(N).$$

For any $p > 1$, (v_i) is bounded in $L^p(B(N))$. The Schauder interior estimates [6, Theorem 9.11] imply that

$$\sup_i \|v_i\|_{W^{2,p}(B(N))} < \infty.$$

We take p so large as to get $1 - 3/p > \beta > 0$. Then $W^{2,p}(B(N))$ is compactly embedded in $C^{1,\beta}(\overline{B(N)})$. By taking a subsequence (v_i) converges in $C^{1,\beta}(\overline{B(N)})$ as $i \to \infty$. Applying the Schauder interior estimates [6, Corollary 6.3] again, we find that (v_i) converges in $C^{2,\beta}(\overline{B(N)})$. By the diagonal argument we find a subsequence (v_i) that converges in $C^{2,\beta}_{\text{loc}}(\mathbb{R}^3)$. Let V_* be the limit function. Then it satisfies (13). Since $(v_i)_z < 0$ in $\mathbb{R}^3$, we have $(V_*)_z \le 0$ in $\mathbb{R}^3$. From Lemma 8 we have $(v_i)_{xx}(0,0,0) \ge 0$ and $(v_i)_{yy}(0,0,0) \ge 0$ and thus $(V_*)_{xx}(0,0,0) \ge 0$ and $(V_*)_{yy}(0,0,0) \ge 0$. If $(V_*)_z \equiv 0$, we obtain a contradiction by $G'(\lambda_1) < 0$ and $\mathcal{L}_0[V_*] = 0$ at the origin. By the strong maximum principle, we get $(V_*)_z < 0$ in $\mathbb{R}^3$. ∎

This V_* might inherit pyramidal structures, or it might not. This problem is yet to be studied. If we replace $h_\delta(x,y)$ by $(\sqrt{1-\delta^2}/\delta)\max_{1\le j\le n}(A_jx + B_jy)$ with (7) and (8), we get the associated limit traveling waves from the argument stated above and also find interesting open problems. The classification and the stability of all traveling waves for unbalanced and balanced nonlinearity have a wide variety of unknown problems and are left for further studies.

Acknowledgments. The author thanks Professor Hirokazu Ninomiya of Ryukoku University for many stimulating and valuable discussions. Special thanks go to the referees, who carefully read the manuscript and gave important and useful comments and suggestions.

REFERENCES

[1] D. G. Aronson and H. F. Weinberger, *Nonlinear diffusion in population genetics*, in Partial Differential Equations and Related Topics, Lecture Notes in Math. 446, J. A. Goldstein, ed., Springer, New York, 1975., pp. 5–49.

[2] X. Chen, *Existence, uniqueness, and asymptotic stability of traveling waves in nonlocal evolution equations*, Adv. Differential Equations, 2 (1997), pp. 125–160.

[3] X. Chen, J-S. Guo, F. Hamel, H. Ninomiya, and J-M Roquejoffre, *Traveling waves with paraboloid like interfaces for balanced bistable dynamics*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 24 (2007), pp. 369–393.

[4] P. C. Fife, *Dynamics of internal layers and diffusive interfaces*, in CBMS-NSF Regional Conf. Ser. in Appl. Math. 53, SIAM, Philadelphia, 1988.

[5] P. C. Fife and J. B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Ration. Mech. Anal., 65 (1977), pp. 335–361.

[6] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin, 1983.

[7] F. Hamel, R. Monneau, and J.-M. Roquejoffre, *Stability of travelling waves in a model for conical flames in two space dimensions*, Ann. Sci. École Norm. Sup. (4), 37 (2004), pp. 469–506.

[8] F. Hamel, R. Monneau and J.-M. Roquejoffre, *Existence and qualitative properties of multidimensional conical bistable fronts*, Discrete Contin. Dyn. Syst., 13 (2005), pp. 1069–1096.

[9] F. Hamel, R. Monneau, and J.-M. Roquejoffre, *Asymptotic properties and classification of bistable fronts with Lipschitz level sets*, Discrete Contin. Dyn. Syst., 14 (2006), pp. 75–92.

[10] Y. I. Kanel', *Certain problems on equations in the theory of burning*, Sov. Math. Dokl., 2 (1961), pp. 48–51.

[11] Y. I. Kanel', *Stabilization of solutions of the Cauchy problem for equations encountered in combustion theory*, Mat. Sb., 59 (1962), pp. 245–288.

[12] F. Hamel and N. Nadirashvili, *Travelling fronts and entire solutions of the Fisher-KPP equation in* $\mathbb{R}^N$, Arch. Ration. Mech. Anal., 157 (2001), pp. 91–163.

[13] M. Haragus and A. Scheel, *Corner defects in almost planar interface propagation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 23 (2006), pp. 283–329.

[14] J. Nagumo, S. Yoshizawa, and S. Arimoto, *Bistable transmission lines*, IEEE Trans. Circuit Theory, CT-12, No. 3, 1965, pp. 400–412.

[15] H. Ninomiya and M. Taniguchi, *Existence and global stability of traveling curved fronts in the Allen-Cahn equations*, J. Differential Equations, 213 (2005), pp. 204–233.

[16] H. Ninomiya and M. Taniguchi, *Global stability of traveling curved fronts in the Allen-Cahn equations*, Discrete Contin. Dyn. Syst., 15 (2006), pp. 819–832.

[17] D. H. Sattinger, *Monotone methods in nonlinear elliptic and parabolic boundary value problems*, Indiana Univ. Math. J., 21 (1972), pp. 979–1000.

[18] D. Terman, *Directed graphs and traveling waves*, Trans. Amer. Math. Soc., 289 (1985), pp. 809–847.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 345–378

SELF-SIMILAR SOLUTIONS TO THE OORT–HULST–SAFRONOV COAGULATION EQUATION*

VÉRONIQUE BAGLAND† AND PHILIPPE LAURENÇOT‡

Abstract. The existence of self-similar solutions with a finite first moment is established for the Oort–Hulst–Safronov coagulation equation when the coagulation kernel is given by $a(y, y_*) = y^\lambda + y_*^\lambda$ for some $\lambda \in (0, 1)$. The corresponding self-similar profiles are compactly supported and have a discontinuity at the edge of their support.

Key words. coagulation equation, Oort–Hulst–Safronov model, self-similar solution, compact support

AMS subject classifications. 45K05, 45M05, 82C21

DOI. 10.1137/060658333

1. Introduction.

Coagulation models provide a mean-field description of particle growth, the particles increasing their size as a consequence of successive binary mergers. Assuming the particles to be fully identified by their size, the resulting equations determine the dynamics of the size distribution function $g(t, v) \geq 0$ of particles of size $v \in \mathbb{R}_+ := (0, \infty)$ at time $t \geq 0$. Of particular interest are the predictions concerning the large time behavior of the size distribution of the particles which can be drawn from these models. It is actually commonly expected that, for coagulation kernels with a moderate growth for large sizes, the size distribution function g should approach a mass-conserving self-similar function g_s for large times; that is,

$$g(t, v) \sim g_s(t, v) = \frac{1}{s(t)^2}\, \xi\left(\frac{v}{s(t)}\right) \quad \text{as} \quad t \to \infty, \tag{1.1}$$

where $s(t)$ denotes the mean particle size at time $t > 0$ and ξ is a nonnegative function in $L^1(\mathbb{R}_+, vdv)$ (see the extensive discussion in [4, 19] for Smoluchowski's coagulation equation [24, 25]). However, the validity of the dynamical scaling hypothesis (1.1) has up to now mainly been checked by numerical simulations [7, 14, 18] and is not yet established rigorously (except in a few particular cases, [2, 13, 16, 17, 20]). The first difficulty encountered is actually the existence of the profile ξ, and such a result has been obtained only recently for Smoluchowski's coagulation equation [6, 8]. The aim of the present paper is to investigate the existence of self-similar solutions for another coagulation equation which has been proposed in astrophysics by Oort and van de Hulst [21] and Safronov [22] in order to describe the aggregation of stellar objects.

*Received by the editors April 27, 2006; accepted for publication (in revised form) December 19, 2006; published electronically June 5, 2007.
http://www.siam.org/journals/sima/39-2/65833.html

†Laboratoire de Mathématiques, Université Blaise Pascal, Campus Universitaire des Cézeaux, F–63177 Aubière cedex, France (Veronique.Bagland@math.univ-bpclermont.fr).
‡Mathématiques pour l'Industrie et la Physique, CNRS UMR 5640, Université Paul Sabatier – Toulouse 3, 118 route de Narbonne, F–31062 Toulouse cedex 9, France (laurenco@mip.ups-tlse.fr).

More precisely, the Oort–Hulst–Safronov (OHS) equation reads

$$\partial_t g = Q_{OHS}(g), \quad (t,v) \in (0,\infty)^2, \tag{1.2}$$

$$Q_{OHS}(g)(v) := -\partial_v \left(\int_0^v v_*\ a(v,v_*)\ g(v_*)\ dv_*\ g(v) \right)$$

$$- \int_v^\infty a(v,v_*)\ g(v_*)\ dv_*\ g(v), \tag{1.3}$$

$$g(0) = g_0, \tag{1.4}$$

where the coagulation kernel a is a nonnegative and symmetric function. On one hand, the first term on the right-hand side of (1.3) describes the formation of clusters of size v from smaller clusters with the growth rate

$$\int_0^v v_*\ a(v,v_*)\ g(v_*)\ dv_*$$

which does not depend on the sizes of the clusters involved in the coagulation reaction but on the size distribution of the small clusters. On the other hand, the second term on the right-hand side of (1.3) accounts for the disappearance of clusters of size v by "sedimentation" on larger clusters. Let us mention at this point that a solution to (1.2), (1.4) is expected to satisfy the conservation of mass

$$\int_0^\infty v\ g(t,v)\ dv = \int_0^\infty v\ g_0(v)\ dv \quad \text{for} \quad t \geq 0, \tag{1.5}$$

a property which holds true if a does not grow too fast for large values of v and v_*. For the coagulation kernels we consider in this paper, the weak solutions constructed in [15] enjoy the mass conservation property.

For a homogeneous coagulation kernel satisfying

$$a(uv, uv_*) = u^\lambda a(v,v_*), \qquad (u,v,v_*) \in (0,\infty)^3,$$

for some $\lambda \in (-\infty, 1)$, it is rather natural to expect that the self-similar function g_s is in fact a self-similar solution to (1.2). Inserting the self-similar ansatz (1.1) in (1.2), we obtain $s(t) := (w(1-\lambda)t)^{1/(1-\lambda)}$, while ξ satisfies

$$w\,[y\,\partial_y \xi(y) + 2\,\xi(y)] + Q_{OHS}(\xi)(y) = 0, \qquad y \in (0,\infty), \tag{1.6}$$

for some positive constant w. In addition, in view of (1.5), we require that $\xi \in L^1(0,\infty; y\,dy)$ with

$$\int_0^\infty \xi(y)\,y\,dy = \varrho \tag{1.7}$$

for some positive constant ϱ. We first observe that, if ξ is a solution to (1.6), (1.7) for the parameters (w,ϱ), then the function $\xi_{A,B}$ defined by $\xi_{A,B}(y) = A\ \xi(B\,y)$, $A > 0$, $B > 0$, is a solution to (1.6), (1.7) for $\left(w\ A\ B^{-(1+\lambda)}, \varrho\ A\ B^{-2}\right)$. Choosing

$$A = \frac{1}{\varrho}\ \left(\frac{\varrho}{w\,(1-\lambda)}\right)^{2/(1-\lambda)}, \qquad B = \left(\frac{\varrho}{w\,(1-\lambda)}\right)^{1/(1-\lambda)},$$

then the function $\xi_{A,B}$ is a solution to (1.6), (1.7) for $(1/(1-\lambda), 1)$. We therefore may assume without loss of generality that

$$w = \gamma := 1/(1-\lambda) \qquad \text{and} \qquad \varrho = 1.$$

From now on, we assume further that the coagulation kernel is given by

$$a(y,y_*) = y^\lambda + y_*^\lambda, \quad (y,y_*) \in (0,\infty)^2, \tag{1.8}$$

for some $\lambda \in (0,1)$ and prove the following result.

Theorem 1.1. *There exists a nonnegative function $\xi \in L^1(0,\infty;(y+y^\lambda)\ dy)$, $y_0 \in (0,\infty)$, and $q \in (0,\infty)$ such that*

$$\int_0^\infty y\ \xi(y)\ dy = 1, \quad M_\lambda := \int_0^\infty y^\lambda\ \xi(y)\ dy \in (1,\gamma), \tag{1.9}$$

$$\xi \in \mathcal{C}((0,\infty)\setminus\{y_0\}) \quad \textit{with} \quad \mathrm{Supp}\ \xi = [0,y_0] \quad \textit{and} \quad \xi(y_0-) > 0, \tag{1.10}$$

$$\lim_{y\to 0} y^\tau\ \xi(y) = q \quad \textit{with} \quad \tau := 2 - \frac{M_\lambda}{\gamma} \in (1,1+\lambda) \quad \textit{and} \quad \gamma = \frac{1}{1-\lambda}, \tag{1.11}$$

and satisfying

$$\left(\gamma\ y - \int_0^y \left(y^\lambda + y_*^\lambda\right)\ y_*\ \xi(y_*)\ dy_*\right)\ \xi(y) = \left(\gamma - \int_y^\infty y_*^\lambda\ \xi(y_*)\ dy_*\right)\ \int_y^\infty \xi(y_*)\ dy_* \tag{1.12}$$

for $y \in (0,y_0)$.

We first note that (1.12) is in fact a weak formulation of (1.6) but that ξ satisfies (1.6) in $(0,y_0)$ and (y_0,∞). Next, on one hand, we point out that the profile ξ has a singularity at $y=0$ and is actually not integrable near $y=0$ by (1.11): a similar property is enjoyed by the profile of self-similar solutions to Smoluchowski's coagulation equation (see [4, 19] for formal arguments and [5, 9] for a rigorous proof). By the way, the exponent of the singularity of ξ can be predicted by formal computations similar to the ones done for the Smoluchowski coagulation equation [4, 19]. On the other hand, a striking difference between the profiles of self-similar solutions to the OHS and Smoluchowski coagulation equations is that the former are compactly supported with a discontinuity at the edge of the support while the latter belong to $\mathcal{C}^1((0,\infty))$ with infinite support [9]. Such a difference had already been noticed for the constant coagulation kernel ($a \equiv 1$), for which self-similar profiles to Smoluchowski's coagulation equation are explicitly given by $y \mapsto (4/\varrho)\ e^{-2y/\varrho}$ [4, 19], while self-similar profiles to (1.2) are also explicit and given by $y \mapsto (2/\varrho)\mathbf{1}_{[0,\varrho]}(y)$ [16].

The function ξ constructed in Theorem 1.1 allows us to construct a self-similar solution to (1.2) (in a weak sense).

Corollary 1.2. *Let ξ be the function constructed in Theorem* 1.1 *and put*

$$g_s(t,v) := t^{-2\gamma}\ \xi\left(vt^{-\gamma}\right), \qquad (t,v) \in (0,\infty)^2.$$

Then g_s is a weak solution to (1.2)*, that is,*

$$\begin{aligned}&\frac{d}{dt}\int_0^\infty g_s(t,v)\ \vartheta(v)\ dv \\ &\quad = \int_0^\infty \int_0^v \left(v_*\ \partial_v\vartheta(v) - \vartheta(v_*)\right)\ \left(v^\lambda + v_*^\lambda\right)\ g_s(t,v_*)\ g_s(t,v)\ dv_* dv\end{aligned}$$

for every $t \in (0,\infty)$ and $\vartheta \in \mathcal{C}_0^\infty((0,\infty))$. In addition, for $t > 0$,

$$\int_0^\infty v\ g_s(t,v)\ dv = \int_0^\infty v\ \xi(v)\ dv = 1.$$

Remark 1.3. More generally, given $w > 0$ and $\varrho > 0$, the function

$$(t,v) \longmapsto \frac{1}{s_w(t)^2}\ A\ \xi\left(\frac{B\ v}{s_w(t)}\right)$$

with

$$s_w(t) := (w(1-\lambda)t)^{1/(1-\lambda)}, \quad A := \varrho\ \left(\frac{w(1-\lambda)}{\varrho}\right)^{2/(1-\lambda)}, \quad B = \left(\frac{w(1-\lambda)}{\varrho}\right)^{1/(1-\lambda)},$$

is also a weak (self-similar) solution to (1.2) in the sense of Corollary 1.2 with a first moment equal to ϱ for each $t > 0$. Observe that all the profiles $y \mapsto A\xi(By)$ of these self-similar solutions have the same singular behavior as $y \to 0$. In particular, it does not depend on the parameter ϱ.

Owing to the nonintegrability of ξ, $v \longmapsto g_s(t,v)$ does not belong to $L^1(\mathbb{R}_+)$ for $t > 0$ and actually solves (1.2) in a weaker sense than the solutions constructed in [15]. Since $g_s(t)$ has a finite moment of order λ for each $t > 0$, the above formulation actually makes sense for any test function $\vartheta \in \mathcal{C}^1(\mathbb{R}_+)$ such that $|\partial_v\vartheta(v)| \le C\ v^{\lambda-1}$ and $\vartheta(0) = 0$. In fact, in view of the well-posedness results obtained recently for the Smoluchowski coagulation equation [10], it is likely that the existence of (weak) solutions can be shown for nonnegative initial data in $L^1(\mathbb{R}_+, v^\lambda dv)$.

To prove Theorem 1.1, we use a dynamical approach as in [6, 8, 11] relying on the fact that finding profiles ξ that satisfy (1.6), (1.7) amounts to finding steady states to

$$\partial_t f = \gamma\ \ (y\ \partial_y f + 2\ f) + Q_{OHS}(f), \quad (t,y) \in \mathbb{R}_+^2. \tag{1.13}$$

To this aim, we recall the following result.

THEOREM 1.4. *Let X be a locally convex topological vector space and K be a nonempty compact and convex subset of X. If $\mathcal{F}\ : [0,\infty)\times K \to X$ is a semiflow on X for which $\mathcal{F}(t,K) \subset K$ for each $t \ge 0$, then there is $x_0 \in K$ such that $\mathcal{F}(t,x_0) = x_0$ for each $t \ge 0$. In other words, x_0 is a steady state for the semiflow $\mathcal{F}$ in K.*

The proof of Theorem 1.4 relies on the Tychonov–Schauder fixed point theorem (or Brouwer fixed point theorem if $\dim X < \infty$) [1, 6, 11]. For the sake of completeness, the proof of Theorem 1.4 is given in Appendix A.

Applying Theorem 1.4 thus requires finding a functional setting in which (1.13) is well-posed and possesses a compact and convex invariant set as well. The existence of mass-conserving weak solutions to (1.2), (1.4) in $L^1(\mathbb{R}_+, (1+v)\ dv)$ has been established in [15]: noting that, if f is a solution to (1.13), (1.4), then

$$g(t,v) = \frac{1}{(1+t)^{2\gamma}}\ f\left(\ln(1+t), \frac{v}{(1+t)^\gamma}\right)$$

is a solution to (1.2), (1.4), we thus also obtain weak solutions to (1.13), (1.4) in the same functional setting. However, since the function ξ constructed in Theorem 1.1 does not belong to $L^1(\mathbb{R}_+)$, and g_s defined in Corollary 1.2 is not a weak solution to (1.2) in the sense of [15], we thus have to find a more appropriate functional setting. Furthermore, as far as uniqueness and continuous dependence are concerned, it turns out that it is more convenient to construct a semiflow in $L^1(\mathbb{R}_+)$ for the cumulative distribution function F defined by

$$F(t,y) = \int_y^\infty f(t,y_*)\,dy_*, \qquad (t,y) \in \mathbb{R}_+^2.$$

But at this point another serious difficulty to be bypassed arises: indeed, for each $r > 0$, $\gamma\ r^{-\lambda}\ \delta_{y=r}$ is a measure-valued stationary solution to (1.13), $\delta_{y=r}$ denoting the Dirac measure at $y = r$. A consequence of this remark and Theorem 1.1 is that (1.12) has singular (measure-valued) and "regular" solutions. As the singular ones are explicit, our aim is to construct a "regular" solution, and thus the functional setting to be used must exclude the singular solutions. Unfortunately, we have been unable to find a suitable functional framework for either f or F in which the application of Theorem 1.4 warrants the existence of a "regular" solution to (1.12).

One difference between the singular solutions $\gamma\ r^{-\lambda}\ \delta_{y=r}$, $r > 0$ and the "regular" solutions we are looking for is that the latter are expected to have an invertible primitive while the primitive of the former has only a generalized inverse. This property leads us to study the inverse function $\Phi(t,.)$ of $F(t,.)$, hoping that a similar difficulty will not show up with this alternative formulation. Formally, Φ is a solution to

$$\begin{aligned}\partial_t\Phi(t,x) &= -\gamma\ \left(x\ \partial_x\Phi(t,x)+\Phi(t,x)\right)+x\ \left(\int_0^x \Phi(t,x_*)^\lambda\ dx_*\right)\ \partial_x\Phi(t,x)\\ &\quad+\int_x^\infty \left(\Phi(t,x)^\lambda+\Phi(t,x_*)^\lambda\right)\ \Phi(t,x_*)\ dx_*. \end{aligned}\tag{1.14}$$

A weak stationary solution Ψ to (1.14) then solves

$$x\ \left(\gamma-\int_0^x \Psi(x_*)^\lambda\ dx_*\right)\ \Psi(x)=\int_x^\infty\int_0^x \left(\Psi(x_*)^\lambda+\Psi(x')^\lambda\right)\ \Psi(x_*)\ dx'dx_*$$

for $x > 0$ and this equation has no solution of the form $A\ \mathbf{1}_{[0,B]}$.

A natural functional setting for the well-posedness of (1.14) is the set

$$K_R := \left\{\begin{array}{l} U\ \in\ L^1(0,\infty) \text{ is a nonnegative and nonincreasing function such that}\\ \|U\|_1=1,\ \|U\|_\infty\le R \text{ and } \|U^\lambda\|_1\le R \end{array}\right\},$$

which is also invariant for sufficiently large R. Unfortunately, K_R is not convex since $\lambda \in (0,1)$. Therefore we cannot use Theorem 1.4, and this leads us to introduce the following modified equation:

$$\begin{aligned}\partial_t\Phi(t,x) &= -\gamma\ \left(x\ \partial_x\Phi(t,x)+\Phi(t,x)\right)+x\ \left(\int_0^x \Phi(t,x_*)^\lambda\ dx_*+\delta x\right)\ \partial_x\Phi(t,x)\\ &\quad+\int_x^\infty \left(\Phi(t,x)^\lambda+\Phi(t,x_*)^\lambda+2\delta\right)\ \Phi(t,x_*)\ dx_*, \end{aligned}\tag{1.15}$$

where $\delta \in (0,1)$. Let us mention at this point that (1.15) can be obtained from (1.13) with the coagulation kernel $a_\delta(y,y_*) = y^\lambda + y_*^\lambda + 2\delta$ by the same procedure as (1.14). We then investigate the existence of steady states to (1.15). For that purpose, we first study in section 2 the well-posedness of (1.15) supplemented with the initial condition

$$\Phi(0,x)=\Phi_0(x).\tag{1.16}$$

We then determine in section 3 a compact and convex set which is left invariant by the semiflow induced by (1.15), (1.16). The existence of a stationary solution to (1.15) then follows by applying Theorem 1.4. We obtain a stationary solution to (1.14) by letting $\delta \to 0$. Section 4 is then devoted to the analysis of the smoothness of this stationary solution. The proof of Theorem 1.1 is carried out in section 5.

Remark 1.5. As a final comment, let us point out that Theorem 1.1 and Corollary 1.2 are only a first step towards the study of the validity of the *dynamical scaling hypothesis* (1.1). Indeed, our analysis shows the existence of at least one self-similar solution with a first moment being constant through time evolution and equal to ϱ for each $\varrho > 0$ (see Remark 1.3) but we do not have any clue concerning the stability of this solution. A related open question is the uniqueness of the profile ξ given by Theorem 1.1 (in the class of functions).

Finally, we introduce some notations: for any $u, v \in \mathbb{R}$, we define

$$\begin{aligned} u \wedge v &= \min\{u,v\}, \qquad u \vee v = \max\{u,v\}, \\ u_+ &= \max\{u,0\}, \qquad \operatorname{sign}_+(u) = \operatorname{sign}(u_+). \end{aligned}$$

For any $p \in [1,\infty]$ and $\zeta \in L^p(0,\infty)$, we set

$$\|\zeta\|_p = \|\zeta\|_{L^p(0,\infty)}.$$

2. Well-posedness of (1.15), (1.16). In this section we prove the following theorem.

THEOREM 2.1. *Assume that*

$$\begin{cases} \Phi_0 \in L^\infty(\mathbb{R}_+) \text{ is a nonnegative and nonincreasing compactly supported function} \\ \text{such that } Supp\ \Phi_0 \subset [0,R_0] \text{ for some } R_0 > 0. \end{cases} \tag{2.1}$$

Then there is a unique function $\Phi \in \mathcal{C}([0,\infty); L^1(0,\infty))$ such that

- *$\Phi(t,.)$ is nonnegative and nonincreasing with compact support in $[0, e^{\gamma t} R_0]$,*
- *$\|\Phi(t)\|_1 = \|\Phi_0\|_1$ and $\Phi \in L^\infty((0,t)\times(0,\infty))$,*

and Φ satisfies (1.15) *in the following weak sense:*

$$\begin{aligned} &\frac{d}{dt}\int_0^\infty \Phi(t,y)\ \vartheta(y)\ dy \\ &\quad = \int_0^\infty \partial_y\vartheta(y)\ \left(\gamma\ y - \delta\ y^2 - y\int_0^y \Phi(t,y_*)^\lambda\ dy_*\right)\ \Phi(t,y)\ dy \\ &\quad - \int_0^\infty \partial_y\vartheta(y)\ \int_y^\infty\int_0^y \left(\Phi(t,y_*)^\lambda + \Phi(t,y')^\lambda + 2\delta\right)\ \Phi(t,y_*)\ dy' dy_*\ dy \end{aligned} \tag{2.2}$$

for every $t \geq 0$ and $\vartheta \in \mathcal{C}_0^\infty((0,\infty))$.

Before proving Theorem 2.1, we first observe that, if φ is a weak solution to

$$\begin{aligned} \partial_t\varphi(t,x) &= x\ \left(\int_0^x \varphi(t,x_*)^\lambda\ dx_* + \delta x(1+t)^{\lambda\gamma}\right)\ \partial_x\varphi(t,x) \\ &\quad + \int_x^\infty \left(\varphi(t,x)^\lambda + \varphi(t,x_*)^\lambda + 2\delta(1+t)^{\lambda\gamma}\right)\ \varphi(t,x_*)\ dx_*, \end{aligned} \tag{2.3}$$

$$\varphi(0,x) = \Phi_0(x), \tag{2.4}$$

then the function Φ defined by

$$\Phi(t,x) = e^{-\gamma t}\varphi(e^t - 1, e^{-\gamma t}x), \qquad (t,x) \in (0,\infty)^2, \tag{2.5}$$

is a weak solution to (1.15), (1.16).

We now consider the existence part of Theorem 2.1 and show that there exists a weak solution to (2.3), (2.4). It then implies the existence of a weak solution to (1.15), (1.16) by (2.5).

Remark 2.2. Formally, if Φ is the solution to (1.15), (1.16) given by Theorem 2.1, then $\tilde{f}(t,.) = -\frac{d}{dy}\Phi(t,.)^{-1}$ is a solution to (1.13) with coagulation kernel $a_\delta(y,y_*) = y^\lambda + y_*^\lambda + 2\delta$ and vice versa. A rigorous justification of this fact does not seem to be obvious and prevents us from using [15] to prove the existence part of Theorem 2.1.

2.1. The regularized problem. We first investigate a regularized problem and prove the existence of a solution by the method of characteristics. Let $\varepsilon > 0$, $R > 1$, and $\chi_R \in \mathcal{C}^1(\mathbb{R}_+)$ be such that

$$\chi_R(x) = \begin{cases} x & \text{if} \quad x \le R+1, \\ 2R & \text{if} \quad x \ge 4R, \end{cases}$$

and

$$0 \le \chi_R(x) \le 2R, \qquad 0 \le \chi_R'(x) \le 1 \qquad \text{for} \quad x \ge 0.$$

We consider the following equation:

$$\begin{aligned} \partial_t\varphi(t,x) = {} & x \left(\int_0^x \mathcal{R}_\varepsilon(\varphi(t,x_*))\, dx_* + \delta\,(1+t)^{\lambda\gamma}\chi_{R_0}(x) \right) \partial_x\varphi(t,x) \\ & + \int_x^\infty \left(\mathcal{R}_\varepsilon(\varphi(t,x)) + \mathcal{R}_\varepsilon(\varphi(t,x_*)) + 2\,\delta\,(1+t)^{\lambda\gamma} \right) \varphi(t,x_*)\, dx_*, \end{aligned} \tag{2.6}$$

$$\varphi(0,x) = \Phi_0(x), \tag{2.7}$$

where the initial condition Φ_0 fulfills (2.1) (R_0 is given in (2.1)) and

$$\mathcal{R}_\varepsilon(z) := (\varepsilon + z)^\lambda - \varepsilon^\lambda, \qquad z \in \mathbb{R}_+$$

is a $\mathcal{C}^1$-smooth approximation of $z \mapsto z^\lambda$. In particular, $\mathcal{R}_\varepsilon$ enjoys the following properties:

$$0 \le \mathcal{R}_\varepsilon(z) \le \left(\lambda\,\varepsilon^{\lambda-1}\,z\right) \wedge z^\lambda, \qquad |\mathcal{R}_\varepsilon(z) - \mathcal{R}_\varepsilon(z_*)| \le \lambda\,\varepsilon^{\lambda-1}|z-z_*|, \quad (z,z_*) \in \mathbb{R}_+^2. \tag{2.8}$$

We now establish the existence of a solution to (2.6), (2.7) by a fixed point method. Let M, C_1, L_1, and T_ε be four positive real numbers, the values of which we will specify later. We denote by $\mathcal{H}_\varepsilon$ the set of nonnegative functions $h \in \mathcal{C}([0,T_\varepsilon]; L^1(0,\infty))$ such that, for every $t \in [0,T_\varepsilon]$,

- $h(t,.)$ is nonincreasing with compact support $\operatorname{Supp} h(t,.) \subset [0,R_0]$,
- $\|h(t)\|_\infty \le M$, $\|h(t)\|_1 \le C_1$, and $|h(t,x) - h(t,x_*)| \le L_1|x-x_*|$ for every $(x,x_*) \in \mathbb{R}_+^2$.

For $h \in \mathcal{H}_\varepsilon$, we consider the following transport equation:

$$\partial_t\varphi(t,x) + A_h(t,x)\,\partial_x\varphi(t,x) = B_h(t,x), \tag{2.9}$$

where

$$\begin{aligned} A_h(t,x) &= -x \int_0^x \mathcal{R}_\varepsilon(h(t,x_*))\, dx_* - \delta\, x\,(1+t)^{\lambda\gamma}\chi_{R_0}(x), \\ B_h(t,x) &= \int_x^\infty \left(\mathcal{R}_\varepsilon(h(t,x)) + \mathcal{R}_\varepsilon(h(t,x_*)) + 2\,\delta\,(1+t)^{\lambda\gamma} \right) h(t,x_*)\, dx_*. \end{aligned}$$

Since $h(t,.)$ is nonincreasing for every $t \in [0,T_\varepsilon]$, we have

$$x\,h(t,x) \le \int_0^x h(t,x_*)\,dx_*, \qquad (t,x) \in [0,T_\varepsilon] \times (0,\infty). \tag{2.10}$$

Owing to (2.8) and (2.10), A_h and B_h enjoy the following properties.

LEMMA 2.3. *For $h \in \mathcal{H}_\varepsilon$, A_h is continuous on $[0,T_\varepsilon]\times[0,\infty)$ and, for every $(t,x)\in[0,T_\varepsilon]\times[0,\infty)$, we have*

$$-x\,\lambda\,\varepsilon^{\lambda-1}\,C_1 - 2\,\delta\,R_0\,x\,(1+T_\varepsilon)^{\lambda\gamma} \le\ A_h(t,x)\ \le 0,$$
$$-2\,\lambda\varepsilon^{\lambda-1}C_1 - 6\,\delta\,R_0\,(1+T_\varepsilon)^{\lambda\gamma} \le \partial_x A_h(t,x) \le 0.$$

Moreover, $B_h(t,.)$ is a nonincreasing function with compact support in $[0,R_0]$ for every $t\in[0,T_\varepsilon]$ and

$$0\ \le\ B_h(t,x) \le 2\,\left(M^\lambda + \delta\,(1+T_\varepsilon)^{\lambda\gamma}\right)\,C_1, \qquad (t,x)\in[0,T_\varepsilon]\times[0,\infty).$$

We then construct the characteristic curves associated with (2.9): for $h\in\mathcal{H}_\varepsilon$, $t\in[0,T_\varepsilon]$, and $x\in[0,\infty)$, it follows from the continuity of A_h and the boundedness of $\partial_x A_h$ established in Lemma 2.3 and the Cauchy–Lipschitz theorem that the ordinary differential equation

$$\frac{dX}{ds}(s;t,x) = A_h(s,X(s;t,x)), \tag{2.11}$$
$$X(t;t,x) = x, \tag{2.12}$$

has a unique global solution $X(.;t,x)\in\mathcal{C}^1([0,T_\varepsilon];\mathbb{R})$. Furthermore, Lemma 2.3 warrants that

$$x\,e^{-(\lambda\varepsilon^{\lambda-1}C_1+2\,\delta\,R_0\,(1+T_\varepsilon)^{\lambda\gamma})(s-t)} \le X(s;t,x)\le x \quad \text{for} \quad s\in[t,T_\varepsilon], \tag{2.13}$$
$$x\le X(s;t,x)\le x\,e^{(\lambda\varepsilon^{\lambda-1}C_1+2\,\delta\,R_0\,(1+T_\varepsilon)^{\lambda\gamma})(t-s)} \quad \text{for} \quad s\in[0,t]. \tag{2.14}$$

PROPOSITION 2.4. *Consider $\Phi_0\in W^{1,\infty}(\mathbb{R}_+)$ satisfying* (2.1) *and $h\in\mathcal{H}_\varepsilon$. Setting*

$$\varphi(t,x) = \Phi_0(X(0;t,x)) + \int_0^t B_h(s,X(s;t,x))\,ds, \qquad (t,x)\in[0,T_\varepsilon]\times[0,\infty), \tag{2.15}$$

then φ is the unique weak solution to (2.9) *with initial condition Φ_0. In addition, $\varphi(t,.)$ is nonincreasing and* Supp $\varphi(t,.)\subset[0,R_0]$ *for each $t\in[0,T_\varepsilon]$. Moreover, if $\varepsilon\in(0,\lambda^\gamma)$, $M=1+2\|\Phi_0\|_\infty$, $C_1=2\|\Phi_0\|_1$, $L_1=4(\|\partial_x\Phi_0\|_\infty+M+(M^2/C_1))$, and*

$$T_\varepsilon = \min\left(2^{1/(\lambda\gamma)}-1,\ \frac{1}{8\,\lambda\,\varepsilon^{\lambda-1}\,\|\Phi_0\|_1 + 24\,\delta\,R_0}\right),$$

then $\varphi\in\mathcal{H}_\varepsilon$.

Proof. The first assertion of Proposition 2.4 is classical and the compactness of the support of $\varphi(t,.)$ follows from that of Φ_0 and B_h (see Lemma 2.3). We next investigate the behavior of the L^∞- and L^1-norms of φ. We deduce from Lemma 2.3 that

$$0\le\varphi(t,x)\le\|\Phi_0\|_\infty + 2\left(M^\lambda+\delta\,(1+T_\varepsilon)^{\lambda\gamma}\right)C_1\;T_\varepsilon \tag{2.16}$$

for every $(t,x)\in[0,T_\varepsilon]\times[0,\infty)$. Next, the change of variables $x_*=X(s;t,x)$ is a diffeomorphism for every $(s,t)\in[0,T_\varepsilon]^2$ and we have $x=X(t;s,x_*)$ with

$$\partial_x X(t;s,x_*) = \exp\left(\int_s^t \partial_x A_h(\sigma,X(\sigma;s,x_*))\,d\sigma\right). \tag{2.17}$$

Consequently, we deduce from (2.13), (2.14), and (2.15) that

$$\begin{aligned}\int_0^\infty \varphi(t,x)\,dx &= \int_0^\infty \Phi_0(x_*) \exp\left(\int_0^t \partial_x A_h(\sigma, X(\sigma;0,x_*))\,d\sigma\right)\,dx_* \\ &\quad + \int_0^t \int_0^\infty B_h(s,x_*) \exp\left(\int_s^t \partial_x A_h(\sigma, X(\sigma;s,x_*))\,d\sigma\right)\,dx_*\,ds.\end{aligned}$$

The nonpositivity of $\partial_x A_h$ (see Lemma 2.3) implies that

$$\int_0^\infty \varphi(t,x)\,dx \le \int_0^\infty \Phi_0(x_*)\,dx_* + \int_0^t \int_0^\infty B_h(s,x_*)\,dx_*\,ds.$$

But, by (2.8), we get the following equation by the monotonicity and the compactness of the support of $h(s,.)$ for $s \in [0,T_\varepsilon]$:

$$\begin{aligned}\int_0^t \int_0^\infty B_h(s,x_*)\,dx_*\,ds &\le 2\,\lambda\,\varepsilon^{\lambda-1} \int_0^t \int_0^\infty h(s,x_*) \int_{x_*}^\infty h(s,x)\,dx\,dx_*\,ds \\ &\quad + 2\,\delta(1+T_\varepsilon)^{\lambda\gamma} \int_0^t \int_0^{R_0} \int_{x_*}^{R_0} h(s,x)\,dx\,dx_*\,ds \\ &\le \lambda\,\varepsilon^{\lambda-1}\,T_\varepsilon\,C_1^2 + 2\,\delta\,R_0\,C_1\,T_\varepsilon\,(1+T_\varepsilon)^{\lambda\gamma}.\end{aligned}$$

Thus,

$$\|\varphi(t)\|_1 \le \|\Phi_0\|_1 + 2\,C_1\,T_\varepsilon\left(\lambda\,\varepsilon^{\lambda-1}\,C_1 + \delta\,R_0\,(1+T_\varepsilon)^{\lambda\gamma}\right), \qquad t \in [0,T_\varepsilon]. \tag{2.18}$$

We next turn to the Lipschitz property and put $L_0 := \|\partial_x \Phi_0\|_\infty$. For every $(x,x_*) \in (0,\infty)^2$, we have

$$\begin{aligned}|\varphi(t,x) - \varphi(t,x_*)| &\le |\Phi_0(X(0;t,x)) - \Phi_0(X(0;t,x_*))| \\ &\quad + \int_0^t |B_h(s,X(s;t,x)) - B_h(s,X(s;t,x_*))|\,ds.\end{aligned}$$

But, by (2.8),

$$\begin{aligned}&|B_h(s,X(s;t,x)) - B_h(s,X(s;t,x_*))| \\ &\quad \le 2\,M\left(\lambda\,\varepsilon^{\lambda-1} M + \delta\,(1+T_\varepsilon)^{\lambda\gamma}\right)|X(s;t,x) - X(s;t,x_*)| \\ &\quad + \lambda\,\varepsilon^{\lambda-1}\,C_1\,|h(s,X(s;t,x)) - h(s,X(s;t,x_*))|.\end{aligned}$$

Consequently,

$$\begin{aligned}|\varphi(t,x) - \varphi(t,x_*)| &\le L_0\,|X(0;t,x) - X(0;t,x_*)| \\ &\quad + \left(\lambda\,\varepsilon^{\lambda-1}(L_1\,C_1 + 2M^2) + 2\,\delta\,M\,(1+T_\varepsilon)^{\lambda\gamma}\right) \int_0^t |X(s;t,x) - X(s;t,x_*)|\,ds.\end{aligned}$$

We next deduce from (2.17) and Lemma 2.3 that, for $s \in [0,t]$,

$$|X(s;t,x) - X(s;t,x_*)| \le \exp\left(\left(2\,\lambda\,\varepsilon^{\lambda-1} C_1 + 6\,\delta\,R_0\,(1+T_\varepsilon)^{\lambda\gamma}\right)(t-s)\right)|x - x_*|.$$

Consequently,

$$
\begin{aligned}
&|\varphi(t,x)-\varphi(t,x_*)| \\
&\quad \le |x-x_*| \Bigg[\left(L_0 + \frac{M^2}{C_1} + \frac{M}{3\,R_0} \right) \exp\left\{ (2\lambda\,\varepsilon^{\lambda-1}\,C_1 + 6\,\delta\,R_0\,(1+T_\varepsilon)^{\lambda\gamma})\,T_\varepsilon \right\} \\
&\quad + \frac{L_1}{2} \left(\exp\left\{ (2\lambda\,\varepsilon^{\lambda-1}\,C_1 + 6\,\delta\,R_0\,(1+T_\varepsilon)^{\lambda\gamma})\,T_\varepsilon \right\} - 1 \right) \Bigg].
\end{aligned} \tag{2.19}
$$

Moreover, (2.17) implies that the function $x \mapsto X(s;t,x)$ is nondecreasing for every $(s,t) \in [0,T_\varepsilon]^2$. It thus follows from the monotonicity of Φ_0 and $B_h(s,.)$ (see Lemma 2.3) that $x \mapsto \varphi(t,x)$ is a nonincreasing function.

Finally, if $\varepsilon \in (0,\lambda^\gamma)$, we put $M = 1 + 2\|\Phi_0\|_\infty$, $C_1 = 2\|\Phi_0\|_1$,

$$
L_1 = 4\left(L_0 + \frac{M^2}{C_1} + M \right), \qquad T_\varepsilon = \min\left(2^{1/(\lambda\gamma)} - 1,\ \frac{1}{4\,\lambda\,\varepsilon^{\lambda-1}\,C_1 + 24\,\delta\,R_0} \right),
$$

and notice that $C_1 \le R_0 M$. It then readily follows from (2.16), (2.18), and (2.19) that φ belongs to $\mathcal{H}_\varepsilon$. ∎

THEOREM 2.5. *Consider $\varepsilon \in (0,\lambda^\gamma)$ and $\Phi_0 \in W^{1,\infty}(\mathbb{R}_+)$ satisfying* (2.1)*. Then there exists a nonnegative weak solution $\varphi^\varepsilon \in \mathcal{C}([0,\infty);L^1(0,\infty))$ to* (2.6)*,* (2.7) *such that $\varphi^\varepsilon(t,.)$ is a nonincreasing function for every $t \ge 0$ and φ^ε satisfies*

$$
\mathrm{Supp}\ \varphi^\varepsilon(t,.) \subset [0,R_0], \qquad \|\varphi^\varepsilon(t)\|_1 = \|\Phi_0\|_1, \qquad t \in (0,\infty).
$$

To prove the existence of a weak solution to (2.6), (2.7), we need a preliminary lemma.

LEMMA 2.6. *Consider $\varepsilon \in (0,\lambda^\gamma)$ and $\Phi_0 \in W^{1,\infty}(\mathbb{R}_+)$ satisfying* (2.1)*. Assume that the parameters M, C_1, L_1, and T_ε are given by Proposition* 2.4*. Let h_1 and h_2 be two functions in $\mathcal{H}_\varepsilon$ and denote by X_1 and X_2 the associated characteristic curves. Setting $z := X_1 - X_2$, we have*

$$
|z(s;t,x)| \le \lambda\,\varepsilon^{\lambda-1}\,T_\varepsilon\,\|h_1-h_2\|_{\mathcal{C}([0,T_\varepsilon];L^1(0,\infty))}\,X_i(s;t,x)\,\exp\left(2\,T_\varepsilon\,(\lambda\,\varepsilon^{\lambda-1}\,C_1+6\,\delta\,R_0)\right), \tag{2.20}
$$

for $0 \le s \le t \le T_\varepsilon$, $x \in [0,\infty)$ and $i = 1,2$.

Proof. By (2.11), for $s \le t$, we have

$$
\begin{aligned}
|z(s;t,x)| \le {} & \int_s^t |A_{h_1}(\sigma,X_1(\sigma;t,x)) - A_{h_1}(\sigma,X_2(\sigma;t,x))|\,d\sigma \\
& + \int_s^t |(A_{h_1}-A_{h_2})(\sigma,X_2(\sigma;t,x))|\,d\sigma.
\end{aligned}
$$

Since $(1+T_\varepsilon)^{\lambda\gamma} \le 2$, it follows from Lemma 2.3 that

$$
\begin{aligned}
|z(s;t,x)| \le {} & 2\,(\lambda\,\varepsilon^{\lambda-1}\,C_1 + 6\,\delta\,R_0) \int_s^t |z(\sigma;t,x)|\,d\sigma \\
& + \int_s^t X_2(\sigma;t,x) \int_0^{X_2(\sigma;t,x)} |\mathcal{R}_\varepsilon(h_1(\sigma,x_*)) - \mathcal{R}_\varepsilon(h_2(\sigma,x_*))|\,dx_*\,d\sigma.
\end{aligned}
$$

For $s \le t$, (2.8) thus leads to

$$|z(s;t,x)| \le 2\,(\lambda\,\varepsilon^{\lambda-1}\,C_1 + 6\,\delta\,R_0) \int_s^t |z(\sigma;t,x)|\,d\sigma \\ + \lambda\,\varepsilon^{\lambda-1}\,\|h_1 - h_2\|_{\mathcal{C}([0,T_\varepsilon];L^1(0,\infty))} \int_s^t X_2(\sigma;t,x)\,d\sigma.$$

The Gronwall lemma and the monotonicity of $X_2(.;t,x)$ then imply that, for $s \le t$, (2.20) holds for $i=2$. By symmetry of X_1 and X_2, we infer that (2.20) also holds for $i=1$. $\square$

Proof of Theorem 2.5. Let M, C_1, L_1, and T_ε be the four parameters given by Proposition 2.4. We consider the map $\mathcal{T} : \mathcal{H}_\varepsilon \longrightarrow \mathcal{H}_\varepsilon$ defined by $\mathcal{T}(h) = \varphi$, where φ is given by (2.15). Let us check that $\mathcal{T}$ is continuous and compact for the topology of $\mathcal{C}([0,T_\varepsilon];L^1(0,\infty))$.

Continuity of $\mathcal{T}$

Let $(h_n)_{n\ge1}$ be a sequence in $\mathcal{H}_\varepsilon$ that converges to $h \in \mathcal{H}_\varepsilon$. Let X_n and X be the characteristic curves defined by (2.11), (2.12) associated to A_{h_n} and A_h, respectively. We set $z_n = X_n - X$, $\varphi_n = \mathcal{T}(h_n)$ and $\varphi = \mathcal{T}(h)$ for $n \ge 1$. Then,

$$\|\varphi_n(t) - \varphi(t)\|_1 \le \int_0^\infty |\Phi_0(X_n(0;t,x)) - \Phi_0(X(0;t,x))|\,dx \quad (=: J_1(t)) \\ + \int_0^\infty\!\!\int_0^t |B_{h_n}(s,X_n(s;t,x)) - B_h(s,X_n(s;t,x))|\,ds\,dx \quad (=: J_2(t)) \\ + \int_0^\infty\!\!\int_0^t |B_h(s,X_n(s;t,x)) - B_h(s,X(s;t,x))|\,ds\,dx \quad (=: J_3(t)).$$

Owing to the compactness of the support of Φ_0, we have

$$J_1(t) = \int_0^{R_0} |\Phi_0(X_n(0;t,x)) - \Phi_0(X(0;t,x))|\,dx \le \|\partial_x\Phi_0\|_\infty \int_0^{R_0} |z_n(0;t,x)|\,dx,$$

and we deduce from (2.14) and (2.20) that

$$J_1(t) \le \frac{1}{2}\,\lambda\,\varepsilon^{\lambda-1}\,T_\varepsilon\,\|\partial_x\Phi_0\|_\infty\,R_0^2\,\exp\left(T_\varepsilon\,(3\,\lambda\,\varepsilon^{\lambda-1}\,C_1 + 16\,\delta\,R_0)\right)\|h_n - h\|_{\mathcal{C}([0,T_\varepsilon];L^1(0,\infty))}. \tag{2.21}$$

Let us now consider $J_2(t)$. A change of variables yields

$$J_2(t) = \int_0^t\int_0^\infty |B_{h_n}(s,x_*) - B_h(s,x_*)|\ \partial_x X_n(t;s,x_*)\ dx_* ds \\ \le \int_0^t\int_0^\infty |B_{h_n}(s,x) - B_h(s,x)|\ dx ds,$$

the last inequality being a consequence of the nonpositivity of $\partial_x A_{h_n}$ and (2.17). Since both $h(s,.)$ and $h_n(s,.)$ are nonincreasing and compactly supported for $s \in [0,T_\varepsilon]$, we obtain

$$J_2(t) \le \int_0^t\int_0^\infty |\mathcal{R}_\varepsilon(h_n(s,x)) - \mathcal{R}_\varepsilon(h(s,x))| \int_x^\infty h_n(s,x_*)\,dx_*\,dx\,ds \\ + 2\int_0^t\int_0^{R_0} \left(\mathcal{R}_\varepsilon(h(s,x)) + \delta\,(1+s)^{\lambda\gamma}\right) \int_x^\infty |(h_n - h)(s,x_*)|\ dx_*\,dx\,ds \\ + \int_0^t\int_0^\infty h_n(s,x) \int_x^\infty |\mathcal{R}_\varepsilon(h_n(s,x_*)) - \mathcal{R}_\varepsilon(h(s,x_*))|\,dx_*\,dx\,ds.$$

By (2.8), we get

$$\begin{aligned} J_2(t) &\le \lambda\,\varepsilon^{\lambda-1}\,C_1 \int_0^t\int_0^\infty |(h_n-h)(s,x)|\,dx\,ds \\ &\quad + 2\,\delta\,R_0\,T_\varepsilon\,(1+T_\varepsilon)^{\lambda\gamma}\,\|h_n-h\|_{\mathcal{C}([0,T_\varepsilon];L^1(0,\infty))} \\ &\quad + \lambda\,\varepsilon^{\lambda-1}\,\|h_n-h\|_{\mathcal{C}([0,T_\varepsilon];L^1(0,\infty))} \int_0^t\int_0^{R_0} (2h(s,x)+h_n(s,x))\,dx\,ds, \end{aligned}$$

from whence

$$J_2(t) \le 4\,T_\varepsilon\big(\lambda\,\varepsilon^{\lambda-1}\,C_1 + \delta\,R_0\big)\,\|h_n-h\|_{\mathcal{C}([0,T_\varepsilon];L^1(0,\infty))}. \tag{2.22}$$

It remains to handle $J_3(t)$. By the compactness of the support of $h(s,.)$ for $s\in[0,T_\varepsilon]$, we have

$$\begin{aligned} &J_3(t) \\ &\le \int_0^t\int_0^{R_0} |\mathcal{R}_\varepsilon(h(s,X_n(s;t,x))) - \mathcal{R}_\varepsilon(h(s,X(s;t,x)))| \int_{X_n(s;t,x)}^\infty h(s,x_*)\,dx_*\,dx\,ds \\ &+ \int_0^t\int_0^{R_0} \big(\mathcal{R}_\varepsilon(h(s,X(s;t,x))) + 2\,\delta\,(1+s)^{\lambda\gamma}\big) \left| \int_{X_n(s;t,x)}^{X(s;t,x)} h(s,x_*)\,dx_* \right| dx\,ds \\ &+ \int_0^t\int_0^{R_0} \left| \int_{X_n(s;t,x)}^{X(s;t,x)} h(s,x_*)\,\mathcal{R}_\varepsilon(h(s,x_*))\,dx_* \right| dx\,ds. \end{aligned}$$

Thanks to (2.8), we obtain

$$\begin{aligned} J_3(t) &\le \lambda\,\varepsilon^{\lambda-1}\,C_1 \int_0^t\int_0^{R_0} |h(s,X_n(s;t,x)) - h(s,X(s;t,x))|\,dx\,ds \\ &\quad + 2\,M\,\big(\lambda\,\varepsilon^{\lambda-1}\,M + \delta\,(1+T_\varepsilon)^{\lambda\gamma}\big) \int_0^t\int_0^{R_0} |z_n(s;t,x)|\,dx\,ds. \end{aligned}$$

The Lipschitz continuity of h, (2.14) and (2.20) then imply that

$$\begin{aligned} J_3(t) &\le \left(\lambda\,\varepsilon^{\lambda-1}\,M^2 + 2\,\delta\,M + \frac{1}{2}\,\lambda\,\varepsilon^{\lambda-1}\,C_1\,L_1\right)\lambda\,\varepsilon^{\lambda-1}\,T_\varepsilon^2\,R_0^2 \\ &\quad \times \exp\big(T_\varepsilon\,(3\,\lambda\,\varepsilon^{\lambda-1}\,C_1 + 16\,\delta\,R_0)\big)\,\|h_n-h\|_{\mathcal{C}([0,T_\varepsilon];L^1(0,\infty))}. \end{aligned} \tag{2.23}$$

Finally, since

$$\|\varphi_n(t)-\varphi(t)\|_1 \le J_1(t)+J_2(t)+J_3(t),$$

we infer from (2.21), (2.22), and (2.23) that $\mathcal{T}$ is continuous.

COMPACTNESS OF $\mathcal{T}$

Let $(h_n)_{n\ge1}$ be a sequence in $\mathcal{H}_\varepsilon$ and put $\varphi_n = \mathcal{T}(h_n)$ for $n\ge1$. On one hand, since φ_n belongs to $\mathcal{H}_\varepsilon$ for each $n\ge1$, the sequence $(\varphi_n)_{n\ge1}$ is bounded in $L^\infty(0,T_\varepsilon;W^{1,\infty}(0,R_0))$. On the other hand, we have $\partial_t\varphi_n = -A_{h_n}\,\partial_x\varphi_n + B_{h_n}$ by Lemma 2.3 and Proposition 2.4, from which we readily conclude that $(\partial_t\varphi_n)_{n\ge1}$ is bounded in $L^\infty((0,T_\varepsilon)\times(0,R_0))$. By the Arzelà–Ascoli theorem, $(\varphi_n)_{n\ge1}$ is then relatively compact in $\mathcal{C}([0,T_\varepsilon]\times[0,R_0])$, from whence $\mathcal{C}([0,T_\varepsilon];L^1(0,\infty))$ thanks to the compactness of the support of $\varphi_n(t,.)$ for each $t\in[0,T_\varepsilon]$ and $n\ge1$.

Therefore, $\mathcal{H}_\varepsilon$ is a nonempty, convex, closed, and bounded subset of $\mathcal{C}([0,T_\varepsilon];L^1(0,\infty))$, and $\mathcal{T}$ is a compact and continuous map from $\mathcal{H}_\varepsilon$ into $\mathcal{H}_\varepsilon$. The Schauder fixed point theorem ensures the existence of a fixed point of $\mathcal{T}$ that is a weak solution $\varphi^{\varepsilon,1}\in\mathcal{C}([0,T_\varepsilon];L^1(0,\infty))$ to (2.6), (2.7) such that $\varphi^{\varepsilon,1}(t,.)$ is nonnegative and nonincreasing with compact support in $[0,R_0]$ for each $t\in[0,T_\varepsilon]$.

Now, since Supp $\varphi^{\varepsilon,1}(t,.)\subset[0,R_0]$ for $t\in[0,T_\varepsilon]$ and $\chi_{R_0}(x)=x$ for $x\le R_0+1$, we deduce from (2.6) that

$$\frac{d}{dt}\int_0^\infty \varphi^{\varepsilon,1}(t,x)\,dx=0,$$

from whence $\|\varphi^{\varepsilon,1}(t)\|_1=\|\Phi_0\|_1$ for $t\in[0,T_\varepsilon]$. Observing that T_ε depends only on R_0 and $\|\Phi_0\|_1$, we may thus proceed as before with $\varphi^{\varepsilon,1}(T_\varepsilon)$ instead of Φ_0 and deduce the existence of a solution $\varphi^{\varepsilon,2}\in\mathcal{C}([0,T_\varepsilon];L^1(0,\infty))$ to (2.6) with initial condition $\varphi^{\varepsilon,1}(T_\varepsilon)$. Repeating this argument yields the existence of a solution $\varphi^\varepsilon\in\mathcal{C}([0,\infty);L^1(0,\infty))$ to (2.6), (2.7) that satisfies the desired properties. ∎

The next task is to pass to the limit as $\varepsilon\to0$. For that purpose, we need the following estimates.

PROPOSITION 2.7. *Let $\varepsilon\in(0,\lambda^\gamma)$ and $\Phi_0\in W^{1,\infty}(0,\infty)$ satisfy* (2.1). *The weak solution φ^ε to* (2.6), (2.7) *given by Theorem* 2.5 *is nonnegative, $\varphi^\varepsilon(t,.)$ is nonincreasing with compact support in* $[0,R_0]$, *and*

$$\|\varphi^\varepsilon(t)\|_1=\|\Phi_0\|_1, \tag{2.24}$$

$$\|\varphi^\varepsilon(t)\|_\infty\le(1+\|\Phi_0\|_\infty)\,e^{2\|\Phi_0\|_1t\,(1+\delta\,(1+t)^{\lambda\gamma})}, \tag{2.25}$$

for every $t\ge0$.

Proof. All statements of Proposition 2.7 are actually a consequence of Theorem 2.5, except the L^∞-bound which we establish now. Let $p>2$. Multiplying (2.6) by $p\,\varphi^\varepsilon(t,x)^{p-1}$ and recalling that Supp $\varphi^\varepsilon(t,.)\subset[0,R_0]$ for $t\ge0$, we get

$$\begin{aligned}
&\frac{d}{dt}\int_0^\infty \varphi^\varepsilon(t,x)^p\,dx\\
&\quad= -\int_0^\infty \varphi^\varepsilon(t,x)^p\left(\int_0^x \mathcal{R}_\varepsilon(\varphi^\varepsilon(t,x_*))\,dx_*+\delta\,(1+t)^{\lambda\gamma}x\right)dx\\
&\quad- \int_0^\infty \varphi^\varepsilon(t,x)^p\left(x\,\mathcal{R}_\varepsilon(\varphi^\varepsilon(t,x))+\delta\,(1+t)^{\lambda\gamma}\,x\right)\,dx\\
&\quad+\ p\int_0^\infty \varphi^\varepsilon(t,x)\int_0^x \varphi^\varepsilon(t,x_*)^{p-1}\,\mathcal{R}_\varepsilon(\varphi^\varepsilon(t,x_*))\,dx_*\,dx\\
&\quad+\ p\int_0^\infty \left(\mathcal{R}_\varepsilon(\varphi^\varepsilon(t,x))+2\,\delta\,(1+t)^{\lambda\gamma}\right)\,\varphi^\varepsilon(t,x)\int_0^x \varphi^\varepsilon(t,x_*)^{p-1}\,dx_*\,dx.
\end{aligned}$$

By (2.8), it follows from the nonnegativity and the monotonicity of φ^ε that, for

$t \in [0,T]$,

$$\begin{aligned}
&\frac{d}{dt}\int_0^\infty \varphi^\varepsilon(t,x)^p\,dx \\
&\quad\le 2p\int_0^\infty \varphi^\varepsilon(t,x)\int_0^x \left(\varphi^\varepsilon(t,x_*)^{p+\lambda-1} + \delta\,(1+t)^{\lambda\gamma}\,\varphi^\varepsilon(t,x_*)^{p-1}\right)\,dx_*\,dx \\
&\quad\le 2p\,\|\Phi_0\|_1\left[\int_0^\infty \varphi^\varepsilon(t,x_*)^{p(p+\lambda-2)/(p-1)}\,\varphi^\varepsilon(t,x_*)^{(1-\lambda)/(p-1)}\,dx_*\right. \\
&\quad\left. +\,\delta\,(1+t)^{\lambda\gamma}\int_0^\infty \varphi^\varepsilon(t,x_*)^{p(p-2)/(p-1)}\,\varphi^\varepsilon(t,x_*)^{1/(p-1)}\,dx_*\right].
\end{aligned}$$

Since $p>2$, the Young inequality implies that

$$\frac{d}{dt}\int_0^\infty \varphi^\varepsilon(t,x)^p\,dx \le 2p\,\|\Phi_0\|_1(1+\delta\,(1+t)^{\lambda\gamma})\left[\int_0^\infty \varphi^\varepsilon(t,x)^p\,dx + \|\Phi_0\|_1\right].$$

Thus, we infer from the Gronwall lemma that

$$\|\varphi^\varepsilon(t)\|_p^p \;\le\; (\|\Phi_0\|_p^p + \|\Phi_0\|_1)\,e^{2p\|\Phi_0\|_1 t\,(1+\delta\,(1+t)^{\lambda\gamma})}.$$

Hence,

$$\|\varphi^\varepsilon(t)\|_p \;\le\; (\|\Phi_0\|_p + \|\Phi_0\|_1^{1/p})\,e^{2\|\Phi_0\|_1 t\,(1+\delta\,(1+t)^{\lambda\gamma})}.$$

Letting $p \longrightarrow \infty$ leads to (2.25). □

2.2. Proof of Theorem 2.1. The existence part of Theorem 2.1 is a straightforward consequence of (2.5) and the following proposition.

PROPOSITION 2.8. *Let $\Phi_0 \in L^\infty(0,\infty)$ satisfy* (2.1)*. Then, there exists a weak solution $\varphi \in \mathcal{C}([0,\infty);L^1(0,\infty))$ to* (2.3), (2.4) *such that $\varphi(t,.)$ is nonnegative and nonincreasing with compact support in* $[0,R_0]$,

$$\|\varphi(t)\|_1 = \|\Phi_0\|_1 \qquad \textit{and} \qquad \sup_{0\le s\le t}\|\varphi(s)\|_\infty < \infty,$$

for every $t \ge 0$.

Proof. We fix $k_0 \ge 1$ such that $k_0 > \lambda^{-\gamma}$ and $T>0$. Let $(\Phi_0^k)_{k\ge k_0}$ be a sequence of functions from $W^{1,\infty}(\mathbb{R}_+)$ such that Φ_0^k is nonnegative and nonincreasing with compact support in $[0,R_0]$, $\Phi_0^k(x) \le 2\Phi_0(x)$ a.e., and $(\Phi_0^k)_{k\ge k_0}$ converges towards Φ_0 in $L^1(\mathbb{R}_+)$.

For all $k \ge k_0$, we set $\varepsilon_k = 1/k < \lambda^\gamma$ and denote by $\varphi^k = \varphi^{\varepsilon_k}$ the solution to (2.6) with initial condition Φ_0^k given by Theorem 2.5. It follows from (2.6) and Proposition 2.7 that

$$\begin{aligned}
&(\varphi^k)_{k\ge k_0}\text{is bounded in } L^\infty((0,T)\times(0,R_0)) \cap L^\infty(0,T;BV(0,R_0)), \\
&(\partial_t\varphi^k)_{k\ge k_0}\text{is bounded in } L^\infty(0,T;W^{1,1}(0,R_0)').
\end{aligned}$$

Since $L^\infty(0,R_0)\cap BV(0,R_0)$ is compactly embedded in $L^1(0,R_0)$ [12] and $L^1(0,R_0)$ is continuously embedded in $W^{1,1}(0,R_0)'$, we infer from [23, Corollary 4] that $(\varphi^k)_{k\ge k_0}$ is relatively compact in $\mathcal{C}([0,T];L^1(0,R_0))$, from whence in $\mathcal{C}([0,T];L^1(0,\infty))$ since φ^k identically vanishes in $(0,\infty)\times(R_0,\infty)$ for each $k \ge k_0$. Consequently, there exists

$\varphi \in \mathcal{C}([0,T];L^1(0,\infty))$ such that, up to an extraction, (φ^k) converges towards φ in $\mathcal{C}([0,T];L^1(0,\infty))$ and a.e. on $[0,T]\times(0,\infty)$.

Let $\vartheta \in \mathcal{C}_0^\infty((0,\infty))$. Then, recalling that $\mathrm{Supp}\,\varphi^k(t,.) \subset [0,R_0]$ and $\chi_{R_0}(x)=x$ for $x \le R_0+1$, we deduce from (2.6) that, for every $t\in[0,T]$,

$$
\begin{aligned}
&\int_0^\infty \varphi^k(t,x)\,\vartheta(x)\,dx - \int_0^\infty \Phi_0^k(x)\,\vartheta(x)\,dx \\
&= -\int_0^t\int_0^\infty \varphi^k(s,x)\,\partial_x\vartheta(x)\,x\left(\int_0^x \mathcal{R}_{\varepsilon_k}(\varphi^k(s,x_*))\,dx_* + \delta\,x\,(1+s)^{\lambda\gamma}\right)dx\,ds \\
&- \int_0^t\int_0^\infty \varphi^k(s,x)\,\vartheta(x)\left(\int_0^x \mathcal{R}_{\varepsilon_k}(\varphi^k(s,x_*))\,dx_* + x\,\mathcal{R}_{\varepsilon_k}(\varphi^k(s,x)) + 2\,\delta\,x\,(1+s)^{\lambda\gamma}\right)dx\,ds \\
&+ \int_0^t\int_0^\infty \vartheta(x)\int_x^\infty \varphi^k(s,x_*)\,\left(\mathcal{R}_{\varepsilon_k}(\varphi^k(s,x)) + \mathcal{R}_{\varepsilon_k}(\varphi^k(s,x_*)) + 2\,\delta\,(1+s)^{\lambda\gamma}\right)\,dx_*\,dx\,ds.
\end{aligned}
$$

Thanks to the dominated convergence theorem, we let $k\to\infty$ and get that φ is a weak solution to (2.3), (2.4). The properties satisfied by φ follow easily from Proposition 2.7. $\square$

The uniqueness assertion of Theorem 2.1 is actually a consequence of the following result.

Proposition 2.9. *Consider two functions Φ_0 and $\hat{\Phi}_0$ fulfilling the assumptions* (2.1). *If Φ and $\hat{\Phi}$ are weak solutions to* (1.15), (1.16) *with initial data Φ_0 and $\hat{\Phi}_0$, respectively, and $T>0$, then there exists $C(T)$ depending only on λ, $\|\Phi_0\|_\infty$, $\|\hat{\Phi}_0\|_\infty$, R_0, and T such that*

$$\left\|\Phi(t)-\hat{\Phi}(t)\right\|_1 \le C(T)\ \left\|\Phi_0-\hat{\Phi}_0\right\|_1 \quad \textit{for} \quad t\in[0,T]. \tag{2.26}$$

Proof. Let $T>0$. By Proposition 2.8 and (2.5), the support of $\Phi(t,.)$ and $\hat{\Phi}(t,.)$ is contained in $[0,R_0\ e^{\gamma t}]$ for each $t\in[0,T]$ and both Φ and $\hat{\Phi}$ belong to $L^\infty((0,T)\times(0,\infty))$. Consequently, we have

$$\Lambda := \sup_{t\in[0,T]}\left\{\left\|\Phi(t)^\lambda\right\|_1 \vee \left\|\hat{\Phi}(t)^\lambda\right\|_1\right\} \le R_0\ e^{\gamma T}\ \sup_{t\in[0,T]}\left\{\|\Phi(t)\|_\infty^\lambda \vee \left\|\hat{\Phi}(t)\right\|_\infty^\lambda\right\} < \infty,$$

and notice that the monotonicity of Φ and $\hat{\Phi}$ imply that

$$x\ \Phi(t,x)^\lambda \le \int_0^x \Phi(t,x_*)^\lambda\ dx_* \le \Lambda \quad\text{and}\quad x\ \hat{\Phi}(t,x)^\lambda \le \int_0^x \hat{\Phi}(t,x_*)^\lambda\ dx_* \le \Lambda \tag{2.27}$$

for $x>0$. We put $E:=\Phi-\hat{\Phi}$ and $\sigma = \ \mathrm{sign}(E)$ and give only a formal proof of (2.26) below as both Φ and $\hat{\Phi}$ do not have the required smoothness to justify the forthcoming computations. Nevertheless, a rigorous proof can be performed by approximation arguments as in [3]. We infer from (1.15) that

$$
\begin{aligned}
\frac{d}{dt}\|E(t)\|_1 = & -\gamma \int_0^\infty \left(x\ \partial_x|E(t,x)| + |E(t,x)|\right)\ dx \\
& + \delta \int_0^\infty x^2\ \partial_x|E(t,x)|\ dx + 2\delta \int_0^\infty \sigma(t,x) \int_x^\infty E(t,x_*)\ dx_* dx \\
& + \frac{1}{2} \int_0^\infty x \left(\int_0^x \left(\Phi^\lambda + \hat{\Phi}^\lambda \right)(t,x_*)\ dx_* \right)\ \partial_x|E(t,x)|\ dx \\
& + \frac{1}{2} \int_0^\infty x\ \sigma(t,x) \left(\int_0^x \left(\Phi^\lambda - \hat{\Phi}^\lambda \right)(t,x_*)\ dx_* \right)\ \partial_x \left(\Phi + \hat{\Phi} \right)(t,x)\ dx \\
& + \int_0^\infty \sigma(t,x) \int_x^\infty \left(\Phi(t,x)^\lambda\ \Phi(t,x_*) - \hat{\Phi}(t,x)^\lambda\ \hat{\Phi}(t,x_*) \right)\ dx_* dx \\
& + \int_0^\infty \sigma(t,x) \int_x^\infty \left(\Phi^{1+\lambda} - \hat{\Phi}^{1+\lambda} \right)(t,x_*)\ dx_* dx.
\end{aligned}
$$

Integrating by parts the first, third, and fifth terms of the right-hand side of the above equality and using the fact that $|\sigma(t,x)| \le 1$, we obtain

$$
\begin{aligned}
\frac{d}{dt}\|E(t)\|_1 \le & \left[(-\gamma\ x + \delta\ x^2)\ |E(t,x)| \right]_0^\infty \\
& + \frac{1}{2} \left[x \left(\int_0^x \left(\Phi(t,x_*)^\lambda + \hat{\Phi}(t,x_*)^\lambda \right)\ dx_* \right)\ |E(t,x)| \right]_0^\infty \\
& + 2\delta \int_0^\infty \int_x^\infty |E(t,x_*)|\ dx_* dx - 2\delta \int_0^\infty x\ |E(t,x)|\ dx \quad (=: -I_1(t)) \\
& - \frac{1}{2} \int_0^\infty \int_0^x \left(\Phi^\lambda + \hat{\Phi}^\lambda \right)(t,x_*)\ dx_*\ |E(t,x)|\ dx \quad (=: -I_2(t)) \\
& - \frac{1}{2} \int_0^\infty x \left(\Phi^\lambda + \hat{\Phi}^\lambda \right)(t,x)\ |E(t,x)|\ dx \quad (=: -I_3(t)) \\
& + \frac{1}{2} \int_0^\infty x \left(\int_0^x \left| \Phi^\lambda - \hat{\Phi}^\lambda \right|(t,x_*)\ dx_* \right) \left| \partial_x \left(\Phi + \hat{\Phi} \right)(t,x) \right|\ dx \quad (=: I_4(t)) \\
& + \int_0^\infty \int_x^\infty \left| \Phi(t,x)^\lambda\ \Phi(t,x_*) - \hat{\Phi}(t,x)^\lambda\ \hat{\Phi}(t,x_*) \right|\ dx_* dx \quad (=: I_5(t)) \\
& + \int_0^\infty \int_x^\infty \left| \Phi^{1+\lambda} - \hat{\Phi}^{1+\lambda} \right|(t,x_*)\ dx_* dx. \quad (=: I_6(t)).
\end{aligned}
$$

Owing to the compactness of the support of $\Phi(t,.)$ and $\hat{\Phi}(t,.)$, the boundary terms in the previous inequality vanish. Also, we clearly have $I_1(t) \ge 0$ by the Fubini theorem. Consequently,

$$
\frac{d}{dt}\|E(t)\|_1 \le -I_2(t) - I_3(t) + I_4(t) + I_5(t) + I_6(t). \tag{2.28}
$$

Thanks to the monotonicity of $\Phi(t,.)$ and $\hat{\Phi}(t,.)$ with respect to x, $I_4(t)$ also reads

$$\begin{aligned}
2\ I_4(t) &= -\int_0^\infty x\ \left(\int_0^x \left|\Phi^\lambda - \hat{\Phi}^\lambda\right|(t,x_*)\ dx_*\right)\ \partial_x \left(\Phi+\hat{\Phi}\right)(t,x)\ dx \\
&= -\left[x\ \left(\int_0^x \left|\Phi^\lambda - \hat{\Phi}^\lambda\right|(t,x_*)\ dx_*\right)\ \left(\Phi+\hat{\Phi}\right)(t,x)\right]_0^\infty \\
&+ \int_0^\infty x\ \left|\Phi^\lambda - \hat{\Phi}^\lambda\right|(t,x)\ \left(\Phi+\hat{\Phi}\right)(t,x)\ dx \\
&+ \int_0^\infty \int_0^x \left|\Phi^\lambda - \hat{\Phi}^\lambda\right|(t,x_*)\ dx_*\ \left(\Phi+\hat{\Phi}\right)(t,x)\ dx.
\end{aligned}$$

As in the above computation, the boundary terms vanish. Since

$$(\Phi+\hat{\Phi})(t,x) = 2\ \left(\Phi\wedge\hat{\Phi}\right)(t,x) + |E(t,x)|, \tag{2.29}$$

we infer from the mean value theorem that

$$\begin{aligned}
2\ I_4(t) &\le 2\lambda\ \int_0^\infty x\ \left(\Phi\wedge\hat{\Phi}\right)^\lambda(t,x)\ |E(t,x)|\ dx \\
&+ \int_0^\infty x\ \left|\Phi^\lambda - \hat{\Phi}^\lambda\right|(t,x)\ |E(t,x)|\ dx \\
&+ 2\lambda\ \int_0^\infty \int_0^x \left(\Phi\wedge\hat{\Phi}\right)^{\lambda-1}(t,x_*)\ |E(t,x_*)|\ dx_*\ \left(\Phi\wedge\hat{\Phi}\right)(t,x)\ dx \\
&+ \int_0^\infty \int_0^x \left|\Phi^\lambda - \hat{\Phi}^\lambda\right|(t,x_*)\ dx_*\ |E(t,x)|\ dx.
\end{aligned}$$

Using the monotonicity of $\Phi(t,.)$ and $\hat{\Phi}(t,.)$ with respect to x and (2.27), we further obtain

$$\begin{aligned}
2\ I_4(t) &\le 2\lambda\ \Lambda\ \|E(t)\|_1 + 2\ \Lambda\ \|E(t)\|_1 \\
&+ 2\lambda\ \int_0^\infty \int_0^x \left(\Phi\wedge\hat{\Phi}\right)^{\lambda-1}(t,x)\ |E(t,x_*)|\ dx_*\ \left(\Phi\wedge\hat{\Phi}\right)(t,x)\ dx \\
&+ 2\ \Lambda\ \|E(t)\|_1 \\
&\le 2(\lambda+2)\ \Lambda\ \|E(t)\|_1 + 2\lambda\ \int_0^\infty \int_{x_*}^\infty \left(\Phi\wedge\hat{\Phi}\right)^\lambda(t,x)\ dx\ |E(t,x_*)|\ dx_*
\end{aligned}$$

$$I_4(t) \le 2(\lambda+1)\ \Lambda\ \|E(t)\|_1. \tag{2.30}$$

Next, by (2.29), the Fubini theorem and the mean value theorem, we have

$$\begin{aligned}
I_5(t) \leq & \frac{1}{2} \int_0^\infty \int_x^\infty \left(\Phi^\lambda + \hat{\Phi}^\lambda\right)(t,x) \ |E(t,x_*)| \ dx_* dx \\
& + \frac{1}{2} \int_0^\infty \int_x^\infty \left(\Phi + \hat{\Phi}\right)(t,x_*) \ \left|\Phi^\lambda - \hat{\Phi}^\lambda\right|(t,x) \ dx_* dx \\
\leq & \frac{1}{2} \int_0^\infty \int_0^x \left(\Phi^\lambda + \hat{\Phi}^\lambda\right)(t,x_*) \ dx_* \ |E(t,x)| \ dx \\
& + \lambda \int_0^\infty \int_x^\infty \left(\Phi \wedge \hat{\Phi}\right)(t,x_*) \ \left(\Phi \wedge \hat{\Phi}\right)^{\lambda-1}(t,x) \ |E(t,x)| \ dx_* dx \\
& + \frac{1}{2} \int_0^\infty \int_x^\infty \left|\Phi^\lambda - \hat{\Phi}^\lambda\right|(t,x) \ |E(t,x_*)| \ dx_* dx \\
\leq & I_2(t) + \lambda \int_0^\infty \int_x^\infty \left(\Phi \wedge \hat{\Phi}\right)^\lambda (t,x_*) \ |E(t,x)| \ dx_* dx + \Lambda \ \|E(t)\|_1,
\end{aligned}$$

the last inequality resulting from the monotonicity of $\Phi(t,.)$ and $\hat{\Phi}(t,.)$ with respect to x. We therefore end up with

$$I_5(t) \leq I_2(t) + (\lambda+1) \ \Lambda \ \|E(t)\|_1. \tag{2.31}$$

Next, using once more the monotonicity of $\Phi(t,.)$ and $\hat{\Phi}(t,.)$ with respect to x and the mean value theorem, we obtain

$$\begin{aligned}
I_6(t) & \leq (1+\lambda) \int_0^\infty \int_x^\infty \left(\Phi(t,x_*) \vee \hat{\Phi}(t,x_*)\right)^\lambda \ |E(t,x_*)| \ dx_* dx \\
& \leq (1+\lambda) \int_0^\infty \left(\Phi(t,x) \vee \hat{\Phi}(t,x)\right)^\lambda \int_x^\infty |E(t,x_*)| \ dx_* dx \\
& \leq 2 \ (1+\lambda) \ \Lambda \ \|E(t)\|_1. \qquad (2.32)
\end{aligned}$$

Since I_3 is nonnegative, we infer from (2.28), (2.30), (2.31), and (2.32) that

$$\frac{d}{dt}\|E(t)\|_1 \leq 5 \ (1+\lambda) \ \Lambda \ \|E(t)\|_1$$

for $t \in [0,T]$, from whence (2.26). $\square$

3. Stationary solutions to (1.14). To establish the existence of a steady state Ψ to (1.14) satisfying $\|\Psi\|_1 = 1$, we proceed in two steps and first show that, for each $\delta \in (0,1)$, there is a stationary solution Ψ_δ to (1.15) such that $\|\Psi_\delta\|_1 = 1$. We next prove that the family $(\Psi_\delta)_{\delta\in(0,1)}$ belongs to a compact subset of $L^1(0,\infty)$ and that the cluster points of $(\Psi_\delta)_{\delta\in(0,1)}$ are stationary solutions to (1.14) satisfying the required L^1-constraint.

In order to apply Theorem 1.4 to the semiflow associated to (1.15), (1.16), we have to identify a compact and convex subset of $L^1(0,\infty)$ which is left invariant by the semiflow. We first recall that when Φ_0 satisfies (2.1), then $\Phi(t,.)$ is compactly supported with Supp $\Phi(t,.) \subset [0, R_0 \ e^{\gamma t}]$ and

$$\int_0^\infty \Phi(t,x) \ dx = 1 := \int_0^\infty \Phi_0(x) \ dx \quad \text{for} \quad t \geq 0. \tag{3.1}$$

We next investigate the time evolution of the L^∞-norm.

Lemma 3.1. *Consider $\delta \in (0,1)$ and assume that Φ_0 satisfies* (2.1). *Denoting by Φ the corresponding solution to* (1.15), (1.16), *we have*

$$\|\Phi(t)\|_\infty \le m(t), \quad t \ge 0, \tag{3.2}$$

where m is the solution to

$$\frac{dm}{dt}(t) = 2\left(m(t)^\lambda + \delta\right) - \gamma\, m(t), \quad m(0) = \|\Phi_0\|_\infty. \tag{3.3}$$

Proof. We fix $T > 0$ and recall that $\operatorname{Supp}\Phi(t,.) \subset [0, x_T]$ for $t \in [0,T]$ with $x_T := R_0\, e^{\gamma T}$. For $t \in [0,T]$, it follows from (1.15), (3.1) and the monotonicity and nonnegativity of Φ that

$$\begin{aligned}
&\frac{d}{dt}\int_0^\infty (\Phi(t,x)-m(t))_+\ dx = \frac{d}{dt}\int_0^{x_T} (\Phi(t,x)-m(t))_+\ dx \\
&\quad \le \gamma \int_0^{x_T} \operatorname{sign}_+(\Phi(t,x)-m(t))\ (\Phi(t,x)-m(t)-\Phi(t,x))\ dx \\
&\quad - 2\delta \int_0^{x_T} x\ (\Phi(t,x)-m(t))_+\ dx \\
&\quad + 2\delta \int_0^{x_T} \operatorname{sign}_+(\Phi(t,x)-m(t)) \int_x^\infty \Phi(t,x_*)\ dx_* dx \\
&\quad + \int_0^{x_T} \operatorname{sign}_+(\Phi(t,x)-m(t)) \int_x^\infty \Phi(t,x_*)\ \left(\Phi(t,x)^\lambda + \Phi(t,x_*)^\lambda\right)\ dx_* dx \\
&\quad - \int_0^{x_T} \operatorname{sign}_+(\Phi(t,x)-m(t))\ \frac{dm}{dt}(t)\ dx \\
&\quad \le \int_0^{x_T} \operatorname{sign}_+(\Phi(t,x)-m(t))\ \left(2\delta - \gamma\, m(t) - \frac{dm}{dt}(t)\right)\ dx \\
&\quad + 2\int_0^{x_T} \operatorname{sign}_+(\Phi(t,x)-m(t))\ \Phi(t,x)^\lambda\ dx.
\end{aligned}$$

By (3.3), we have the lower bound $m(t) \ge m(0)\ e^{-\gamma t} \ge m(0)\ e^{-\gamma T}$, and we obtain

$$\begin{aligned}
\frac{d}{dt}\int_0^\infty (\Phi(t,x)-m(t))_+\ dx &\le 2\int_0^{x_T} \operatorname{sign}_+(\Phi(t,x)-m(t))\ \left(\Phi(t,x)^\lambda - m(t)^\lambda\right)\ dx \\
&\le 2\ \lambda\ m(t)^{\lambda-1} \int_0^{x_T} (\Phi(t,x)-m(t))_+\ dx \\
&\le 2\ \lambda\ m(0)^{\lambda-1}\ e^T \int_0^{x_T} (\Phi(t,x)-m(t))_+\ dx \\
&\le 2\ \lambda\ m(0)^{\lambda-1}\ e^T \int_0^\infty (\Phi(t,x)-m(t))_+\ dx.
\end{aligned}$$

Consequently,

$$\int_0^\infty (\Phi(t,x)-m(t))_+\ dx \le C(T) \int_0^\infty (\Phi_0(x)-m(0))_+\ dx = 0,$$

from which the inequality (3.2) readily follows for $t \in [0,T]$. Since T was arbitrarily chosen, we obtain the expected result. ∎

Having excluded the occurrence of large values of Φ throughout time evolution, we next turn to a refined estimate on the propagation of the support of Φ.

LEMMA 3.2. *Consider* $\delta \in (0,1)$ *and assume that* Φ_0 *satisfies* (2.1). *Denoting by* Φ *the corresponding solution to* (1.15), (1.16), *we have*

$$\text{Supp}\ \Phi(t,.) \subset [0,R(t)], \quad t \geq 0, \tag{3.4}$$

where R *is the solution to*

$$\frac{dR}{dt}(t) = \gamma\ R(t) - \delta\ R(t)^2, \quad R(0) = R_0. \tag{3.5}$$

Proof. For $t \in (0,\infty)$, it follows from (1.15) and the Fubini theorem that

$$\begin{aligned}
&\frac{d}{dt}\int_{R(t)}^\infty \Phi(t,x)\ dx \\
&= -\frac{dR}{dt}(t)\ \Phi(t,R(t)) - \gamma\ \Big[x\ \Phi(t,x)\Big]_{R(t)}^\infty \\
&+ \left[x\ \Phi(t,x)\left(\delta\ x + \int_0^x \Phi(t,x_*)^\lambda\ dx_*\right)\right]_{R(t)}^\infty \\
&- \int_{R(t)}^\infty \left(\delta\ x + \int_0^x \Phi(t,x_*)^\lambda\ dx_*\right)\ \Phi(t,x)\ dx \\
&- \int_{R(t)}^\infty x\ \left(\delta + \Phi(t,x)^\lambda\right)\ \Phi(t,x)\ dx \\
&+ \int_{R(t)}^\infty \int_{R(t)}^x \left(\Phi(t,x)^\lambda + \Phi(t,x_*)^\lambda + 2\delta\right)\ \Phi(t,x)\ dx_* dx \\
&= \left(-\frac{dR}{dt}(t) + \gamma\ R(t) - \delta\ R(t)^2 - R(t)\ \int_0^{R(t)} \Phi(t,x_*)^\lambda\ dx_*\right)\Phi(t,R(t)) \\
&- \int_{R(t)}^\infty \int_0^x \left(2\delta + \Phi(t,x)^\lambda + \Phi(t,x_*)^\lambda\right)\ \Phi(t,x)\ dx_* dx \\
&+ \int_{R(t)}^\infty \int_{R(t)}^x \left(\Phi(t,x)^\lambda + \Phi(t,x_*)^\lambda + 2\delta\right)\ \Phi(t,x)\ dx_* dx \\
&\leq 0,
\end{aligned}$$

from which we deduce (3.4) by integration. □

Observe that the estimate on the expansion on the support of Φ obtained in the previous lemma heavily depends on δ and will thus not be useful to pass to the limit as $\delta \to 0$. For that purpose, a control on the behavior of Φ for large x which does not depend on δ is obtained in the next lemma.

LEMMA 3.3. *Consider* $\delta \in (0,1)$ *and assume that* Φ_0 *satisfies* (2.1). *Denoting by* Φ *the corresponding solution to* (1.15), (1.16), *we have*

$$\mathcal{L}(t) := \int_0^\infty x^{(1-\lambda)/\lambda}\ \Phi(t,x)\ dx \leq \ell(t), \quad t \geq 0, \tag{3.6}$$

where

$$\frac{d\ell}{dt}(t) = \frac{1}{\lambda}\ \ell(t) - \frac{1-\lambda}{(1+\lambda)\lambda^{1+\lambda}}\ \ell(t)^{1+\lambda}, \quad \ell(0) = \mathcal{L}(0) = \int_0^\infty x^{(1-\lambda)/\lambda}\ \Phi_0(x)\ dx.$$

Proof. For $t \in (0,\infty)$, it follows from (1.15), the compactness of the support of $\Phi(t,.)$ and the Fubini theorem that

$$\begin{aligned}
\frac{d\mathcal{L}}{dt}(t) &= \frac{\mathcal{L}(t)}{\lambda} - \frac{1}{\lambda} \int_0^\infty x^{(1-\lambda)/\lambda} \left(\delta\ x + \int_0^x \Phi(t,x_*)^\lambda\ dx_*\right)\ \Phi(t,x)\ dx \\
&\quad - \int_0^\infty x^{1/\lambda} \left(\delta + \Phi(t,x)^\lambda\right) \Phi(t,x)\ dx \\
&\quad + \int_0^\infty \int_0^x x_*^{(1-\lambda)/\lambda} \left(2\delta + \Phi(t,x)^\lambda + \Phi(t,x_*)^\lambda\right)\ \Phi(t,x)\ dx_* dx \\
&= \frac{\mathcal{L}(t)}{\lambda} + \left(2\lambda - 1 - \frac{1}{\lambda}\right)\ \delta \int_0^\infty x^{1/\lambda}\ \Phi(t,x)\ dx \\
&\quad + \int_0^\infty \int_0^x \left(x_*^{(1-\lambda)/\lambda} - \frac{x^{(1-\lambda)/\lambda}}{\lambda}\right)\ \Phi(t,x_*)^\lambda\ \Phi(t,x)\ dx_* dx \\
&\quad - (1-\lambda) \int_0^\infty x^{1/\lambda}\ \Phi(t,x)^{1+\lambda}\ dx.
\end{aligned}$$

Since $\lambda \in (0,1)$, we end up with

$$\frac{d\mathcal{L}}{dt}(t) \le \frac{\mathcal{L}(t)}{\lambda} - \frac{1-\lambda}{\lambda} \int_0^\infty x^{(1-\lambda)/\lambda} \int_0^x \Phi(t,x_*)^\lambda\ dx_*\ \Phi(t,x)\ dx.$$

Using Lemma B.1, we further obtain

$$\begin{aligned}
\frac{d\mathcal{L}}{dt}(t) &\le \frac{\mathcal{L}(t)}{\lambda} - \frac{1-\lambda}{\lambda^{1+\lambda}} \int_0^\infty x^{(1-\lambda)/\lambda}\ \Phi(t,x) \left(\int_0^x x_*^{(1-\lambda)/\lambda}\ \Phi(t,x_*)\ dx_*\right)^\lambda\ dx \\
&\le \frac{\mathcal{L}(t)}{\lambda} - \frac{1-\lambda}{(1+\lambda)\lambda^{1+\lambda}}\ \mathcal{L}(t)^{1+\lambda},
\end{aligned}$$

from whence (3.6) by the comparison principle. ∎

We are now in a position to construct stationary solutions to (1.15).

PROPOSITION 3.4. *Given $\delta \in (0,1)$, there exists a nonnegative and nonincreasing function $\Psi_\delta \in L^1(0,\infty) \cap L^\infty(0,\infty)$ such that* $\text{Supp}\ \Psi_\delta \subset [0,\gamma/\delta]$,

$$\|\Psi_\delta\|_1 = 1, \quad \|\Psi_\delta\|_\infty \le A(\lambda) + 2\ \delta, \quad \int_0^\infty x^{(1-\lambda)/\lambda}\ \Psi_\delta(x)\ dx \le B(\lambda), \tag{3.7}$$

and

$$\begin{aligned}
& x \left(\gamma - \delta\ x - \int_0^x \Psi_\delta(x_*)^\lambda\ dx_*\right)\ \Psi_\delta(x) \\
&\qquad = \int_x^\infty \int_0^x \left(\Psi_\delta(x_*)^\lambda + \Psi_\delta(x')^\lambda + 2\delta\right)\ \Psi_\delta(x_*)\ dx' dx_*
\end{aligned} \tag{3.8}$$

for almost every $x \in (0,\gamma/\delta)$. The parameters $A(\lambda)$ and $B(\lambda)$ are given by

$$A(\lambda) := \left(\frac{2}{\gamma}\right)^\gamma, \quad B(\lambda) := \lambda\ \left(\frac{1+\lambda}{1-\lambda}\right)^{1/\lambda}.$$

Proof. Given $\delta \in (0,1)$, we introduce the set $\mathcal{K}_\delta$ defined by

$$\mathcal{K}_\delta := \left\{ \begin{array}{l} U \in L^1(0,\infty) \text{ is a nonnegative and nonincreasing compactly sup-} \\ \text{ported function such that } \text{Supp}\ U \subset [0,\gamma/\delta],\ \|U\|_1 = 1,\ \|U\|_\infty \le \\ z(\delta) \text{ and } \displaystyle\int_0^\infty x^{(1-\lambda)/\lambda}\ U(x)\ dx \le B(\lambda) \end{array} \right\},$$

where $z(\delta)$ is the unique positive zero of $z \longmapsto 2(\delta + z^\lambda) - \gamma z$. We note that

$$z(0) = A(\lambda) \le z(\delta) \le A(\lambda) + 2\delta. \tag{3.9}$$

Then, $\mathcal{K}_\delta$ is a closed convex subset of $L^1(0,\infty)$. In addition, if $(U_n)_{n\ge1}$ is a sequence in $\mathcal{K}_\delta$, then $(U_n)_{n\ge1}$ is bounded in $BV(0,\infty)$ and there is a subsequence of $(U_n)_{n\ge1}$ (not relabeled) and a function U such that $(U_n(x))$ converges towards $U(x)$ for almost every $x \in (0,\infty)$ as $n \to \infty$ [12]. On one hand, this convergence and the Fatou lemma imply that U is a nonnegative and nonincreasing function with compact support in $[0,\gamma/\delta]$ and satisfies

$$\|U\|_\infty \le z(\delta) \quad \text{and} \quad \int_0^\infty x^{(1-\lambda)/\lambda}\ U(x)\ dx \le B(\lambda).$$

On the other hand, since $(U_n)_{n\ge1}$ is bounded in $L^\infty(0,\infty)$ with $\text{Supp}\ U_n \subset [0,\gamma/\delta]$, we deduce from the Lebesgue dominated convergence theorem that (U_n) converges towards U in $L^1(0,\infty)$, from whence $\|U\|_1 = 1$. Therefore, $U \in \mathcal{K}_\delta$, and we have thus shown that $\mathcal{K}_\delta$ is a closed convex and compact subset of $L^1(0,\infty)$.

We now claim that, if $\Phi_0 \in \mathcal{K}_\delta$, then $\Phi(t,.) \in \mathcal{K}_\delta$ for each $t \ge 0$, Φ being the corresponding solution to (1.15), (1.16). Indeed, consider $\Phi_0 \in \mathcal{K}_\delta$. From the analysis of the previous section and (3.1), we know that $\Phi(t,.)$ is a nonnegative and nonincreasing function in $L^1(0,\infty)$ with $\|\Phi(t)\|_1 = \|\Phi_0\|_1 = 1$ for $t \ge 0$. Next, it readily follows from (3.3) that $m(t) \le \|\Phi_0\|_\infty \vee z(\delta)$, from whence $\|\Phi(t)\|_\infty \le z(\delta)$ for $t \ge 0$ by Lemma 3.1. Similarly, as $\text{Supp}\ \Phi_0 \subset [0,\gamma/\delta]$, the function R defined by (3.5) is bounded from above by $R(0) \vee (\gamma/\delta) = \gamma/\delta$ and we infer from Lemma 3.2 that $\text{Supp}\ \Phi(t,.) \subset [0,\gamma/\delta]$ for $t \ge 0$. Finally, Lemma 3.3 implies that

$$\int_0^\infty x^{(1-\lambda)/\lambda}\ \Phi(t,x)\ dx \le \left(\int_0^\infty x^{(1-\lambda)/\lambda}\ \Phi_0(x)\ dx \right) \vee B(\lambda) = B(\lambda), \quad t \ge 0.$$

Consequently, $\mathcal{K}_\delta$ is a closed convex and compact subset of $L^1(0,\infty)$ which is left invariant by the semiflow associated to (1.15), (1.16). Applying Theorem 1.4 with $X = L^1(0,\infty)$ and $K = \mathcal{K}_\delta$, we obtain the existence of a stationary solution Ψ_δ to (1.15) which belongs to $\mathcal{K}_\delta$. This last property and (3.9) yield the bounds (3.7) while (3.8) follows from (2.2). ∎

We are thus left to pass to the limit as $\delta \to 0$ to construct a stationary solution to (1.14) and this is the purpose of the next proposition.

PROPOSITION 3.5. *There exists a nonnegative and nonincreasing function* $\Psi \in L^1(0,\infty) \cap L^\infty(0,\infty)$ *such that*

$$\|\Psi\|_1 = 1, \quad \|\Psi\|_\infty \le A(\lambda), \quad \int_0^\infty x^{(1-\lambda)/\lambda}\ \Psi(x)\ dx \le B(\lambda), \tag{3.10}$$

and

$$x\ \left(\gamma - \int_0^x \Psi(x_*)^\lambda\ dx_* \right)\ \Psi(x) = \int_x^\infty \int_0^x \left(\Psi(x_*)^\lambda + \Psi(x')^\lambda \right)\ \Psi(x_*)\ dx' dx_* \tag{3.11}$$

for almost every $x \in (0,\infty)$*, the parameters* $A(\lambda)$ *and* $B(\lambda)$ *being defined in Proposition* 3.4.

Proof. For $\delta \in (0,1)$, let Ψ_δ be a stationary solution to (1.15) given by Proposition 3.4. We infer from Proposition 3.4 that $(\Psi_\delta)_{\delta\in(0,1)}$ is bounded in $BV(0,\infty) \cap$

$L^1(0,\infty;x^{(1-\lambda)/\lambda}\,dx)\cap L^\infty(0,\infty)$ which is clearly compactly embedded in $L^1(0,\infty)$ as $(1-\lambda)/\lambda>0$. Consequently, there is a sequence $(\delta_n)_{n\geq 1}$, $\delta_n\to 0$, and $\Psi\in L^1(0,\infty)$ such that Ψ_{δ_n} converges to Ψ in $L^1(0,\infty)$ and a.e. in $(0,\infty)$. Passing to the limit as $\delta_n\to 0$ in (3.7), we obtain (3.10) (with the help of the Fatou lemma for the moment estimate and a weak convergence argument for the L^∞-bound). Owing to the boundedness of (Ψ_{δ_n}) and the convergence of (Ψ_{δ_n}) towards Ψ in $L^1(0,\infty)$ and a.e., it is straightforward to pass to the limit as $\delta_n\to 0$ in (3.8) and obtain (3.11). □

4. Properties of stationary solutions to (1.14). We next turn to the study of properties of stationary solutions Ψ to (1.14) given by Proposition 3.5. We first establish their $\mathcal{C}^1$-smoothness on $[0,\infty)$.

PROPOSITION 4.1. *Let Ψ be a stationary solution to* (1.14) *given by Proposition* 3.5. *Then, $\Psi\in\mathcal{C}^1([0,\infty))$ and $\Psi^\lambda\in L^1(0,\infty)$ with*

$$L_\lambda := \|\Psi^\lambda\|_1 < \gamma. \tag{4.1}$$

Proof. We first observe that, since $\Psi\in L^\infty(0,\infty)$, we have

$$x_0 := \sup\left\{x\geq 0 \ \text{ such that } \ \int_0^x \Psi(x_*)^\lambda\ dx_* < \gamma\right\}\in(0,\infty].$$

It then clearly follows from (3.10) and (3.11) that $\Psi\in\mathcal{C}^1((0,x_0))$ with

$$\begin{aligned} & x\ \left(\gamma-\int_0^x \Psi(x_*)^\lambda\ dx_*\right)\ \frac{d\Psi}{dx}(x) \\ & = -\gamma\ \Psi(x)+\int_x^\infty \Psi(x_*)^{\lambda+1}\ dx_* + \Psi(x)^\lambda\ \int_x^\infty \Psi(x_*)\ dx_* \end{aligned} \tag{4.2}$$

for $x\in(0,x_0)$.

Assume for contradiction that $x_0<\infty$. Then

$$\int_0^{x_0}\Psi(x_*)^\lambda\ dx_* = \gamma \tag{4.3}$$

and the nonnegativity of Ψ and (3.11) imply that $\Psi(x)=0$ for $x>x_0$. It next follows from (4.2) and the monotonicity of Ψ that, if $x\in(0,x_0)$, we have

$$\begin{aligned} x\ \left(\gamma-\int_0^x \Psi(x_*)^\lambda\ dx_*\right)\ \frac{d\Psi}{dx}(x) & \leq -\gamma\ \Psi(x)+2\ \Psi(x)\ \int_x^{x_0}\Psi(x_*)^\lambda\ dx_* \\ & \leq -\gamma\ \Psi(x)+2\ \Psi(x)\ \left(\gamma-\int_0^x \Psi(x_*)^\lambda\ dx_*\right), \end{aligned}$$

from whence

$$x\ \frac{d\Psi}{dx}(x)\leq -\gamma\ \Psi(x)\ \left(\gamma-\int_0^x\Psi(x_*)^\lambda\ dx_*\right)^{-1}+2\ \Psi(x).$$

Therefore

$$\begin{aligned} \frac{d}{dx}\left\{x\ \Psi(x)^\lambda\right\} & = \Psi(x)^\lambda+\lambda\ x\ \Psi(x)^{\lambda-1}\ \frac{d\Psi}{dx}(x) \\ & \leq \Psi(x)^\lambda-\gamma\ \lambda\ \Psi(x)^\lambda\ \left(\gamma-\int_0^x\Psi(x_*)^\lambda\ dx_*\right)^{-1}+2\ \lambda\ \Psi(x)^\lambda \\ & \leq \frac{d}{dx}\left\{-(1+2\lambda)\ \int_x^{x_0}\Psi(x_*)^\lambda\ dx_*+\lambda\ \gamma\ \log\left(\gamma-\int_0^x\Psi(x_*)^\lambda\ dx_*\right)\right\}, \end{aligned}$$

from which we deduce after integration over $(0,x)$, $x\in(0,x_0)$ that

$$\begin{aligned} & x\ \Psi(x)^\lambda + (1+2\lambda)\ \int_x^{x_0} \Psi(x_*)^\lambda\ dx_* - \lambda\ \gamma\ \log\left(\gamma - \int_0^x \Psi(x_*)^\lambda\ dx_*\right) \\ & \le (1+2\lambda)\ \int_0^{x_0} \Psi(x_*)^\lambda\ dx_* - \lambda\ \gamma\ \log\gamma \\ & \le (1+2\lambda)\ \gamma - \lambda\ \gamma\ \log\gamma. \end{aligned} \tag{4.4}$$

But the right-hand side of (4.4) is finite while the left-hand side of (4.4) diverges to infinity as $x\to x_0$ by (4.3) and a contradiction. Therefore, $x_0=\infty$, $\Psi\in\mathcal{C}^1((0,\infty))$ and

$$\int_0^x \Psi(x_*)^\lambda\ dx_* < \gamma \quad \text{for each} \quad x\in(0,\infty).$$

In particular, $\Psi^\lambda\in L^1(0,\infty)$ and $L_\lambda=\|\Psi^\lambda\|_1\le\gamma$.

Suppose now for contradiction that $\|\Psi^\lambda\|_1=\gamma$. Arguing as before, we realize that (4.4) is valid for every $x\in(0,\infty)$. Then, letting $x\to\infty$ in (4.4) yields that the left-hand side of (4.4) diverges to infinity while the right-hand side is finite, from whence a contradiction. We have thus shown that $L_\lambda<\gamma$, from whence the claim (4.1).

We next turn to the regularity of Ψ at $x=0$. Since Ψ is a nonincreasing and bounded function, the limit $\Psi(0+)$ of $\Psi(x)$ as $x\to0$ exists and is finite and we may actually set $\Psi(0)=\Psi(0+)$. Then $\Psi\in\mathcal{C}([0,\infty))$. In addition, since $\|\Psi\|_1=1>0$, Ψ is not identically equal to zero and thus $\Psi(0)>0$ by the monotonicity of Ψ. Since the identity (3.11) also reads

$$\left(\gamma - \int_0^x \Psi(x_*)^\lambda\ dx_*\right)\ \Psi(x) = \int_x^\infty \Psi(x_*)^{\lambda+1}\ dx_* + \frac{1}{x}\ \int_0^x \Psi(x_*)^\lambda\, dx_*\ \int_x^\infty \Psi(x_*)\, dx_*,$$

we may let $x\to0$ in the previous inequality and obtain

$$\gamma\ \Psi(0) = \int_0^\infty \Psi(x_*)^{\lambda+1}\ dx_* + \Psi(0)^\lambda. \tag{4.5}$$

It follows from (3.11), (4.5) and the monotonicity of Ψ that

$$\begin{aligned} \gamma\ (\Psi(0)-\Psi(x)) &= \int_0^x \Psi(x_*)^{1+\lambda}\ dx_* - \Psi(x)\ \int_0^x \Psi(x_*)^\lambda\ dx_* \\ &\quad + \Psi(0)^\lambda - \frac{1}{x}\ \int_0^x \Psi(x_*)^\lambda\ dx_*\ \int_x^\infty \Psi(x_*)\ dx_* \\ &\le 2\ x\ \Psi(0)^{1+\lambda} + \left(\Psi(0)^\lambda - \frac{1}{x}\ \int_0^x \Psi(x_*)^\lambda\ dx_*\right) \\ &\le 2\ x\ \Psi(0)^{1+\lambda} + \left(\Psi(0)^\lambda - \Psi(x)^\lambda\right) \\ &\le 2\ x\ \Psi(0)^{1+\lambda} + \lambda\ \Psi(x)^{\lambda-1}\ (\Psi(0)-\Psi(x)), \end{aligned}$$

from whence

$$\left(\gamma - \lambda\ \Psi(x)^{\lambda-1}\right)\ \frac{\Psi(0)-\Psi(x)}{x} \le 2\ \Psi(0)^{1+\lambda}.$$

As $\lambda \in (0,1)$ and $\gamma \ge \Psi(0)^{\lambda-1}$ by (4.5), we have $\gamma > \lambda\ \Psi(0)^{\lambda-1}$ and we infer from the continuity of Ψ at $x=0$ that there are $\delta_1 > 0$ and $x_1 > 0$ such that

$$\gamma - \lambda\ \Psi(x)^{\lambda-1} \ge \delta_1 \quad \text{for} \quad x \in (0,x_1).$$

Combining the above two inequalities and the monotonicity of Ψ yields

$$0 \le \frac{\Psi(0)-\Psi(x)}{x} \le \frac{2}{\delta_1}\ \Psi(0)^{1+\lambda} \quad \text{for} \quad x \in (0,x_1), \tag{4.6}$$

so that $x \longmapsto (\Psi(0)-\Psi(x))/x$ belongs to $L^\infty(0,x_1)$.

Another consequence of (4.5) is that we may pass to the limit as $x \to 0$ in (4.2) to deduce that

$$\lim_{x\to 0} x\ \frac{d\Psi}{dx}(x) = 0. \tag{4.7}$$

Using once more (4.5), the identity (4.2) also reads

$$\begin{aligned}\gamma\ x\ \frac{d\Psi}{dx}(x) &= x \int_0^x \Psi(x_*)^\lambda\ dx_*\ \frac{d\Psi}{dx}(x) - \gamma\ (\Psi(x)-\Psi(0)) - \int_0^x \Psi(x_*)^{1+\lambda}\ dx_* \\ &\quad + \left(\Psi(x)^\lambda - \Psi(0)^\lambda\right) - \Psi(x)^\lambda \int_0^x \Psi(x_*)\ dx_*,\end{aligned}$$

from which we deduce that

$$\begin{aligned}&\gamma\ \frac{d\Psi}{dx}(x) + \left(\gamma - \lambda\ \Psi(0)^{\lambda-1}\right)\ \frac{\Psi(x)-\Psi(0)}{x} \\ &= \left(\frac{1}{x}\int_0^x \Psi(x_*)^\lambda\ dx_*\right)\left(x\ \frac{d\Psi}{dx}(x)\right) - \frac{1}{x}\int_0^x \Psi(x_*)^{1+\lambda}\ dx_* - \frac{\Psi(x)^\lambda}{x}\int_0^x \Psi(x_*)\ dx_* \\ &+ \frac{1}{x}\ \left(\Psi(x)^\lambda - \Psi(0)^\lambda - \lambda\ \Psi(0)^{\lambda-1}(\Psi(x)-\Psi(0))\right),\end{aligned}$$

On one hand, by (4.7) and the continuity of Ψ at $x=0$, we have

$$\begin{aligned}&\lim_{x\to 0}\left(\frac{1}{x}\int_0^x \Psi(x_*)^\lambda\ dx_*\right)\left(x\ \frac{d\Psi}{dx}(x)\right) = 0, \\ &\lim_{x\to 0}\frac{1}{x}\int_0^x \Psi(x_*)^{1+\lambda}\ dx_* = \lim_{x\to 0}\frac{\Psi(x)^\lambda}{x}\int_0^x \Psi(x_*)\ dx_* = \Psi(0)^{1+\lambda}.\end{aligned}$$

On the other hand, it follows from the concavity of $r \mapsto r^\lambda$, the monotonicity of Ψ, and (4.6) that, if $x \in (0,x_1)$, then

$$\begin{aligned}&\left|\frac{1}{x}\ \left(\Psi(x)^\lambda - \Psi(0)^\lambda - \lambda\ \Psi(0)^{\lambda-1}(\Psi(x)-\Psi(0))\right)\right| \\ &\quad = \frac{1}{x}\ \left(\Psi(0)^\lambda - \Psi(x)^\lambda - \lambda\ \Psi(0)^{\lambda-1}(\Psi(0)-\Psi(x))\right) \\ &\quad \le \frac{\lambda}{x}\ \left(\Psi(x)^{\lambda-1} - \Psi(0)^{\lambda-1}\right)\ (\Psi(0)-\Psi(x)) \\ &\quad \le \lambda \sup_{x_*\in(0,x_1)}\left(\frac{\Psi(0)-\Psi(x_*)}{x_*}\right)\ \left(\Psi(x)^{\lambda-1} - \Psi(0)^{\lambda-1}\right) \\ &\quad \le \frac{2\ \lambda}{\delta_1}\ \Psi(0)^{1+\lambda}\ \left(\Psi(x)^{\lambda-1} - \Psi(0)^{\lambda-1}\right) \underset{x\to 0}{\longrightarrow} 0\end{aligned}$$

by the continuity of Ψ at $x=0$. Consequently,

$$\lim_{x\to 0}\left\{\gamma\ \frac{d\Psi}{dx}(x)+\left(\gamma-\lambda\ \Psi(0)^{\lambda-1}\right)\ \frac{\Psi(x)-\Psi(0)}{x}\right\}=-2\ \Psi(0)^{1+\lambda}. \tag{4.8}$$

Introducing $\omega:=1-\lambda\ \Psi(0)^{\lambda-1}\ \gamma^{-1}$, we have $\omega>0$ by (4.5) and the previous limit also reads

$$\lim_{x\to 0}\left\{x^{-\omega}\ \frac{d}{dx}\left(x^{\omega}\ (\Psi(x)-\Psi(0))\right)\right\}=-\frac{2\ \Psi(0)^{1+\lambda}}{\gamma},$$

from whence, by integration,

$$\lim_{x\to 0}\frac{\Psi(x)-\Psi(0)}{x}=-\frac{2\ \Psi(0)^{1+\lambda}}{\gamma\ (\omega+1)}.$$

Therefore, Ψ is differentiable at $x=0$ with

$$\frac{d\Psi}{dx}(0)=-\frac{2\ \Psi(0)^{1+\lambda}}{2\ \gamma-\lambda\ \Psi(0)^{\lambda-1}}<0, \tag{4.9}$$

and (4.8) ensures the continuity of $d\Psi/dx$ at $x=0$. ∎

We next turn to the positivity and monotonicity properties of stationary solutions to (1.14).

PROPOSITION 4.2. *Let Ψ be a stationary solution to* (1.14) *given by Proposition* 3.5. *Then*

$$\Psi(x)>0\quad \text{and}\quad \frac{d\Psi}{dx}(x)<0\quad \text{for}\quad x\ge 0. \tag{4.10}$$

Proof. Recalling that $L_\lambda:=\|\Psi^\lambda\|_1<\gamma$ by Proposition 4.1, we have for $\delta>0$ and $x\ge\delta$

$$\delta\ (\gamma-L_\lambda)\le x\ \left(\gamma-\int_0^x \Psi(x_*)^\lambda\ dx_*\right).$$

We then infer from (4.2) and the nonpositivity of $d\Psi/dx\le 0$ that

$$\delta\ (\gamma-L_\lambda)\ \frac{d\Psi}{dx}(x)\ge -\gamma\ \Psi(x),\quad x\ge\delta.$$

Therefore,

$$\Psi(x)\ge\Psi(\delta)\ \exp\left\{\frac{\gamma\ (\delta-x)}{\delta\ (\gamma-L_\lambda)}\right\},\quad x\ge\delta.$$

Owing to the continuity of Ψ at $x=0$ and the positivity of $\Psi(0)$, we also have $\Psi(\delta)>0$ for δ sufficiently small, which, together with the above lower bound for Ψ, entail the positivity of Ψ in $[0,\infty)$. Similarly, it follows from (4.2) that Ψ is twice differentiable in $(0,\infty)$ with

$$\begin{aligned}
&x\ \left(\gamma-\int_0^x \Psi(x_*)^\lambda\ dx_*\right)\ \frac{d^2\Psi}{dx^2}(x)\\
&=\left(x\ \Psi(x)^\lambda+\lambda\ \Psi(x)^{\lambda-1}\ \int_x^\infty \Psi(x_*)\ dx_*+\int_0^x \Psi(x_*)^\lambda\ dx_*-2\ \gamma\right)\ \frac{d\Psi}{dx}(x)\\
&-2\ \Psi(x)^{1+\lambda}\\
&\le -2\ \gamma\ \frac{d\Psi}{dx}(x),
\end{aligned}$$

the last inequality being a consequence of the positivity and monotonicity of Ψ. Consequently, if $\delta > 0$ and $x \geq \delta$,

$$\begin{aligned}\frac{d^2\Psi}{dx^2}(x) &\leq -\frac{2\,\gamma}{x}\,\frac{d\Psi}{dx}(x)\,\left(\gamma - \int_0^x \Psi(x_*)^\lambda\,dx_*\right)^{-1}\\ &\leq -\frac{2\,\gamma}{\delta\,(\gamma - L_\lambda)}\,\frac{d\Psi}{dx}(x),\end{aligned}$$

from whence

$$\frac{d\Psi}{dx}(x) \leq \frac{d\Psi}{dx}(\delta)\,\exp\left\{\frac{2\,\gamma\,(\delta - x)}{\delta\,(\gamma - L_\lambda)}\right\}, \quad x \geq \delta.$$

Recalling that $d\Psi/dx \in \mathcal{C}([0,\infty))$ with $d\Psi(0)/dx < 0$ by (4.9), we easily deduce from the previous inequality that $d\Psi(x)/dx < 0$ for $x \geq 0$. □

We finally identify the behavior of Ψ as $x \to \infty$.

PROPOSITION 4.3. *Let Ψ be a stationary solution to* (1.14) *given by Proposition* 3.5. *Then $L_\lambda := \left\|\Psi^\lambda\right\|_1 > 1$ and there is a positive constant $b > 0$ such that*

$$\lim_{x\to\infty} x^\alpha\,\Psi(x) = b \quad \textit{and} \quad \lim_{x\to\infty} x^{1+\alpha}\,\frac{d\Psi}{dx}(x) = -\alpha\,b, \tag{4.11}$$

with $\alpha := \gamma/(\gamma - L_\lambda) > 0$.

Proof. We first establish that

$$L_\lambda = \left\|\Psi^\lambda\right\|_1 > 1. \tag{4.12}$$

Indeed, since $\Psi(x) > 0$ for $x \geq 0$, we multiply (4.2) by $\lambda\,\Psi(x)^{\lambda-1}$ and integrate over $(0,\infty)$ to obtain

$$\begin{aligned}&\left[x\,\left(\gamma - \int_0^x \Psi(x_*)^\lambda\,dx_*\right)\,\Psi(x)^\lambda\right]_0^\infty\\ &\qquad - \int_0^\infty \Psi(x)^\lambda\,\left(\gamma - \int_0^x \Psi(x_*)^\lambda\,dx_* - x\,\Psi(x)^\lambda\right)\,dx\\ &\qquad = -\lambda\,\gamma\,L_\lambda + \lambda\,\int_0^\infty \int_x^\infty \left(\Psi(x)^{\lambda-1}\,\Psi(x_*)^{\lambda+1} + \Psi(x)^{2\lambda-1}\,\Psi(x_*)\right)\,dx_* dx\\ &\qquad = -\lambda\,\gamma\,L_\lambda + \lambda\,\int_0^\infty \int_0^x \left(\Psi(x)^{\lambda+1}\,\Psi(x_*)^{\lambda-1} + \Psi(x)\,\Psi(x_*)^{2\lambda-1}\right)\,dx_* dx.\end{aligned}$$

Since Ψ is nonincreasing and $\Psi^\lambda \in L^1(0,\infty)$, the boundary terms vanish and we end up with

$$\begin{aligned}(1-\lambda)\,\gamma\,L_\lambda &= \int_0^\infty \int_0^x \left(\Psi(x)^\lambda\,\Psi(x_*)^\lambda + \Psi(x)^{2\lambda}\right)\,dx_* dx\\ &\quad - \lambda\,\int_0^\infty \int_0^x \left(\Psi(x)^{\lambda+1}\,\Psi(x_*)^{\lambda-1} + \Psi(x)\,\Psi(x_*)^{2\lambda-1}\right)\,dx_* dx,\end{aligned}$$

$$L_\lambda = \int_0^\infty \int_0^x \left(\Psi(x)^\lambda + \Psi(x_*)^\lambda\right)\,\left(\Psi(x)^{\lambda-1} - \lambda\,\Psi(x_*)^{\lambda-1}\right)\,\Psi(x)\,dx_* dx.$$

Since $\Psi(x) \le \Psi(x_*)$ for $x_* \in (0,x)$, it follows from Lemma B.2 and the elementary inequality $\Psi(x)^\lambda + \Psi(x_*)^\lambda \le 2^{1-\lambda}\ (\Psi(x)+\Psi(x_*))^\lambda$ that

$$L_\lambda \le 2^{1-\lambda} \int_0^\infty \int_0^x \Psi(x)^\lambda\ \Psi(x_*)^\lambda dx_* dx = 2^{-\lambda}\ L_\lambda^2.$$

Therefore, $L_\lambda \ge 2^\lambda > 1$ which completes the proof of (4.12).

We next infer from (4.2) and the positivity and monotonicity of Ψ that

$$x\ (\gamma - L_\lambda)\ \frac{d\Psi}{dx}(x) \ge x\ \left(\gamma - \int_0^x \Psi(x_*)^\lambda\ dx_*\right)\ \frac{d\Psi}{dx}(x) \ge -\gamma\ \Psi(x),$$

from whence

$$x\ \left|\frac{d\Psi}{dx}(x)\right| \le \frac{\gamma}{\gamma - L_\lambda}\ \Psi(x),$$

and

$$(\gamma - L_\lambda)\ x\ \frac{d\Psi}{dx}(x) + \gamma\ \Psi(x) \le \left(2\ \Psi(x) + x\ \left|\frac{d\Psi}{dx}(x)\right|\right)\ \int_x^\infty \Psi(x_*)^\lambda\ dx_*$$

for $x > 0$. We combine the previous inequalities to obtain

$$(\gamma - L_\lambda)\ x\ \frac{d\Psi}{dx}(x) + \gamma\ \Psi(x) \le \left(2 + \frac{\gamma}{\gamma - L_\lambda}\right)\ \Psi(x)\ \int_x^\infty \Psi(x_*)^\lambda\ dx_*.$$

Recall that $\alpha = \gamma/(\gamma - L_\lambda) > 0$ and fix $\varepsilon \in (0,\alpha)$. Since $\Psi^\lambda \in L^1(0,\infty)$, there is $x_\varepsilon > 0$ such that

$$\int_x^\infty \Psi(x_*)^\lambda\ dx_* \le \frac{\gamma - L_\lambda}{2+\alpha}\ \varepsilon \quad \text{for} \quad x \ge x_\varepsilon.$$

Therefore, for $x \ge x_\varepsilon$, we have

$$x\ \frac{d\Psi}{dx}(x) + \alpha\ \Psi(x) \le \varepsilon\ \Psi(x),$$

from which we deduce by integration that

$$\Psi(x) \le \frac{x_\varepsilon^{\alpha-\varepsilon}\ \Psi(x_\varepsilon)}{x^{\alpha-\varepsilon}} \quad \text{for} \quad x \ge x_\varepsilon.$$

Recalling that $\Psi \in L^\infty(0,\infty)$, we have thus established that, for each $\varepsilon \in (0,\alpha)$, there is $\kappa_\varepsilon > 0$ such that

$$\Psi(x) \le \kappa_\varepsilon\ x^{-\alpha+\varepsilon} \quad \text{for} \quad x > 0. \tag{4.13}$$

Once more we use (4.2) to obtain that

$$\begin{aligned}(\gamma - L_\lambda)\ \frac{d}{dx}\left\{x^\alpha\ \Psi(x)\right\} &= (\gamma - L_\lambda)\ x^\alpha\ \frac{d\Psi(x)}{dx} + \gamma\ x^{\alpha-1}\ \Psi(x) \\ &= x^{\alpha-1}\ \left(\int_x^\infty \Psi(x_*)^{1+\lambda}\ dx_* + \Psi(x)^\lambda\ \int_x^\infty \Psi(x_*)\ dx_*\right) \\ &\quad - x^\alpha\ \frac{d\Psi}{dx}(x)\ \int_x^\infty \Psi(x_*)^\lambda\ dx_*. \end{aligned} \tag{4.14}$$

Now, since $L_\lambda > 1$ by (4.12), we have $\alpha > \lambda\alpha > 1$ and there exists $\varepsilon \in (0, \alpha - 1)$ such that $\lambda\alpha > 1 + (1+\lambda)\varepsilon$. Then, on one hand, by the monotonicity of Ψ and (4.13), an integration by parts yields

$$\begin{aligned}
0 &\le -\int_1^\infty x^\alpha \ \frac{d\Psi}{dx}(x) \ \int_x^\infty \Psi(x_*)^\lambda \ dx_* dx \\
&\le \Psi(1) \int_1^\infty \Psi(x_*)^\lambda \ dx_* + \int_1^\infty \Psi(x) \ \left(\alpha \ x^{\alpha-1} \ \int_x^\infty \Psi(x_*)^\lambda \ dx_* - x^\alpha \ \Psi(x)^\lambda \right) \ dx \\
&\le C + \frac{\alpha \ \kappa_\varepsilon^{1+\lambda}}{\lambda\alpha - 1 - \lambda\varepsilon} \ \int_1^\infty x^{(1+\lambda)\varepsilon - \lambda\alpha} \ dx \\
&\le C(\varepsilon).
\end{aligned}$$

On the other hand, it follows from the monotonicity of Ψ and (4.13) that

$$\begin{aligned}
&\int_1^\infty x^{\alpha-1} \ \left(\int_x^\infty \Psi(x_*)^{1+\lambda} \ dx_* + \Psi(x)^\lambda \ \int_x^\infty \Psi(x_*) \ dx_* \right) \ dx \\
&\le 2 \ \frac{\kappa_\varepsilon^{1+\lambda}}{\alpha - 1 - \varepsilon} \ \int_1^\infty x^{(1+\lambda)\varepsilon - \lambda\alpha} \ dx \le C(\varepsilon).
\end{aligned}$$

Consequently, the right-hand side of (4.14) belongs to $L^1(1,\infty)$ and is positive, from which we conclude that $x \longmapsto x^\alpha \ \Psi(x)$ has a positive limit b as $x \to \infty$. We have thus proved the first assertion in (4.11).

As for $d\Psi/dx$, we note that the large x-behavior of Ψ ensures that there is x_∞ large enough such that $x^\alpha \ \Psi(x) \le 2b$ for $x \ge x_\infty$. Consequently, for $x \ge x_\infty$, we have

$$\begin{aligned}
x^\alpha \ \left(\int_x^\infty \Psi(x_*)^{\lambda+1} \ dx_* + \Psi(x)^\lambda \ \int_x^\infty \Psi(x_*) \ dx_* \right) &\le 2 \ x^\alpha \ \Psi(x)^\lambda \ \int_x^\infty \Psi(x_*) \ dx_* \\
&\le 2 \ (2b)^{\lambda+1} \ x^{(1-\lambda)\alpha} \int_x^\infty x_*^{-\alpha} \ dx_* \\
&\le \frac{2 \ (2b)^{\lambda+1}}{\alpha - 1} \ x^{1-\lambda\alpha}.
\end{aligned}$$

Recalling that $\lambda\alpha > 1$ by (4.12), we conclude that

$$\lim_{x\to\infty} x^\alpha \ \left(\int_x^\infty \Psi(x_*)^{\lambda+1} \ dx_* + \Psi(x)^\lambda \ \int_x^\infty \Psi(x_*) \ dx_* \right) = 0.$$

We now multiply (4.2) by x^α and let $x \to \infty$ in the resulting identity with the help of the previous limit and the first statement in (4.11) to complete the proof of (4.11). $\square$

5. Proofs of Theorem 1.1 and Corollary 1.2. By Proposition 3.5 there exists a nonnegative function $\Psi \in L^1(0,\infty)$ satisfying (3.10) and (3.11). In addition, $\Psi \in \mathcal{C}^1([0,\infty))$ is a positive and decreasing function from $[0,\infty)$ onto $(0, y_0]$ with $y_0 := \Psi(0)$ by Propositions 4.1 and 4.2. We then denote its inverse function by $\Xi : (0, y_0] \longrightarrow [0,\infty)$ which is also a decreasing and nonnegative function in $\mathcal{C}^1((0,y_0])$ and put $\xi := -d\Xi/dy > 0$. We extend Ξ and ξ to (y_0,∞) by setting $\Xi(y) = \xi(y) = 0$ for $y > y_0$. Then $\xi \in \mathcal{C}((0,\infty) \setminus \{y_0\})$ with $\xi(y_0-) = -(d\Psi/dx(0))^{-1} > 0$ by (4.9) and $\xi(y_0+) = 0$. Also, since $\Xi(y) \longrightarrow \infty$ as $y \to 0$, we infer from Propositions 4.1 and 4.3 that

$$\lim_{y\to 0}\ y\ \Xi(y)^{\gamma/(\gamma-L_\lambda)} = b, \quad \lim_{y\to 0}\ -\frac{\Xi(y)^{1+(\gamma/(\gamma-L_\lambda))}}{\xi(y)} = -\frac{\gamma\ b}{\gamma - L_\lambda}$$

with $L_\lambda := \left\|\Psi^\lambda\right\|_1 \in (1,\gamma)$, from whence

$$\lim_{y\to 0}\ y^{2-(L_\lambda/\gamma)}\ \xi(y) = \frac{\gamma - L_\lambda}{\gamma}\ b^{1-L_\lambda/\gamma} > 0.$$

In addition, Ξ being a $\mathcal{C}^1$-diffeomorphism from $(0,y_0]$ onto $[0,\infty)$, a simple change of variables yields

$$\int_0^\infty y\ \xi(y)\ dy = \|\Psi\|_1 = 1 \quad \text{and} \quad \int_0^\infty y^\lambda\ \xi(y)\ dy = \left\|\Psi^\lambda\right\|_1 = L_\lambda.$$

We have thus shown that ξ enjoys all the properties (1.9), (1.10), and (1.11) listed in Theorem 1.1. To check (1.12), we consider $y \in (0,y_0)$ and take $x = \Xi(y)$ in (4.2) to obtain

$$-\left(\gamma - \int_0^{\Xi(y)} \Psi(x_*)^\lambda\ dx_*\right)\ \frac{\Xi(y)}{\xi(y)} = -\gamma\ y + \int_{\Xi(y)}^\infty \left(\Psi(x_*)^{\lambda+1} + y^\lambda\ \Psi(x_*)\right)\ dx_*,$$

from whence (1.12) after performing the change of variables $x_* = \Xi(y_*)$ in the integrals. The proof of Theorem 1.1 is now complete and we show that Corollary 1.2 follows. Let $t \in (0,\infty)$ and $\vartheta \in \mathcal{C}_0^\infty((0,\infty))$. Then

$$\begin{aligned}
&\frac{d}{dt}\int_0^\infty g_s(t,y)\ \vartheta(y)\ dy \\
&\quad = \frac{d}{dt}\left(\frac{1}{t^{2\gamma}}\ \int_0^\infty \xi\left(\frac{y}{t^\gamma}\right)\ \vartheta(y)\ dy\right) \\
&\quad = \frac{d}{dt}\left(\frac{1}{t^{\gamma}}\ \int_0^\infty \xi(y)\ \vartheta\left(y\ t^\gamma\right)\ dy\right) \\
&\quad = -\frac{\gamma}{t^{1+\gamma}}\ \int_0^\infty \xi(y)\ \vartheta\left(y\ t^\gamma\right)\ dy + \frac{\gamma}{t}\ \int_0^\infty y\ \xi(y)\ \partial_y\vartheta\left(y\ t^\gamma\right)\ dy \\
&\quad = \frac{\gamma}{t}\int_0^\infty y\ \xi(y)\ \partial_y\vartheta\left(y\ t^\gamma\right)\ dy - \frac{\gamma}{t}\int_0^\infty \xi(y)\ \int_0^y \partial_y\vartheta\left(y_*\ t^\gamma\right)\ dy_* dy,
\end{aligned}$$

from whence

$$t\ \frac{d}{dt}\int_0^\infty g_s(t,y)\ \vartheta(y)\ dy = \gamma\ \int_0^\infty \left(y\ \xi(y) - \int_y^\infty \xi(y_*)\ dy_*\right)\partial_y\vartheta\left(y\ t^\gamma\right)\ dy_* dy.$$

Using (1.12), we deduce that

$$
\begin{aligned}
t\ \frac{d}{dt}\int_0^\infty & g_s(t,y)\ \vartheta(y)\ dy \\
&= \int_0^\infty \partial_y\vartheta\left(y\ t^\gamma\right)\ \int_0^y \left(y^\lambda + y_*^\lambda\right)\ y_*\ \xi(y_*)\ dy_*\ \xi(y)\ dy \\
&- \int_0^\infty \partial_y\vartheta\left(y\ t^\gamma\right)\ \left(\int_y^\infty y_*^\lambda\ \xi(y_*)\ dy_*\right)\ \left(\int_y^\infty \xi(y')\ dy'\right)\ dy \\
&= \frac{1}{t^{(3+\lambda)\gamma}}\ \int_0^\infty \partial_y\vartheta(y)\ \int_0^y \left(y^\lambda + y_*^\lambda\right)\ y_*\ \xi\left(\frac{y_*}{t^\gamma}\right)\ dy_*\ \xi\left(\frac{y}{t^\gamma}\right)\ dy \\
&- \frac{1}{t^\gamma}\ \int_0^\infty \vartheta\left(y\ t^\gamma\right)\ \int_y^\infty \left(y^\lambda + y_*^\lambda\right)\ \xi(y_*)\ \xi(y)\ dy_* dy \\
&= t^{(1-\lambda)\gamma}\ \int_0^\infty \partial_y\vartheta(y)\ \int_0^y \left(y^\lambda + y_*^\lambda\right)\ y_*\ g_s(t,y_*)\ g_s(t,y)\ dy_* dy \\
&- t^{(1-\lambda)\gamma}\ \int_0^\infty \vartheta(y)\ \int_y^\infty \left(y^\lambda + y_*^\lambda\right)\ g_s(t,y_*)\ g_s(t,y)\ dy_* dy \\
&= t\int_0^\infty \int_0^y \left(y_*\,\partial_y\vartheta(y) - \vartheta(y_*)\right)\ \left(y^\lambda + y_*^\lambda\right)\ g_s(t,y_*)\, g_s(t,y)\, dy_* dy,
\end{aligned}
$$

since $\gamma = 1/(1-\lambda)$. Dividing the above equality by t yields that g_s is a weak solution to (1.2) and completes the proof of Corollary 1.2. ∎

Appendix A. Proof of Theorem 1.4.

We first recall the definition of a semiflow.

DEFINITION A.1. *Let X be a topological vector space. A semiflow $\mathcal{F} : [0,\infty) \times X \to X$ is a continuous mapping such that $\mathcal{F}(0,x) = x$ and $\mathcal{F}(t,\mathcal{F}(s,x)) = \mathcal{F}(t+s,x)$ for every $(s,t,x) \in [0,\infty)^2 \times X$.*

We first need a preliminary result.

PROPOSITION A.2. *Let X be a topological vector space and $\mathcal{F}$ a semiflow on X. For $T > 0$, we denote the set of T-periodic orbits by K_T, that is,*

$$K_T := \{x \in X, \quad \mathcal{F}(T,x) = x\}.$$

If there is $T > 0$ such that

$$K_T \text{ is compact} \qquad \text{and} \qquad K_{T2^{-k}} \neq \emptyset \qquad \text{for every } k \in \mathbb{N},$$

then $K := \{x \in X \text{ such that } \mathcal{F}(t,x) = x \text{ for all } t \geq 0\}$ is a nonempty compact subset of X and $K = \cap_{k\geq 0} K_{T2^{-k}}$.

Proof. Let $k \geq 0$. If $x \in K_{T2^{-(k+1)}}$, then, by definition of a semiflow,

$$
\begin{aligned}
\mathcal{F}(T2^{-k},x) &= \mathcal{F}(T2^{-(k+1)},\mathcal{F}(T2^{-(k+1)},x)) \\
&= \mathcal{F}(T2^{-(k+1)},x) = x,
\end{aligned}
$$

from whence $K_{T2^{-(k+1)}} \subset K_{T2^{-k}}$. Moreover, $K_t = (Id - \mathcal{F}(t,.))^{-1}(\{0\})$ is a closed set for every $t \geq 0$ and $K_{T2^{-k}}$ is thus compact for every $k \geq 0$. Therefore $K_{stat} := \cap_{k\geq 0} K_{T2^{-k}}$ is a nonempty compact subset of X.

Consider now $x \in K_{stat}$ for which we have

$$\mathcal{F}(T2^{-k},x) = x \quad \text{for every} \quad k \in \mathbb{N}.$$

Let $t \in [0,T]$. For every $k \geq 1$, there is $i_k \in \{0,\dots,2^k-1\}$ such that

$$\frac{i_k}{2^k} \leq \frac{t}{T} < \frac{1+i_k}{2^k} \qquad \text{and} \qquad \lim_{k\to\infty} \frac{i_k}{2^k} = \frac{t}{T}.$$

But, by induction,

$$\mathcal{F}\left(\frac{Ti_k}{2^k}, x\right) = \mathcal{F}\left(\frac{T(i_k-1)}{2^k}, \mathcal{F}\left(\frac{T}{2^k}, x\right)\right) = x,$$

and the continuity of $\mathcal{F}$ leads to $\mathcal{F}(t,x) = x$. We may then extend this result to (T,∞) by the semiflow property. We have thus shown that $K_{stat} = K$. ◻

We are now in a position to prove Theorem 1.4.

Proof of Theorem 1.4. Let $T > 0$. Since X is a locally convex topological vector space and $\mathcal{F}(T,.) : K \to K$ is continuous, the Tychonov–Schauder fixed point theorem ensures the existence of $x_T \in K$ such that

$$\mathcal{F}(T, x_T) = x_T.$$

Then, with the notations introduced in Proposition A.2, K_T is a nonempty closed subset of K and K_T is consequently compact. We thus conclude thanks to Proposition A.2. ◻

Appendix B. Two inequalities.

Lemma B.1. *Consider $\vartheta \in (0,1)$ and a nonnegative and nonincreasing measurable function U such that $U^\vartheta \in L^\infty(0,\infty)$. Then*

$$\int_0^x x_*^{(1-\vartheta)/\vartheta}\ U(x_*)\ dx_* \leq \vartheta\ \left(\int_0^x U(x_*)^\vartheta\ dx_*\right)^{1/\vartheta} \tag{B.1}$$

for $x \in (0,\infty)$. Furthermore, if $U^\vartheta \in L^1(0,\infty)$, then $U \in L^1(0,\infty; x^{(1-\vartheta)/\vartheta}\ dx)$ and

$$\int_0^\infty x^{(1-\vartheta)/\vartheta}\ U(x)\ dx \leq \vartheta\ \left(\int_0^\infty U(x)^\vartheta\ dx\right)^{1/\vartheta}. \tag{B.2}$$

Proof. Consider $x \in (0,\infty)$ and $x_* \in (0,x)$. By the monotonicity of U, we have

$$x_*\ U(x_*)^\vartheta \leq \int_0^{x_*} U(y)^\vartheta\ dy,$$

from whence

$$x_*^{1/\vartheta}\ U(x_*) \leq \left(\int_0^{x_*} U(y)^\vartheta\ dy\right)^{1/\vartheta}.$$

Consequently,

$$\begin{aligned}\int_0^x x_*^{(1-\vartheta)/\vartheta}\ U(x_*)\ dx_* &= \int_0^x x_*^{(1-\vartheta)/\vartheta}\ U(x_*)^{1-\vartheta}\ U(x_*)^\vartheta\ dx_* \\ &\leq \int_0^x U(x_*)^\vartheta\ \left(\int_0^{x_*} U(y)^\vartheta\ dy\right)^{(1-\vartheta)/\vartheta}\ dx_* \\ &= \vartheta\ \left(\int_0^x U(x_*)^\vartheta\ dx_*\right)^{1/\vartheta},\end{aligned}$$

from whence (B.1). Next, if $U^\vartheta \in L^1(0,\infty)$, we may let $x \to \infty$ in (B.1) to obtain (B.2) and thus complete the proof of Lemma B.1. ∎

LEMMA B.2. *Consider $\lambda \in [0,1]$, $r > 0$, and $r_* \in (0,r)$. Then*

$$(1-\lambda)\ 2^\lambda\ \frac{r^\lambda\ r_*^{\lambda-1}}{(r+r_*)^\lambda} \le r_*^{\lambda-1} - \lambda\ r^{\lambda-1} \le \frac{r^\lambda\ r_*^{\lambda-1}}{(r+r_*)^\lambda}. \tag{B.3}$$

Proof. The inequalities (B.3) being obvious for $\lambda \in \{0,1\}$, we restrict ourselves to $\lambda \in (0,1)$ and put

$$p(z) := (1+z)^\lambda\ \left(1 - \lambda\ z^{1-\lambda}\right), \quad q(z) := z^\lambda - z - (1-\lambda)$$

for $z \in (0,1)$. Then

$$p'(z) = \frac{\lambda}{z^\lambda\ (1+z)^{1-\lambda}}\ q(z), \quad q'(z) = \lambda\ z^{\lambda-1} - 1,$$

from which we deduce that $q(z) \le q\left(\lambda^\gamma\right)$ for $z \in (0,1)$. Since $q\left(\lambda^\gamma\right) = \left(\lambda^{\lambda\gamma} - 1\right)\ (1-\lambda) < 0$, we conclude that $p'(z) \le 0$ for $z \in (0,1)$. Consequently, $p(1) \le p(z) \le p(0)$ for $z \in (0,1)$, from whence

$$(1-\lambda)\ 2^\lambda \le p(z) \le 1 \quad \text{for} \quad z \in (0,1).$$

We next consider $r > 0$ and $r_* \in (0,r)$ and take $z = r_*/r$ in the previous inequality to obtain (B.3). ∎

Acknowledgment. We thank the referees for helpful comments.

REFERENCES

[1] H. Amann, *Ordinary Differential Equations. An Introduction to Nonlinear Analysis*, de Gruyter Studies in Mathematics 13, Walter de Gruyter & Co., Berlin, 1990.

[2] M. Deaconu and E. Tanré, *Smoluchowski's coagulation equation: Probabilistic interpretation of solutions for constant, additive and multiplicative kernels*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 29 (2000), pp. 549–579.

[3] R. J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[4] P. G. J. van Dongen and M. H. Ernst, *Scaling solutions of Smoluchowski's coagulation equation*, J. Statist. Phys., 50 (1988), pp. 295–329.

[5] M. Escobedo and S. Mischler, *Dust and self-similarity for the Smoluchowski coagulation equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 23 (2006), pp. 331–362.

[6] M. Escobedo, S. Mischler, and M. Rodriguez-Ricard, *On self-similarity and stationary problems for fragmentation and coagulation models*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 22 (2005), pp. 99–125.

[7] F. Filbet and Ph. Laurençot, *Numerical simulation of the Smoluchowski coagulation equation*, SIAM J. Sci. Comput., 25 (2004), pp. 2004–2028.

[8] N. Fournier and Ph. Laurençot, *Existence of self-similar solutions to Smoluchowski's coagulation equation*, Comm. Math. Phys., 256 (2005), pp. 589–609.

[9] N. Fournier and Ph. Laurençot, *Local properties of self-similar solutions to Smoluchowski's coagulation equation with sum kernels*, Proc. Roy. Soc. Edinburgh Sect. A, 136 (2006), pp. 485–508.

[10] N. Fournier and Ph. Laurençot, *Well-posedness of Smoluchowski's coagulation equation for a class of homogeneous kernels*, J. Funct. Anal., 233 (2006), pp. 351–379.

[11] I. Gamba, V. Panferov, and C. Villani, *On the Boltzmann equation for diffusively excited granular media*, Comm. Math. Phys., 246 (2004), pp. 503–541.

[12] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1970.

[13] M. Kreer and O. Penrose, *Proof of dynamical scaling in Smoluchowski's coagulation equation with constant kernel*, J. Statist. Phys., 75 (1994), pp. 389–407.
[14] D. S. Krivitsky, *Numerical solution of the Smoluchowski kinetic equation and asymptotics of the distribution function*, J. Phys. A, 28 (1995), pp. 2025–2039.
[15] M. Lachowicz, Ph. Laurençot, and D. Wrzosek, *On the Oort-Hulst-Safronov coagulation equation and its relation to the Smoluchowski equation*, SIAM J. Math. Anal., 34 (2003), pp. 1399–1421.
[16] Ph. Laurençot, *Convergence to self-similar solutions for a coagulation equation*, Z. Angew. Math. Phys., 56 (2005), pp. 398–411.
[17] Ph. Laurençot and S. Mischler, *Liapunov functionals for Smoluchowski's coagulation equation and convergence to self-similarity*, Monatsh. Math., 146 (2005), pp. 127–142.
[18] M. H. Lee, *A survey of numerical solutions to the coagulation equation*, J. Phys. A, 34 (2001), pp. 10219–10241.
[19] F. Leyvraz, *Scaling theory and exactly solved models in the kinetics of irreversible aggregation*, Phys. Rep., 383 (2003), pp. 95–212.
[20] G. Menon and R. L. Pego, *Approach to self-similarity in Smoluchowski's coagulation equations*, Comm. Pure Appl. Math., 57 (2004), pp. 1197–1232.
[21] J. H. Oort and H. C. van de Hulst, *Gas and smoke in interstellar space*, Bull. Astron. Inst. Netherland, 10 (1946), pp. 187–210.
[22] V. S. Safronov, *Evolution of the Protoplanetary Cloud and Formation of the Earth and the Planets*, Israel Program for Scientific Translations Ltd., Jerusalem, 1972.
[23] J. Simon, *Compact sets in the space* $L^p(0,T;B)$, Ann. Mat. Pura Appl., 146 (1987), pp. 65–96.
[24] M. Smoluchowski, *Drei Vorträge über Diffusion, Brownsche Molekularbewegung und Koagulation von Kolloidteilchen*, Physik. Zeitschr., 17 (1916), pp. 557–599.
[25] M. Smoluchowski, *Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Lösungen*, Zeitschrift f. Physik. Chemie, 92 (1917), pp. 129–168.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 379–421

L_P-THEORY FOR A CLASS OF NON-NEWTONIAN FLUIDS*

DIETER BOTHE† AND JAN PRÜSS‡

Abstract. Local-in-time well-posedness of the initial-boundary value problem for a class of non-Newtonian Navier–Stokes problems on domains with compact C^{3-}-boundary is proven in an L_p-setting for any space dimension $n \geq 2$. The stress tensor is assumed to be of the generalized Newtonian type, i.e., $\mathcal{S} = 2\mu(|\mathcal{E}|_2^2)\mathcal{E} - \pi I$, $\mathcal{E} = \frac{1}{2}(\nabla u + \nabla u^{\mathsf{T}})$, where $|\mathcal{E}|_2^2 = \sum_{i,j=1}^n \varepsilon_{ij}^2$ denotes the Hilbert–Schmidt norm of the rate of strain tensor $\mathcal{E}$. The viscosity function $\mu \in C^{2-}(\mathbb{R}_+)$ is subject only to the condition $\mu(s) > 0$, $\mu(s) + 2s\mu'(s) > 0$, $s \geq 0$, which for the standard power-law–like function $\mu(s) = \mu_0(1+s)^{\frac{d-2}{2}}$ merely means $\mu_0 > 0$ and $d \geq 1$. This result is based on maximal regularity theory for a suitable linear problem and a contraction argument.

Key words. maximal regularity, non-Newtonian fluids, L_p-theory, $\mathcal{R}$-boundedness, initial boundary value problems, power-law–like fluids, strong ellipticity, Lopatinskii–Shapiro condition

AMS subject classifications. 35Q35, 76D03, 35K50, 47A60, 35K90

DOI. 10.1137/060663635

1. Introduction. In this paper, we study boundary value problems for the Navier–Stokes system of non-Newtonian fluids. By this we mean the following problem:

$$
\begin{aligned}
&\partial_t(\rho u) + \operatorname{div}(\rho u \otimes u) = \operatorname{div} \mathcal{S},\\
&\mathcal{S} = \mu[\nabla u + (\nabla u)^{\mathsf{T}}] - \pi I,\\
&\operatorname{div} u = 0, \quad t > 0,\ x \in G,\\
&u(0,x) = u_0(x), \quad x \in G.
\end{aligned}
\tag{1.1}
$$

Here $G \subset \mathbb{R}^n$, $n \geq 2$, denotes the domain occupied by the fluid, and $\Gamma = \partial G$ means the boundary of G. We assume that Γ is compact and of class C^{3-}. Throughout, $u(t,x)$ means the velocity field of the fluid, $\pi(t,x)$ the pressure, and $\mathcal{S}(t,x)$ the stress tensor. The numbers $\rho > 0$, $\mu > 0$ represent the density and viscosity of the fluid, respectively. We assume that $\rho > 0$ is constant, and without loss of generality (w.l.o.g.) $\rho = 1$.

On the other hand, the (nonconstant) viscosity μ will be taken of *generalized Newtonian type*, i.e.,

$$
\mu = \mu(|\mathcal{E}|_2^2), \quad \mathcal{E} = \frac{1}{2}\,[\nabla u + (\nabla u)^{\mathsf{T}}]. \tag{1.2}
$$

Here $\mathcal{E} = (\varepsilon_{ij})$ denotes the rate of strain tensor and

$$
|\mathcal{E}|_2^2 = \sum_{i,j=1}^n \varepsilon_{ij}^2
$$

*Received by the editors June 23, 2006; accepted for publications (in revised form) January 2, 2007; published electronically June 5, 2007.
http://www.siam.org/journals/sima/39-2/66363.html
†Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen, Lehrstuhl für Mathematik, Center for Computational Engineering Science, Pauwelsstr. 19, 52074 Aachen, Germany (bothe@mathcces.rwth-aachen.de).
‡Martin-Luther-Universität Halle-Wittenberg, Fachbereich Mathematik und Informatik, Institut für Analysis, Theodor-Lieser-Str. 5, 06120 Halle, Germany (jan.pruess@mathematik.uni-halle.de).

its Hilbert–Schmidt norm. Note that the first invariant of $\mathcal{E}$, namely $\operatorname{tr}\mathcal{E} = \operatorname{div} u$, is zero, and hence the Hilbert–Schmidt norm of $\mathcal{E}$ coincides with the second invariant of $\mathcal{E}$ up to a constant factor. It is believed that many *isotropic* fluids which are not subject to viscoelastic memory effects can be described by such a material law. A prominent example is the *power-law fluid* of Ostwald and de Waele,

$$\mu(s) = ms^{d/2-1},$$

where $d \geq 1$ and $m > 0$. Due to the obvious defect of this constitutive equation at $s = 0$, it is common practice to use modified power-laws like the truncated power law of Spriggs, the Eyring model, the Carreau model, etc.; see Chapter 5 of [2].

A standard model in the mathematical literature is

$$\mu(s) = \mu_0(1+s)^{(d-2)/2}, \quad s \geq 0,$$

where $d \geq 1$ and $\mu_0 > 0$. The case $d = 2$ corresponds to the Newtonian case.

At this point it should be observed that such constitutive laws are just the most basic ones beyond the Newtonian case. Such simple stress-strain relations cannot account for many exciting phenomena observed in the world of non-Newtonian fluids; they can only be considered as a starting point. Nevertheless, there has been—and there still is—considerable interest in generalized Newtonian fluids, in the mathematical as well as in the engineering communities, since a nonconstant viscosity that depends on the shear rate covers the most important deviation from the linear behavior for many macromolecular fluids; see [2].

Three natural boundary conditions are of interest, namely the nonslip condition

$$u(t,x) = 0, \quad t > 0,\ x \in \Gamma_0,$$

the pure slip case

$$\mathcal{S}\nu - (\nu|\mathcal{S}\nu)\nu = 0, \qquad (u|\nu) = 0, \quad t > 0,\ x \in \Gamma_s,$$

and the pure Neumann condition

$$\mathcal{S}\nu = 0, \quad x \in \Gamma_n.$$

Here we decompose $\Gamma = \partial G$ disjointly as $\Gamma = \Gamma_0 \cup \Gamma_s \cup \Gamma_n$, where each set Γ_j is open and closed in Γ. The outer normal of G at $x \in \Gamma$ is denoted by $\nu = \nu(x)$. We emphasize that each Γ_j may be empty.

Already in 1890, Schwedoff realized a nonlinearity in the torque-angular velocity relationship for colloidal gelatin solutions in a Couette device; cf. [24] for further historical information about rheology. It was Ladyzhenskaya who in 1964 suggested studying the partial differential equations for incompressible flows of non-Newtonian fluids in a mathematically rigorous way. Due to her also are the first results on weak solutions, dating back to 1969 [14]. More recently, this problem has been studied intensively by the group around Nečas, leading to many interesting results. This group employs various notions of weak solutions, which have a number of advantages in requiring less regularity of the data or allowing one to handle certain degenerate or singular viscosities; they also yield results that hold globally in time and for large data. We do not want to comment on all of the literature here, but concerning weak solutions we refer to the survey article by Málek and Rajagopal [17], and to the papers [6, 8, 9, 12, 13, 16, 18, 19, 20]. However, in any of the papers in the literature there are some restrictions on the exponent d. In space dimension 3, the least restrictive condition we are aware of is $d > 7/5$; this concerns local-in-time existence and uniqueness of strong solutions. This covers the case of *shear-thickening* fluids well, but is quite restrictive or even does not apply in the *shear-thinning* case $d < 2$. From the engineering literature one can infer that $d \geq 1$ is the range of physical interest. A first step toward a larger range of admissible exponents has been achieved in Wielage [25] for the full-space case $G = \mathbb{R}^n$.

We also want to mention that for small initial values u_0 there are well-posedness results for very general stress-strain relations (cf. [1]), and even analyticity of strong solutions has been proved recently; cf. [7]. For such initial values it has already been shown in [1] that for bounded domains unique strong solutions exist which are global in time.

In this paper we show that, in fact, for $d \geq 1$ the problem is well-posed, locally in time, in *any* space dimension without restriction on the size of the initial velocity u_0.

Another aspect are the boundary conditions. Mainly Dirichlet conditions or periodic boundary conditions have been employed in the non-Newtonian case, so far. One exception is [8], where slip boundary conditions are treated. In the engineering literature, in the non-Newtonian case the no-slip condition is no longer well accepted, as the extensive and thorough discussion in [15] shows.

The consideration of pure slip conditions at the boundary requires a more careful inspection of the form of the proper linearization of the Navier–Stokes system, which is a key point of our analysis. In the end, it is this analysis which allows us to generalize previous results to the situation $d \geq 1$ in all space dimensions.

Our tools will be results on maximal L_p-regularity of the following linear problem, which we call the *generalized Stokes problem*:

$$
\begin{aligned}
\partial_t v + \mathcal{A}(x,D)v + \nabla q &= f(t,x), \quad t > 0,\ x \in G,\\
\operatorname{div} v &= g(t,x), \quad t > 0,\ x \in G,\\
\mathcal{B}_j(x,D)v &= h_j(t,x), \quad t > 0,\ x \in \Gamma_j,\ j = 0, s, n,\\
v(0,x) &= v_0(x), \quad x \in G.
\end{aligned}
\tag{1.3}
$$

Here $\mathcal{A}(x,D)$ is a second order differential operator with continuous top order coefficients acting on $\mathbb{C}^n$-valued functions; $\mathcal{B}_j(x,D)$ denote boundary operators of order zero or one, which correspond to the boundary conditions described above; and the data f, g, h_j, u_0 are given. Assuming *strong normal ellipticity* (cf. section 3) of the initial-boundary value problem without pressure and divergence condition, we show that (1.3) is well-posed and has maximal regularity in L_p as well.

It should be mentioned here that, recently, this problem has also been investigated by Solonnikov [23]. However, in his paper only Dirichlet conditions are considered, and—more importantly—only sufficient conditions for solvability are presented. Moreover, additional regularity assumptions on the coefficients a_{ij}^{kl} are imposed, which are too limiting for solving the nonlinear problem.

Based on our maximal regularity result for the generalized Stokes problem, we are able to prove local well-posedness for the nonlinear problem (1.1) by a contraction argument. This way we are able to extend results in the literature for the L_2 case to L_p, and at the same time we improve the conditions on μ in (1.2) to

$$
\mu(s) > 0 \quad \text{and} \quad \mu(s) + 2s\mu'(s) > 0 \quad \text{for all } s \geq 0. \tag{1.4}
$$

In particular, for the model $\mu(s) = \mu_0(1+s)^{(d-2)/2}$ this follows by $\mu_0 > 0$ and $d \geq 1$. Note that the second condition in (1.4) means that the function $s \mapsto \mu(s)\sqrt{s}$ is strictly increasing. Physically, this is a natural assumption, meaning that the viscous stress given by

$$|S_0|_2 = \mu(|\mathcal{E}|_2^2)|\mathcal{E}|_2$$

is increasing with increasing rate of strain $|\mathcal{E}|_2$.

The linear generalized Stokes problem (1.3) is taken here in a fairly general form. Therefore our approach can also be applied to constitutive laws more general than the generalized Newtonian law considered here, in particular to nonisotropic ones. The only restriction we impose is strong normal ellipticity of the proper linearization, which implies parabolicity of the nonlinear problem. More precisely, we can easily generalize our results to the case

$$\mathcal{S}_{ij} = -\pi\delta_{ij} + \partial_{\varepsilon_{ij}}\psi(u,\mathcal{E}),$$

where $\psi : \mathbb{R}^n \times \mathcal{S}ym(n) \to \mathbb{R}$ is of class C^2 and $\mathcal{S}ym(n)$ is the space of real-valued symmetric matrices of dimension n, provided that $a_{ij}^{kl} = \partial_{\varepsilon_{ik}}\partial_{\varepsilon_{jl}}\psi(u,\mathcal{E})$ is strongly normally elliptic as defined in section 3. Roughly speaking, this means that the scalar function ψ is uniformly convex in its second variable.

The boundary conditions considered here seem to be the most interesting ones, but still generalizations are possible. First, an engineer might be confronted with a fast flow through a tube in direction of coordinate z. Then he would of course require the condition $(u|\nu) = 0$, but he would also neglect circumferential convection at the boundary, i.e., $u_\phi = 0$. However, because of fast flow he would omit friction in the axial direction z, i.e., $\partial_z u_z = 0$. Such boundary conditions are covered by our analysis as well, but we will not discuss them here. Secondly, there might be tangential friction which leads to a boundary condition of the form

$$S\nu - (S\nu|\nu)\nu = -\gamma(u - (u|\nu)\nu),$$

where $\gamma \geq 0$ is an empirical constant. This is known as the *Navier condition* in the literature; cf. [15]. We will not deal with it here since the additional term is a lower order perturbation, which even stabilizes to some extent but does not affect the analysis of local well-posedness and regularity.

The plan for this paper is as follows. Section 2 contains the statement of the main result of this paper, namely well-posedness in the L_p-sense. The discussion of normal ellipticity of the linearized problem in section 3 leads to the condition (1.4) on the function $\mu(s)$. The main result on the generalized Stokes problem is presented in section 4. Sections 5 and 6 deal with constant coefficient generalized Stokes problems on $\mathbb{R}^n$ and on the half-space $\mathbb{R}^n_+$, respectively. By means of localization and straightening of the boundary, these results are extended to general domains with compact C^{3-}-boundary in section 7. The contraction argument is carried out in section 8 for the homogeneous slip and nonslip cases, and in section 9 for general inhomogeneous boundary conditions. In an appendix (section 10), we summarize some auxiliary results on optimal regularity for parabolic systems used in previous sections.

2. Main results. With the definition of $\mathcal{S} := \mathcal{S}_0 - \pi I$ we have $\mathcal{S}_0 = 2\mu(|\mathcal{E}|_2^2)\mathcal{E}$,

$$\begin{aligned}[\operatorname{div}\mathcal{S}_0]_i &= \mu(|\mathcal{E}|_2^2)\sum_{k=1}^{n}(\partial_k^2 u_i + \partial_i\partial_k u_k) + 4\mu'(|\mathcal{E}|_2^2)\sum_{j,k,l=1}^{n}\varepsilon_{ij}\varepsilon_{kl}\partial_j\varepsilon_{kl}\\ &= \mu(|\mathcal{E}|_2^2)\sum_{k=1}^{n}(\partial_k^2 u_i + \partial_i\partial_k u_k) + 4\mu'(|\mathcal{E}|_2^2)\sum_{j,k,l=1}^{n}\varepsilon_{ik}(u)\varepsilon_{jl}(u)\partial_k\partial_l u_j\\ &= \sum_{j,k,l=1}^{n} a_{ij}^{kl}(u)\partial_k\partial_l u_j.\end{aligned}$$

Here we have set

$$a_{ij}^{kl}(u) = \mu(|\mathcal{E}|_2^2)(\delta_{kl}\delta_{ij} + \delta_{il}\delta_{jk}) + 4\mu'(|\mathcal{E}|_2^2)\varepsilon_{ik}(u)\varepsilon_{jl}(u). \tag{2.1}$$

Observe that $a_{ij}^{kl}(u)$ are real and that the symmetries $a_{ij}^{kl} = a_{ji}^{lk} = a_{kl}^{ij} = a_{kj}^{il} = a_{il}^{kj}$ are valid.

Define the quasi-linear differential operator $\mathcal{A}(u, D)$ as

$$\mathcal{A}(u, D) = \sum_{k,l=1}^{n} a_{ij}^{kl}(u) D_k D_l,$$

where $D_k = -i\partial_k$. If the function $u \in W_p^{2-2/p}(G;\mathbb{R}^n)$ is known and $p > n+2$, then by the Sobolev embedding $W_p^{2-2/p}(G;\mathbb{R}^n) \hookrightarrow BUC^1(G;\mathbb{R}^n)$ the coefficients of the differential operator $\mathcal{A}(x, D) = \mathcal{A}(u(x), D)$ are uniformly continuous and

$$[\mathcal{A}(\infty, D)v]_i = -\mu(0)\sum_{k=1}^{n}(\partial_k^2 v_i + \partial_i\partial_k v_k)$$

in case G is unbounded, since $|Du(x)| \to 0$ as $|x| \to \infty$. Denoting the boundary operators on Γ_j by $\mathcal{B}_j(u, D)$, problem (1.1), complemented by the boundary conditions discussed above, can be rewritten as

$$\begin{aligned}\partial_t u + \mathcal{A}(u, D)u + \nabla\pi = f(u), &\quad t \in J,\ x \in G,\\ \operatorname{div} u = 0, &\quad t \in J,\ x \in G,\\ \mathcal{B}_j(u, D)u = 0, &\quad t \in J,\ x \in \Gamma_j,\ j = 0, s, n,\\ u|_{t=0} = u_0. &\end{aligned} \tag{2.2}$$

Here the boundary operators $\mathcal{B}_j$ will be $\mathcal{B}_0(u, D)u = u$ on Γ_0, while on Γ_n they will be $S\nu = 0$, i.e.,

$$\mathcal{B}_n(u, D)(u, \pi) = 2\mu(|\mathcal{E}|_2^2)\mathcal{E}\nu - \pi\nu.$$

Similarly, on Γ_s we obtain

$$\mathcal{B}_s(u, D)u = ((u|\nu), 2\mu(|\mathcal{E}|_2^2)[\mathcal{E}\nu - (\mathcal{E}\nu|\nu)\nu]).$$

The nonlinearity $f(u)$ is the convective term $f(u) = -u \cdot \nabla u$.

It is convenient to introduce the projection $\mathcal{P}$ defined by

$$\mathcal{P}w(x) = w(x) - (w(x)|\nu(x))\nu(x), \quad x \in \Gamma,$$

which projects a vector field w on the boundary Γ of G to its tangential part. $\mathcal{P}$ preserves $W^s_p(\Gamma;\mathbb{R}^n)$ for each $s\in[0,2]$, since G is assumed to be of class C^{3-}, and hence $\nu\in C^{2-}(\Gamma;\mathbb{R}^n)$; here a domain is said to be of class C^{k-} if its boundary is locally parameterized by functions which are in C^{k-1} with a Lipschitz continuous $(k-1)$th derivative. Obtaining this property for $\mathcal{P}$ is the main reason that we impose this particular regularity of the boundary. In what follows, we as usual let $\dot H^1_p(G)$ denote the completion of $H^1_p(G)$ w.r.t. the L_p-gradient norm, and $\frac{1}{p}+\frac{1}{p'}=1$.

Our main result is the following.

THEOREM 2.1. *Let $G\subset\mathbb{R}^n$ be a domain with compact boundary $\Gamma=\partial G$ of class C^{3-}, where $\Gamma=\Gamma_0\cup\Gamma_s\cup\Gamma_n$ with disjoint, open and closed Γ_j; let $n+2<p<\infty$; and assume that $\mu\in C^{2-}(\mathbb{R}_+)$ is such that*

$$\mu(s)>0 \quad \textit{and} \quad \mu(s)+2s\mu'(s)>0 \quad \textit{for all } s\geq 0. \tag{2.3}$$

Then for each $u_0\in W^{2-2/p}_p(G;\mathbb{R}^n)$ satisfying the compatibility conditions

$$\operatorname{div} u_0|_G=0,\quad u_0|_{\Gamma_0}=0,\quad (u_0|\nu)|_{\Gamma_s}=0,\quad \mathcal{P}\mathcal{E}(u_0)\nu|_{\Gamma_s\cup\Gamma_n}=0, \tag{2.4}$$

there is a unique solution (u,π) of (2.2) on a maximal time interval $[0,t^*(u_0))$. The solution is in the maximal regularity class

$$u\in H^1_p(J,L_p(G;\mathbb{R}^n))\cap L_p(J,H^2_p(G;\mathbb{R}^n)),\quad \pi\in L_p(J;\dot H^1_p(G)),$$

additionally with

$$\pi|_{\Gamma_n}\in W^{1/2-1/2p}_p(J;L_p(\Gamma_n))\cap L_p(J;W^{1-1/p}_p(\Gamma_n)),$$

for each interval $J=[0,a]$ with $a<t^*(u_0)$. The maximal time $t^*(u_0)$ is characterized by the property

$$\textit{if } t^*(u_0)<\infty, \quad \textit{then} \quad \lim_{t\to t^*(u_0)} u(t) \ \textit{ does not exist in } \ W^{2-2/p}_p(G;\mathbb{R}^n).$$

The solution map $u_0\mapsto u$ generates a local semiflow on

$$X_{pc}:=\{v\in W^{2-2/p}_p(G;\mathbb{R}^n):\ v \textit{ satisfies } (2.4)\},$$

the natural phase space for the problem in the L_p-setting.

Note that in case $\Gamma_n=\emptyset$ uniqueness of π means uniqueness up to a constant. The remainder of this paper deals with the proof of this result. In fact, in section 9 we will even prove well-posedness for the completely inhomogeneous problem.

3. Linearization and strong normal ellipticity. Consider the differential operator $\mathcal{A}(x,D)$ acting on $\mathbb{C}^n$-valued functions as

$$[\mathcal{A}(x,D)v(x)]_i=\sum_{j,k,l=1}^n a^{kl}_{ij}(x)D_kD_lv_j(x),\quad x\in G.$$

Its symbol is defined as the matrix

$$\mathcal{A}(x,\xi)=\sum_{k,l=1}^n a^{kl}_{ij}(x)\xi_k\xi_l.$$

Such a second order differential operator is called *normally elliptic* if the spectrum $\sigma(\mathcal{A}(x,\xi))$ of the matrix $\mathcal{A}(x,\xi)$ is entirely contained in the open right half-plane $\mathbb{C}_+$ for each $x \in \overline{G}$, $\xi \in \mathbb{R}^n$, $|\xi| = 1$; cf. the appendix. It is called *strongly elliptic* if the numerical range of $\mathcal{A}(x,\xi)$ is a subset of $\mathbb{C}_+$ for each $x \in \overline{G}$, $\xi \in \mathbb{R}^n$, $|\xi| = 1$. Thus strongly elliptic means that there is a constant $c > 0$ such that

$$\operatorname{Re}(\mathcal{A}(x,\xi)\eta|\eta) \geq c \quad \text{ for all } \xi \in \mathbb{R}^n,\ \eta \in \mathbb{C}^n,\ |\xi| = |\eta| = 1,\ x \in G.$$

Here $(\cdot|\cdot)$ denotes the inner product on $\mathbb{C}^n$. Obviously, a strongly elliptic differential operator $\mathcal{A}(x,D)$ is also normally elliptic, but the converse is false, in general.

Let us check what this means in our situation, where $\mathcal{A}(x,D) := \mathcal{A}(u,D)$ is defined as in section 2, with a given function $u \in W_p^{2-2/p}(G;\mathbb{R}^n)$. A simple computation shows for $\xi \in \mathbb{R}^n$, $\eta \in \mathbb{C}^n$, $|\xi| = |\eta| = 1$, by the definition of a_{ij}^{kl} in (2.1) and using the sum convention,

$$a_{ij}^{kl}\xi_l\eta_j = \mu(\xi_k\eta_i + \xi_i\eta_k) + 4\mu'\varepsilon_{ik}\varepsilon_{jl}\xi_l\eta_j.$$

Using the symmetry of $\mathcal{E}(u)$, this yields

$$a_{ij}^{kl}\xi_l\eta_j = 2\mu c_{ik} + 4\mu'\varepsilon_{ik}\overline{((\mathcal{E}|\mathcal{C}))},$$

where $\mathcal{C} = (c_{ik}) = \frac{1}{2}(\xi\otimes\eta + \eta\otimes\xi)$ and $((\cdot|\cdot))$ means the inner product in $\mathbb{C}^{n\times n}$. Next observe that this matrix is symmetric; hence we obtain

$$\begin{aligned}(\mathcal{A}(x,\xi)\eta|\eta) &= a_{ij}^{kl}\xi_l\eta_j\xi_k\overline{\eta}_i \\ &= 2\mu c_{ik}\xi_k\overline{\eta}_i + 4\mu'\varepsilon_{ik}\xi_k\overline{\eta}_i\overline{((\mathcal{E}|\mathcal{C}))} \\ &= 2\mu|\mathcal{C}|_2^2 + 4\mu'|((\mathcal{E}|\mathcal{C}))|^2 \\ &= \mu(|\xi|^2|\eta|^2 + |(\xi|\eta)|^2) + 4\mu'|(\mathcal{E}\xi|\eta)|^2.\end{aligned}$$

Notice that $(\mathcal{A}(x,\xi)\eta|\eta)$ is real, for each $x \in G$, $\xi \in \mathbb{R}^n$, $\eta \in \mathbb{C}^n$. To obtain a necessary condition for strong ellipticity, choose ξ as an eigenvector of $\mathcal{E}$ and η perpendicular to ξ. This shows that the condition $\mu(s) > 0$ for each $s \geq 0$ is necessary for strong ellipticity. Obviously this condition is also sufficient in the case $\mu'(|\mathcal{E}|_2^2) \geq 0$. So suppose that $\mu'(|\mathcal{E}|_2^2) < 0$. Then the Cauchy–Schwarz inequality implies

$$(\mathcal{A}(x,\xi)\eta|\eta) \geq 2\mu|\mathcal{C}|_2^2 + 4\mu'|\mathcal{E}|_2^2|\mathcal{C}|_2^2 \geq c|\mathcal{C}|_2^2,$$

provided that we have

$$\mu(s) > 0 \quad \text{ and } \quad \mu(s) + 2s\mu'(s) > 0 \quad \text{ for all } s \geq 0.$$

If $|\mathcal{C}|_2 = 0$, then $(\mathcal{C}\xi|\xi) = 0$, which means $2|\xi|^2(\eta|\xi) = 0$; hence $(\eta|\xi) = 0$ since $|\xi| = 1$ by assumption. But this in turn yields $2\mathcal{C}\xi = \eta|\xi|^2$, and hence the contradiction $\eta = 0$. Thus strong ellipticity is implied by (2.3), independently of the choice of $u \in W_p^{2-2/p}(G;\mathbb{R}^n)$.

We note that the condition $\mu(s) + 2s\mu'(s) > 0$ for $s > 0$ is also necessary, if one allows for *all* symmetric $\mathcal{E}$; choose, e.g., $\mathcal{E} = \operatorname{diag}(\sqrt{s}, 0, \ldots, 0)$ to see this.

Let us next discuss the Lopatinskii–Shapiro condition for the parabolic problem without pressure and incompressibility. For this purpose let $\xi \in \mathbb{R}^n$, $\nu \in \mathbb{R}^n$, $|\nu| = 1$, $(\xi|\nu) = 0$, $\operatorname{Re}\lambda \geq 0$. For convenience, we drop the x-dependence. Consider the linear ode-problem with constant coefficients

$$\lambda w(y) + \mathcal{A}(\xi - \nu D_y)w(y) = 0, \quad y > 0, \quad \mathcal{B}(\xi - \nu D_y)w(0) = 0.$$

The Lopatinskii–Shapiro condition requires that any solution $w \in C_0(\mathbb{R}_+;\mathbb{C}^n)$ of this problem be zero; cf. the appendix. Note that strong ellipticity already implies $w \in H^2_2(\mathbb{R}_+;\mathbb{C}^n)$.

To verify the Lopatinskii–Shapiro condition, we form the inner product of the equation with a solution w, integrate over $\mathbb{R}_+$, and integrate by parts to obtain the result

$$\begin{aligned}
0 &= \lambda|w|_2^2 + \int_0^\infty (\mathcal{A}(\xi - \nu D_y)w(y)|w(y))dy \\
&= \lambda|w|_2^2 + \int_0^\infty \sum_{i,j,k,l=1}^n a_{ij}^{kl}(\xi_l - \nu_l D_y)w_j(y)\overline{(\xi_k - \nu_k D_y)w_i(y)}dy \\
&\quad - i \sum_{i,j,k,l=1}^n \nu_k a_{ij}^{kl}(\xi_l - \nu_l D_y)w_j(0)\overline{w_i(0)} \\
&= \lambda|w|_2^2 + \int_0^\infty \sum_{i,j,k,l=1}^n a_{ij}^{kl}(\xi_l - \nu_l D_y)w_j(y)\overline{(\xi_k - \nu_k D_y)w_i(y)}dy,
\end{aligned}$$

provided that the boundary term vanishes. This is, of course, the case if w is subject to Dirichlet conditions, i.e., on Γ_0, or in case of Neumann conditions on Γ_n,

$$[\mathcal{B}(\xi - \nu D_y)w]_i(0) = \sum_{j,k,l=1}^n \nu_k a_{ij}^{kl}(\xi_l - \nu_l D_y)w_j(0) = 0.$$

But it is also true in the case $(w|\nu) = 0$ and

$$\mathcal{B}(\xi - \nu D_y)w(0) - (\mathcal{B}(\xi - \nu D_y)w(0)|\nu)\nu = 0,$$

which corresponds to the pure slip case, i.e., to Γ_s.

Next consider the generalized Stokes problem (1.3). The same arguments yield

$$\begin{aligned}
0 &= \lambda|w|_2^2 + \int_0^\infty (\mathcal{A}(\xi - \nu D_y)w(y)|w(y))dy + \int_0^\infty ((i\xi - \nu\partial_y)\pi(y)|w(y))dy \\
&= \lambda|w|_2^2 + \int_0^\infty \sum_{i,j,k,l=1}^n [a_{ij}^{kl}(\xi_l - \nu_l D_y)w_j\overline{(\xi_k - \nu_k D_y)w_i} + \pi((i\xi + \nu\partial_y)|w)]dy \\
&\quad - i \sum_{i,j,k,l=1}^n [\nu_k a_{ij}^{kl}(\xi_l - \nu_l D_y)w_j(0) + i\nu_i\pi(0)]\overline{w_i(0)} \\
&= \lambda|w|_2^2 + \int_0^\infty \sum_{i,j,k,l=1}^n a_{ij}^{kl}(\xi_l - \nu_l D_y)w_j(y)\overline{(\xi_k - \nu_k D_y)w_i(y)}dy,
\end{aligned}$$

provided that the boundary term vanishes again, and since $((i\xi + \nu\partial_y)|w) = 0$. Once more, this is certainly true if w is subject to Dirichlet conditions on Γ_0, or in case of the relevant Neumann condition for this situation on Γ_n,

$$\mathcal{B}(\xi - \nu D_y)(w(0), \pi(0)) = \left(\sum_{j,k,l=1}^n \nu_k a_{ij}^{kl}(\xi_l - \nu_l D_y)w_j(0) + i\pi(0)\nu_i \right)_{i=1}^n = 0.$$

However, as before, it is also true in the case $(w|\nu) = 0$ and

$$\mathcal{B}(\xi - \nu D_y)w(0) - (\mathcal{B}(\xi - \nu D_y)w(0)|\nu)\nu = 0,$$

i.e., for the pure slip case on Γ_s.

Thus in all cases we have the identity

$$0 = \lambda|w|_2^2 + \int_0^\infty \sum_{i,j,k,l=1}^n a_{ij}^{kl}(\xi_l - \nu_l D_y)w_j(y)\overline{(\xi_k - \nu_k D_y)w_i(y)}dy. \tag{3.1}$$

To satisfy the Lopatinskii–Shapiro condition, we would like to conclude from this identity that $w \equiv 0$. For this purpose we introduce the following notation.

Definition 3.1. *$\mathcal{A}(D)$ is called* strongly normally elliptic *if it is strongly elliptic and*

$$\mathrm{Re} \sum_{i,j,k,l=1}^n a_{ij}^{kl}(\xi_l u_j - \nu_l v_j)\overline{(\xi_k u_i - \nu_k v_i)} > 0$$

for all $\xi, \nu \in \mathbb{R}^n$, $|\xi| = |\nu| = 1$, $(\xi|\nu) = 0$, $u, v \in \mathbb{C}^n$, $\mathrm{Im}\,(u|v) \neq 0$.

Let us show that this condition implies $w \equiv 0$, for each $w \in H_2^1(\mathbb{R}_+; \mathbb{C}^n)$ satisfying (3.1) with $\mathrm{Re}\,\lambda \geq 0$. In fact, taking real parts in (3.1), we obtain

$$\int_0^\infty \mathrm{Re} \sum_{i,j,k,l=1}^n a_{ij}^{kl}(\xi_l w_j(y) - \nu_l D_y w_j(y))\overline{(\xi_k w_i(y) - \nu_k D_y w_i)(y)}dy = 0,$$

and hence

$$\frac{d}{dy}|w(y)|^2 = 2\,\mathrm{Re}\left(\frac{d}{dy}w(y)|w(y)\right) = -2\,\mathrm{Im}\,(D_y w(y)|w(y)) = 0, \quad y > 0.$$

However, this implies that $|w(y)|$ is constant on $\mathbb{R}_+$, and hence $w(y) = 0$ on $\mathbb{R}_+$. So strong normal ellipticity implies the Lopatinskii–Shapiro condition for all three types of boundary operators introduced above.

Let us check what strong normal ellipticity means for the generalized Stokes problem with a_{ij}^{kl} from (2.1). With $\mathcal{D} = \xi \otimes u + \nu \otimes v$, $\mathcal{C} = \frac{1}{2}(\mathcal{D} + \mathcal{D}^T)$, and using sum convention again, we have

$$a_{ij}^{kl} d_{lj} = \mu(d_{ik} + d_{ki}) + 4\mu' \varepsilon_{ik}\varepsilon_{jl} d_{lj} = 2\mu c_{ik} + 4\mu' \varepsilon_{ik}\overline{((\mathcal{E}|\mathcal{C}))},$$

by symmetry of $\mathcal{E}$. Note that the resulting matrix is symmetric. This yields

$$a_{ij}^{kl} d_{lj}\overline{d}_{ki} = 2\mu|\mathcal{C}|_2^2 + 4\mu'|((\mathcal{E}|\mathcal{C}))|^2;$$

hence this expression is real and

$$a_{ij}^{kl} d_{lj}\overline{d}_{ki} \geq 2\min\{\mu, \mu + 2\mu'|\mathcal{E}|_2^2\}|\mathcal{C}|_2^2.$$

Condition (2.3) implies $a_{ij}^{kl} d_{lj}\overline{d_{ki}} \geq 0$, and $\mathcal{C} = 0$ in case of equality. This yields $\mathcal{C}\xi = 0$ as well as $\mathcal{C}\nu = 0$, and leads to the relations

$$u + (v|\xi)\nu + (u|\xi)\xi = v + (v|\nu)\nu + (u|\nu)\xi = 0.$$

Taking the inner product with ξ (resp., ν), we obtain $(u|\xi) = (v|\nu) = 0$ and $(u|\nu) + (v|\xi) = 0$. We may then conclude

$$u = \alpha\nu, \quad v = -\alpha\xi,$$

and in particular $(u|v) = 0$.

Thus, the Lopatinskii–Shapiro condition also holds for the generalized Stokes problem, as far as $\xi \neq 0$. We want to stress that $\pi = 0$ follows from the equations as long as $\xi \neq 0$. However, if $\xi = 0$, then we obtain $\pi = 0$ from the Neumann boundary condition, but only $\pi = const$ in case of slip or no-slip conditions. This reflects the nonuniqueness of the pressure for no-slip and slip conditions.

4. The generalized Stokes problem on domains. Let $G \subset \mathbb{R}^n$ be a domain with compact boundary $\Gamma = \partial G$, and $J = [0, a]$. We decompose Γ disjointly as $\Gamma = \Gamma_0 \cup \Gamma_s \cup \Gamma_n$, where Γ_j is open and closed in Γ. In this section we consider the fully inhomogeneous initial-boundary value problem

$$\begin{aligned} \partial_t v + \mathcal{A}(t,x,D)v + \nabla q &= f, \quad t \in J,\ x \in G, \\ \operatorname{div} v &= g, \quad t \in J,\ x \in G, \\ \mathcal{B}_j(t,x,D)v &= h_j, \quad t \in J,\ x \in \Gamma_j,\ j = 0, s, n, \\ v|_{t=0} &= v_0. \end{aligned} \tag{4.1}$$

Here $\mathcal{A}(t,x,D) = \sum_{k,l=1}^n \mathcal{A}^{kl}(t,x) D_k D_l$ denotes a strongly normally elliptic differential operator, and the boundary operator $\mathcal{B}_j(t,x,D)$ is $\mathcal{B}_0(t,x,D)v = v$ on Γ_0, while on Γ_n it is given by

$$[\mathcal{B}_n(t,x,D)(v,q)]_i = \sum_{j,k,l=1}^n a_{ij}^{kl}(t,x)\nu_k D_l v_j + iq\nu_i,$$

and on Γ_s it is

$$[\mathcal{B}_s(t,x,D)v]_i = \left((v|\nu),\ \sum_{j,k,l=1}^n a_{ij}^{kl}(t,x)\nu_k D_l v_j - \left[\sum_{r,j,k,l=1}^n a_{rj}^{kl}(t,x)\nu_k\nu_r D_l v_j \right] \nu_i \right).$$

We suppose that the coefficients $\mathcal{A}^{kl}(t,x)$ of $\mathcal{A}(t,x,D)$ are continuous on $J \times \overline{G}$ and have limits $\mathcal{A}^{kl}(t,\infty)$ at $x = \infty$, uniformly in $t \in J$, in case G is unbounded, and that $\mathcal{A}(t,\infty,D)$ is strongly elliptic as well. In addition, we assume

$$a_{ij}^{kl} \in W_s^{1/2-1/2p}(J; L_r(\Gamma_s \cup \Gamma_n)) \cap L_s(J; W_r^{1-1/p}(\Gamma_s \cup \Gamma_n)),$$

for some $r, s \geq p$ such that $\frac{1}{s} + \frac{n-1}{2r} < 1 - \frac{1}{p}$.

Suppose that (4.1) has a solution in the class

$$v \in H_p^1(J; L_p(G;\mathbb{C}^n)) \cap L_p(J; H_p^2(G;\mathbb{C}^n)), \quad q \in L_p(J; \dot{H}_p^1(G)).$$

Then, following the paper of Denk, Hieber, and Prüss [5], we may conclude $f \in L_p(J \times G;\mathbb{C}^n)$, $g \in L_p(J; H_p^1(G))$, $v_0 \in W_p^{2-2/p}(G;\mathbb{C}^n)$, and the compatibility $\operatorname{div} v_0 = g|_{t=0}$. Furthermore,

$$v|_\Gamma \in W_p^{1-1/2p}(J; L_p(\Gamma;\mathbb{C}^n)) \cap L_p(J; W_p^{2-1/p}(\Gamma;\mathbb{C}^n))$$

and

$$\nabla v|_\Gamma \in W_p^{1/2-1/2p}(J;L_p(\Gamma;\mathbb{C}^{n\times n}))\cap L_p(J;W_p^{1-1/p}(\Gamma;\mathbb{C}^{n\times n})).$$

This implies

$$\begin{aligned} h_0 &\in W_p^{1-1/2p}(J;L_p(\Gamma_0;\mathbb{C}^n))\cap L_p(J;W_p^{2-1/p}(\Gamma_0;\mathbb{C}^n)),\\ h_s &\in W_p^{1/2-1/2p}(J;L_p(\Gamma_s;\mathbb{C}^n))\cap L_p(J;W_p^{1-1/p}(\Gamma_s;\mathbb{C}^n)),\\ h_{j2} := (u|\nu) &\in W_p^{1-1/2p}(J;L_p(\Gamma_j))\cap L_p(J;W_p^{2-1/p}(\Gamma_j)), \qquad j=0,s,\\ h_n &\in W_p^{1/2-1/2p}(J;L_p(\Gamma_n;\mathbb{C}^n))\cap L_p(J;W_p^{1-1/p}(\Gamma_n;\mathbb{C}^n)),\end{aligned}$$

and with $h_{s1} := \mathcal{P}h_s$ the compatibility conditions $h_0|_{t=0} = v_0|_{\Gamma_0}$ in case $p>3/2$, $h_{s1}|_{t=0} = \mathcal{P}a_{ij}^{kl}(0,x)\nu_k D_l u_{0j}|_{\Gamma_s}$ in case $p>3$, $h_{s2}|_{t=0} = (v_0|\nu)|_{\Gamma_s}$, and $\mathcal{P}h_n|_{t=0} = \mathcal{P}a_{ij}^{kl}(0,x)\nu_k D_l v_{0j}|_{\Gamma_s}$ in case $p>3$. Note that ν is of class C^{2-} since Γ is assumed to be C^{3-}.

But in addition to these natural conditions on the data, there is a more involved structural property which is due to the divergence equation. To see this, take the L_2-inner product of a function $\phi\in H^1_{p'}(G)$ with the equation $\operatorname{div} v = g$, and integrate to the result

$$\int_G (\operatorname{div} v)\,\phi\,dx - \int_\Gamma (v|\nu)\phi\,d\sigma = -\int_G (v|\nabla\phi)dx.$$

Replacing $\operatorname{div} v$ by g and $(v|\nu)$ by h_ν, this yields

$$\langle F_{g,h_\nu}|\phi\rangle := \int_G g\phi\,dx - \int_\Gamma h_\nu\phi\,d\sigma = -\int_G (v|\nabla\phi)dx.$$

Differentiating w.r.t. t, we get

$$\frac{d}{dt}\langle F_{g,h_\nu}|\phi\rangle = -\int_G (\partial_t v|\nabla\phi)dx.$$

Define $\dot H^1_{p,\Gamma_n}(G)$ as

$$\dot H^1_{p,\Gamma_n}(G) = \{v\in H^1_{p,loc}(G):\ \nabla v\in L_p(G;\mathbb{C}^n), v|_{\Gamma_n}=0\},$$

and set

$$\dot H^{-1}_{p,\Gamma_n}(G) = (\dot H^1_{p',\Gamma_n}(G)/constants)^*.$$

Then we may conclude $F_{g,h_\nu}\in H^1_p(J;\dot H^{-1}_{p,\Gamma_n}(G))$. For this property, we write briefly $(g,h_\nu)\in H^1_p(J;\dot H^{-1}_{p,\Gamma_n}(G))$. Summarizing, we obtain the following necessary conditions on the data.

(D) *Assumptions on the data.*

(a) $f\in L_p(J\times G;\mathbb{C}^n)$.

(b1) $g\in L_p(J;H^1_p(G))$.

(b2) $(g,h_\nu)\in H^1_p(J;\dot H^{-1}_{p,\Gamma_n}(G))$, where $h_\nu = h_{j2}$ on Γ_j, $j=0,s$.

(c) $v_0\in W_p^{2-2/p}(G;\mathbb{C}^n)$ and $\operatorname{div} v_0 = g|_{t=0}$, $h_\nu|_{t=0} = (v_0|\nu)|_{\Gamma_j}$, $j=0,s$.

(d0) $h_0\in W_p^{1-1/2p}(J;L_p(\Gamma_0;\mathbb{C}^n))\cap L_p(J;W_p^{2-1/p}(\Gamma_0;\mathbb{C}^n))$;

$h_0|_{t=0} = v_0|_{\Gamma_0}$ in case $p > 3/2$.

(ds) $h_s \in W_p^{1/2-1/2p}(J; L_p(\Gamma_s; \mathbb{C}^n)) \cap L_p(J; W_p^{1-1/p}(\Gamma_s; \mathbb{C}^n))$,
$h_{s2} \in W_p^{1-1/2p}(J; L_p(\Gamma_0)) \cap L_p(J; W_p^{2-1/p}(\Gamma_0))$;
$h_{s1}|_{t=0} = \mathcal{P}a_{ij}^{kl}(0,x)\nu_k D_l v_{0j}|_{\Gamma_s}$ in case $p > 3$.

(dn) $h_n \in W_p^{1/2-1/2p}(J; L_p(\Gamma_n; \mathbb{C}^n)) \cap L_p(J; W_p^{1-1/p}(\Gamma_n; \mathbb{C}^n))$;
$\mathcal{P}h_n|_{t=0} = \mathcal{P}a_{ij}^{kl}(0,x)\nu_k D_l v_{0j}|_{\Gamma_s}$ in case $p > 3$.

Note that in case G is bounded and $\Gamma_n = \emptyset$, the natural compatibility condition

$$\int_G g\,dx - \int_\Gamma h_\nu\, d\sigma = 0$$

is included in (b2).

Based on the results for problem (4.1) on $G = \mathbb{R}^n$ and $G = \mathbb{R}^n_+$, by means of localization and straightening of the boundary of G, we can prove also the sufficiency of these conditions, which yields the following result.

THEOREM 4.1. *Let G and $\mathcal{A}(t,x,D)$ be as above, and let $1 < p < \infty$, $p \neq 3/2, 3$. Then* (4.1) *has maximal L_p-regularity in the following sense. There is a unique solution (v,q) of* (4.1) *in the class*

$$v \in H_p^1(J; L_p(G; \mathbb{C}^n)) \cap L_p(J; H_p^2(G; \mathbb{C}^n)), \quad q \in L_p(J; \dot{H}_p^1(G)),$$

such that $q \in W_p^{1/2-1/2p}(J; L_p(\Gamma_n)) \cap L_p(J; W_p^{1-1/p}(\Gamma_n))$ if and only if the data f, g, v_0, h_j satisfy conditions (D). *The solution (v,q) depends continuously on the data in the corresponding spaces. For $f \equiv g \equiv h \equiv 0$ and a_{ij}^{kl} independent of t the solutions u generate a semiflow in*

$$X_{pc} = \{v \in W_p^{2-2/p}(G; \mathbb{C}^n) : \operatorname{div} v = 0,\ v|_{\Gamma_0} = (v|\nu)|_{\Gamma_s} = \mathcal{P}a_{ij}^{kl}\nu_k D_l v|_{\Gamma_s \cup \Gamma_n} = 0\},$$

the natural phase space for (4.1) *in the L_p-setting.*

The proof of this result is given in section 7.

It is sometimes convenient to reduce the data to $v_0 = g = h_\nu = 0$. For this we have the following result.

PROPOSITION 4.2. *Let v_0, g, h_ν satisfying the relevant conditions in* (D) *be given. Then there is a function $v_* \in H_p^1(J; L_p(G; \mathbb{C}^n)) \cap L_p(J; H_p^2(G; \mathbb{C}^n))$ such that*

$$\operatorname{div} v_* = g, \quad v_*|_{t=0} = v_0, \quad (v_*|\nu)|_{\Gamma_0 \cup \Gamma_s} = h_\nu,$$

and v_* depends continuously on (v_0, g, h_ν) in the corresponding spaces.

Proof. Since $\nabla : \dot{H}_{p'}^1(G)/constants \to L_{p'}(G; \mathbb{C}^n)$ is injective and has closed range, its dual $\nabla^* := -\mathrm{div} : L_p(G; \mathbb{C}^n) \to \dot{H}_p^{-1}(G)$ is surjective. Choose a bounded right inverse R of ∇^*, which exists since the kernel of ∇^* is complemented in $L_p(G; \mathbb{C}^n)$. This assertion follows from the boundedness of the corresponding Helmholtz projection; cf. [22]. Then $Rg \in H_p^1(J; L_p(G; \mathbb{C}^n))$. Next, solve the parabolic problem

$$\begin{aligned}
(\partial_t - \Delta)w &= \partial_t Rg - \nabla g, \quad t \in J,\ x \in G,\\
w|_{t=0} &= v_0, \quad x \in G,\\
\operatorname{div} w &= g, \quad t \in J,\ x \in \Gamma,\\
\mathcal{P}w &= e^{\Delta_\Gamma t}(\mathcal{P}v_0|_\Gamma),
\end{aligned} \tag{4.2}$$

where Δ_Γ denotes the Laplace–Beltrami operator on Γ. Note that the right-hand side $f = \partial_t Rg - \nabla g$ belongs to $L_p(J \times G)$, $v_0 \in W_p^{2-2/p}(G; \mathbb{C}^n)$ by assumption; hence

$\mathcal{P}v_0|_\Gamma \in W_p^{2-3/p}(\Gamma;\mathbb{C}^n)$, and so the boundary function $h(t) = e^{\Delta_\Gamma t}(\mathcal{P}v_0|_\Gamma)$ belongs to $W_p^{1-1/2p}(J; L_p(\Gamma;\mathbb{C}^n)) \cap L_p(J; W_p^{2-1/p}(\Gamma;\mathbb{C}^n))$, by parabolic regularity theory, and the compatibility condition $\mathcal{P}h|_{t=0} = \mathcal{P}v_0|_\Gamma$ is valid. Similarly, taking traces, $g|_\Gamma$ belongs to $W_p^{1/2-1/2p}(J; L_p(\Gamma;\mathbb{C}^n)) \cap L_p(J; W_p^{1-1/p}(\Gamma;\mathbb{C}^n))$ by assumption and fulfills the compatibility condition $\operatorname{div} v_0 = g|_{t=0}$. Moreover, the Lopatinskii–Shapiro condition is easily seen to be valid (cf. the appendix), and hence problem (4.2) admits a unique solution w in the maximal class of L_p, by Theorem 10.1. Now set $\phi = \operatorname{div} w - g$. Then we see that ϕ solves the problem

$$\begin{aligned} (\partial_t - \Delta)\phi &= 0, \quad t \in J,\ x \in G, \\ \phi|_{t=0} &= 0, \quad x \in G, \\ \phi &= 0, \quad t \in J,\ x \in \Gamma; \end{aligned} \tag{4.3}$$

i.e., $\phi = 0$ by uniqueness of the parabolic problems. Thus the first two conditions in the proposition are valid for w. To obtain v_* we have to modify w in the following way. Solve the problem

$$\begin{aligned} \Delta\psi &= 0, \quad t \in J,\ x \in G, \\ \partial_\nu\psi &= h_\nu - (w|\nu), \quad t \in J,\ x \in \Gamma_0 \cup \Gamma_s, \\ \psi &= 0, \quad t \in J,\ x \in \Gamma_n, \end{aligned} \tag{4.4}$$

and set $v_* = w + \nabla\psi$. This function has the desired properties since for the boundary datum we have $h_\nu \in H_p^1(J; \dot{W}_p^{-1/p}(\Gamma_0 \cup \Gamma_s)) \cap L_p(J; W_p^{2-1/p}(\Gamma_0 \cup \Gamma_s))$, by assumption (b2). Here we once more used $\Gamma \in C^{3-}$. Note that the time trace of $h_\nu - (w|\nu)$ at $t = 0$ is zero by construction, and hence v_* has the right initial condition $v_*(0) = v_0$. ☐

5. The generalized Stokes problem on $\mathbb{R}^n$. Let $J = [0,T]$ be a compact interval, and consider the problem

$$\begin{aligned} \partial_t u + \mathcal{A}(D)u + \nabla\pi &= f(t,x), \quad t \in J,\ x \in \mathbb{R}^n, \\ \operatorname{div} u &= g(t,x), \quad t \in J,\ x \in \mathbb{R}^n, \\ u(0,x) &= u_0(x), \quad x \in \mathbb{R}^n. \end{aligned} \tag{5.1}$$

Here $\mathcal{A}(D) = \sum_{k,l=1}^n \mathcal{A}^{kl} D_k D_l$ denotes a differential operator with constant coefficient matrices $\mathcal{A}^{kl}$ acting on $\mathbb{C}^n$-valued functions. We assume that $\mathcal{A}(D)$ is *strongly elliptic*. This implies (cf. the appendix) that the problem

$$\begin{aligned} \partial_t u + \mathcal{A}(D)u &= f(t,x), \quad t \in J,\ x \in \mathbb{R}^n, \\ u(0,x) &= u_0(x), \quad x \in \mathbb{R}^n, \end{aligned} \tag{5.2}$$

has maximal L_p-regularity, $1 < p < \infty$. In particular, for each $f \in L_p(J \times \mathbb{R}^n; \mathbb{C}^n)$, $u_0 \in W_p^{2-2/p}(\mathbb{R}^n;\mathbb{C}^n)$ there is a unique solution u of (5.2) in the class

$$u \in H_p^1(J; L_p(\mathbb{R}^n;\mathbb{C}^n)) \cap L_p(J; H_p^2(\mathbb{R}^n;\mathbb{C}^n)) \hookrightarrow C(J; W_p^{2-2/p}(\mathbb{R}^n;\mathbb{C}^n)).$$

We want to show that a similar assertion is valid for the generalized Stokes problem (5.1). More precisely, with the definition of

$$\dot{H}_p^{-1}(\mathbb{R}^n) = (\dot{H}_{p'}^1(\mathbb{R}^n)/\mathit{constants})^* = \{f \in \mathcal{S}'(\mathbb{R}^n) : \mathcal{F}^{-1}|\xi|^{-1}\mathcal{F}f \in L_p(\mathbb{R}^n)\},$$

where $\mathcal{F}$ denotes Fourier transform, we have the following result.

THEOREM 5.1. *Let $1 < p < \infty$, and assume that $\mathcal{A}(D) = \sum_{k,l=1}^{n} \mathcal{A}^{kl} D_k D_l$ is strongly elliptic. Then* (5.1) *has maximal L_p-regularity in the following sense. There is a unique solution (u, π) of* (5.1) *in the class*

$$u \in H^1_p(J; L_p(\mathbb{R}^n; \mathbb{C}^n)) \cap L_p(J; H^2_p(\mathbb{R}^n; \mathbb{C}^n)), \quad \pi \in L_p(J; \dot{H}^1_p(\mathbb{R}^n)),$$

if and only if the data f, g, u_0 satisfy the following conditions:

(a) $f \in L_p(J \times \mathbb{R}^n; \mathbb{C}^n)$;

(b) $g \in H^1_p(J; \dot{H}^{-1}_p(\mathbb{R}^n)) \cap L_p(J; H^1_p(\mathbb{R}^n))$;

(c) $u_0 \in W^{2-2/p}_p(\mathbb{R}^n; \mathbb{C}^n)$ *and* $\mathrm{div}\, u_0 = g|_{t=0}$.

The solution (u, π) depends continuously on the data in the corresponding spaces. For $f \equiv g \equiv 0$ the solutions generate a semiflow in

$$X_p = \{v \in W^{2-2/p}_p(\mathbb{R}^n; \mathbb{C}^n) : \ \mathrm{div}\, u = 0\},$$

the natural phase space for (5.1) *in the L_p-setting.*

Proof. To prove sufficiency of the conditions, note that by the open mapping theorem, the continuity assertion follows as soon as the solvability assertion is proved. So let data f, g, u_0 be given, which are subject to conditions (a), (b), and (c). We first solve the parabolic problem

$$\partial_t v + \mathcal{A}(D)v = f, \quad v(0) = u_0,$$

with maximal L_p-regularity, applying the main result of Denk, Hieber, and Prüss [4, section 5]; cf. section 10, Theorem 10.1. Then $w = u - v$ must be a solution of the system

$$\partial_t w + \mathcal{A}(D)w + \nabla \pi = 0, \quad \mathrm{div}\, w = g_0, \quad w(0) = 0,$$

where $g_0 = g - \mathrm{div}\, v$ has the same regularity as g and trace 0 at time $t = 0$.

Suppose that the pressure π is already known. Taking Fourier transform in the space variables and Laplace transform in the time variable, we obtain the system

$$\begin{aligned} \lambda \hat{w} + \mathcal{A}(\xi)\hat{w} &= -i\xi\hat{\pi}, \\ i(\hat{w}|\xi) &= \hat{g}_0. \end{aligned} \tag{5.3}$$

Solving for $\hat{w}$, this yields

$$\hat{w} = -i(\lambda + \mathcal{A}(\xi))^{-1}\xi\hat{\pi},$$

and inserting this relation into the second equation of (5.3), we obtain

$$\hat{g}_0 = ((\lambda + \mathcal{A}(\xi))^{-1}\xi|\xi)\hat{\pi}.$$

Set $\eta = (\lambda + \mathcal{A}(\xi))^{-1}\xi$; then $\eta \neq 0$ unless $\xi = 0$, and

$$\alpha(\lambda, \xi) := ((\lambda + \mathcal{A}(\xi))^{-1}\xi|\xi) = \overline{\lambda}|\eta|^2 + (\eta|\mathcal{A}(\xi)\eta).$$

Hence strong ellipticity of $\mathcal{A}(D)$ implies $\alpha(\lambda, \xi) \neq 0$ for all $\xi \in \mathbb{R}^n$, $\lambda \in \Sigma_\theta$ with $|\xi| + |\lambda| \neq 0$, provided that $\theta > \pi/2$ is sufficiently close to $\pi/2$. We may now solve for $\hat{\pi}$ to the result

$$\hat{\pi} = \frac{\hat{g}_0}{\alpha(\lambda, \xi)},$$

and for $\hat{w}$ we get

$$\hat{w} = -i\frac{(\lambda + \mathcal{A}(\xi))^{-1}\xi}{\alpha(\lambda,\xi)}\hat{g}_0.$$

Choose

$$v_0 \in {}_0H^1_p(J; L_p(\mathbb{R}^n;\mathbb{C}^n)) \cap L_p(J; H^2_p(\mathbb{R}^n;\mathbb{C}^n)), \quad \text{div } v_0 = g_0;$$

this is possible by assumption (b) on the function g. In fact, setting

$$g_1 = (-\Delta)^{-1/2}\partial_t g_0 + (-\Delta)^{1/2} g_0,$$

we obtain $g_0 = -\text{div } R(\partial_t - \Delta)^{-1} g_1$, where R denotes the Riesz transform defined by the symbol $i\xi/|\xi|$; i.e., we may choose $v_0 = -R(\partial_t - \Delta)^{-1} g_1$. Therefore

$$(\partial_t - \Delta)w = T_1(\partial_t - \Delta)v_0, \quad \nabla\pi = T_2(\partial_t - \Delta)v_0,$$

where T_j are defined by means of their Fourier–Laplace symbols

$$\hat{T}_1(\lambda,\xi) = \frac{(\lambda + \mathcal{A}(\xi))^{-1}\xi\otimes\xi}{\alpha(\lambda,\xi)}, \quad \hat{T}_2(\lambda,\xi) = -\frac{\xi\otimes\xi}{(\lambda + |\xi|^2)\alpha(\lambda,\xi)}.$$

Thus, to prove the theorem, it is enough to show that the operators T_j are bounded in L_p.

This in turn will follow by an application of the Kalton–Weis theorem and $\mathcal{R}$-boundedness of families of Fourier multipliers. By the scaling $\mu = \lambda/|\xi|^2$, $\zeta = \xi/|\xi|$, we may rewrite the symbols as

$$\hat{T}_1(\lambda,\xi) = \frac{(\mu + \mathcal{A}(\zeta))^{-1}\zeta\otimes\zeta}{\alpha(\mu,\zeta)}, \quad \hat{T}_2(\lambda,\xi) = -\frac{\zeta\otimes\zeta}{(1+\mu)\alpha(\mu,\zeta)}.$$

By strong ellipticity, we already know $\alpha(\mu,\zeta) \neq 0$ for all $\zeta \in \mathbb{R}^n$, $|\zeta| = 1$, and $\mu \in \Sigma_\phi$, for some $\phi > \pi/2$. As $|\mu| \to \infty$ we have $\mu\alpha(\mu,\zeta) \to 1$, while $\alpha(\mu,\zeta) \to \alpha(0,\zeta) = (\mathcal{A}(\zeta)^{-1}\zeta|\zeta) \neq 0$ as $\mu \to 0$. Therefore, by compactness, $|(1+\mu)\alpha(\mu,\zeta)| \geq \alpha_0 > 0$ for all such ζ and μ, where α_0 denotes a constant. This implies boundedness of the symbols $\hat{T}_j(\mu|\xi|^2,\xi)$, uniformly in ξ and μ. Furthermore, $\hat{T}_j(\mu|\xi|^2,\xi)$ are homogeneous in ξ of degree 0, and so $|\xi|^{|\beta|}D^\beta_\xi \hat{T}_j(\mu|\xi|^2,\xi)$ are also uniformly bounded in ξ and μ, for each multi-index $\beta \in \mathbb{N}^n_0$. The classical Mikhlin multiplier theorem then implies that these symbols are Fourier multipliers in $L_p(\mathbb{R}^n;\mathbb{C}^n)$ w.r.t. ξ, which yields a holomorphic uniformly bounded family $\{T_j(\mu)\}_{\mu\in\Sigma_\phi} \subset \mathcal{B}(L_p(\mathbb{R}^n;E_j))$ for $j = 1,2$, where $E_1 = \mathbb{C}^n$, $E_2 = \mathbb{C}$. Theorem 3.2 of Girardi and Weis [10] shows that this family is also $\mathcal{R}$-bounded in $L_p(\mathbb{R}^n;E_j)$, and hence in $L_p(J;L_p(\mathbb{R}^n;E_j))$ as well. Since the operator $L := \partial_t(-\Delta)^{-1}$ admits an $\mathcal{H}^\infty$-calculus in $L_p(J;L_p(\mathbb{R}^n;E_j))$ of angle $\pi/2$, the Kalton–Weis theorem [11] implies boundedness of $T_j(L)$ in $L_p(J;L_p(\mathbb{R}^n;E_j))$. This completes the proof of Theorem 5.1. ∎

We can easily extend Theorem 5.1 to the case of variable coefficients with small deviation from constant ones. To see this, let $\mathcal{A}(t,x,D) = \mathcal{A}_0(D) + \mathcal{A}_1(t,x,D)$, where

$$\sup\{|a^{kl}_{1,ij}| : k,l,i,j = 1,\ldots,n,\ t\in J,\ x\in\mathbb{R}^n\} \leq \eta.$$

Let S denote the solution operator of the generalized Stokes problem (5.1) from Theorem 5.1 for $\mathcal{A}_0(D)$, and T that of the perturbed problem. Then we obtain the identity

$$T = S + SBT, \qquad \text{where} \qquad B = \begin{bmatrix} -\mathcal{A}_1(t,x,D) & 0 \\ 0 & 0 \end{bmatrix}.$$

The norm of B as an operator from the maximal regularity space into $L_p(J\times\mathbb{R}^n;\mathbb{C}^n)$ is bounded by $C\eta$, where $C>0$ denotes a constant independent of η. Let $|S|$ stand for the norm of the solution operator from the data space to the maximal regularity space. If $|S|C\eta<1$, then a Neumann series argument shows that $T=(I-SB)^{-1}S$ in fact exists and is bounded as a map from the data space to the maximal regularity space as well. Let us record this as the following result.

Corollary 5.2. *The assertions of Theorem* 5.1 *remain valid in the case of variable coefficients* $\mathcal{A}(t,x,D)=\mathcal{A}_0(D)+\mathcal{A}_1(t,x,D)$, *provided that*

$$\sup\{|a^{kl}_{1,ij}(t,x)| : k,l,i,j=1,\dots,n,\ t\in J,\ x\in\mathbb{R}^n\}\le\eta,$$

for some sufficiently small $\eta>0$.

In section 7 we will need a certain decomposition of the solution operator. For this purpose observe that from the proof of Theorem 5.1 we have the representations

$$\hat u=[I-(\lambda+\mathcal{A}(\xi))^{-1}\xi\otimes\xi](\lambda+\mathcal{A}(\xi))^{-1}\hat f-i\alpha^{-1}(\lambda+\mathcal{A}(\xi))^{-1}\xi\hat g$$

and

$$\hat\pi=-i\alpha^{-1}((\lambda+\mathcal{A}(\xi))^{-1}\hat f|\xi)+\frac{\hat g}{\alpha}.$$

Here we have assumed $u_0=0$ for simplicity. Let us have a closer look at the term $1/\alpha(\lambda,\xi)$. We may write

$$\begin{aligned}\frac{1}{\alpha(\lambda,\xi)}&=(\mu+1)\frac{1}{(\mu+1)((\mu+\mathcal{A}(\zeta))^{-1}\zeta|\zeta)}\\&=\mu+1+(\mu+1)\left[\frac{1}{(\mu+1)((\mu+\mathcal{A}(\zeta))^{-1}\zeta|\zeta)}-1\right]\\&=\mu+1+\frac{(\mu+1)[(\mu+\mathcal{A}(\zeta))-(\mu+1)]((\mu+\mathcal{A}(\zeta))^{-1}\zeta|\zeta)}{(\mu+1)((\mu+\mathcal{A}(\zeta))^{-1}\zeta|\zeta)}\\&=\mu+1+\frac{[\mathcal{A}(\zeta)-1](\mu+1)((\mu+\mathcal{A}(\zeta))^{-1}\zeta|\zeta)}{(\mu+1)((\mu+\mathcal{A}(\zeta))^{-1}\zeta|\zeta)}\\&=\frac{\lambda}{|\xi|^2}+1+M_{22}(\lambda,\xi),\end{aligned}$$

where we again used the notation $\mu=\lambda/|\xi|^2$ and $\zeta=\xi/|\xi|$. As in the proof of Theorem 5.1, $\xi\mapsto M_{22}(\mu|\xi|^2,\xi)$ is homogeneous of degree 0 and bounded, uniformly in $\xi\in\mathbb{R}^n$ and $\lambda\in\Sigma_\phi$. The arguments given there apply again to the result that there is an $L_p(J\times\mathbb{R}^n)$-bounded operator S_{22} with symbol $\hat S_{22}=M_{22}$. In a similar way we decompose

$$-i\alpha^{-1}(\lambda+\mathcal{A}(\xi))^{-1}\xi=\frac{-i\xi}{|\xi|^2}+(\lambda+|\xi|^2)^{-1}|\xi|M_{21}(\lambda,\xi),$$

where M_{21} is the symbol of an L_p-bounded operator S_{21}, as well as

$$-i((\lambda+\mathcal{A}(\xi))^{-1}\cdot|\xi)=-i(\xi/|\xi|^2|\cdot)+(\lambda+|\xi|^2)^{-1}|\xi|M_{12}(\lambda,\xi),$$

and M_{12} is the symbol of an L_p-bounded operator S_{12}. Last but not least, in the same way we obtain the decomposition

$$[I-(\lambda+\mathcal{A}(\xi))^{-1}\xi\otimes\xi](\lambda+\mathcal{A}(\xi))^{-1}=(\lambda+|\xi|^2)^{-1}+(\lambda+|\xi|^2)^2|\xi|^{-2}M_{11}(\lambda,\xi),$$

with M_{11} the symbol of an L_p-bounded operator S_{11}. Thus the solution operator S of the generalized Stokes problem splits as $S = S_0 + S_1$, where the symbols of S_j are given by

$$\hat{S}_0 = \begin{bmatrix} (\lambda + |\xi|^2)^{-1} & -i\xi/|\xi|^2 \\ -i(\xi/|\xi|^2|\cdot) & (\lambda + |\xi|^2)/|\xi|^2 \end{bmatrix} \tag{5.4}$$

and

$$\hat{S}_1 = \begin{bmatrix} (\lambda + |\xi|^2)^{-2}|\xi|^2 M_{11}(\lambda,\xi) & (\lambda + |\xi|^2)^{-1}|\xi| M_{12}(\lambda,\xi) \\ (\lambda + |\xi|^2)^{-1}|\xi| M_{21}(\lambda,\xi) & M_{22}(\lambda,\xi) \end{bmatrix}. \tag{5.5}$$

It is important that S_0 is independent of the coefficients of $\mathcal{A}(D)$, and that S_1 factors as

$$\hat{S}_1 = \begin{bmatrix} \frac{1}{\lambda+|\xi|^2} & 0 \\ 0 & \frac{1}{|\xi|} \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \frac{|\xi|^2}{\lambda+|\xi|^2} & 0 \\ 0 & |\xi| \end{bmatrix}.$$

Here $M = [M_{ij}]$ is the symbol of an L_p-bounded operator. It is a remarkable fact that such a decomposition remains valid in the variable coefficient case of Corollary 5.2. This can be seen as follows. We have the Neumann series for T which reads

$$T = S + \sum_{n\geq 1} (SB)^n S = S_0 + S_1 + \sum_{n\geq 1} (SB)^n S.$$

By induction we obtain

$$(SB)^n = \begin{bmatrix} (S_{11}\mathcal{A}_1)^n & 0 \\ S_{21}\mathcal{A}_1(S_{11}\mathcal{A}_1)^{n-1} & 0 \end{bmatrix}$$

and

$$(SB)^n S = \begin{bmatrix} (S_{11}\mathcal{A}_1)^n S_{11} & (S_{11}\mathcal{A}_1)^n S_{12} \\ S_{21}\mathcal{A}_1(S_{11}\mathcal{A}_1)^{n-1} S_{11} & S_{21}\mathcal{A}_1(S_{11}\mathcal{A}_1)^{n-1} S_{12} \end{bmatrix}.$$

In symbolic notation, using the factorization of S, this yields for the first entry

$$(S_{11}\mathcal{A}_1)^n S_{11} = \frac{1}{\lambda + |\xi|^2}\left(1 + \frac{|\xi|^2}{\lambda + |\xi|^2} M_{11}\right)(A_1 S_{11})^{n-1}\mathcal{A}\left(1 + \frac{|\xi|^2}{\lambda + |\xi|^2} M_{11}\right)\frac{|\xi|^2}{\lambda + |\xi|^2}.$$

Similarly, for the second entry we get

$$(S_{11}\mathcal{A}_1)^n S_{12} = \frac{1}{\lambda + |\xi|^2}\left(1 + \frac{|\xi|^2}{\lambda + |\xi|^2} M_{11}\right)(A_1 S_{11})^{n-1}\mathcal{A}_1(\zeta)\left(\frac{-i\xi}{|\xi|} + \frac{|\xi|^2}{\lambda + |\xi|^2} M_{12}\right)|\xi|.$$

In the same way the third entry becomes

$$S_{21}(\mathcal{A}_1 S_{11})^n = \frac{1}{|\xi|}\left(\frac{-i\xi}{|\xi|} + \frac{|\xi|^2}{\lambda + |\xi|^2} M_{21}\right)(A_1 S_{11})^{n-1}\mathcal{A}_1(\zeta)\left(1 + \frac{|\xi|^2}{\lambda + |\xi|^2} M_{11}\right)\frac{|\xi|^2}{\lambda + |\xi|^2},$$

and finally the last entry is

$$\begin{aligned} &S_{21}\mathcal{A}_1(S_{11}\mathcal{A}_1)^{n-1} S_{12} \\ &\quad = \frac{1}{|\xi|}\left(\frac{-i\xi}{|\xi|} + \frac{|\xi|^2}{\lambda + |\xi|^2} M_{12}\right)(A_1 S_{11})^{n-1}\mathcal{A}_1\left(\frac{-i\xi}{|\xi|} + \frac{|\xi|^2}{\lambda + |\xi|^2} M_{12}\right)|\xi|. \end{aligned}$$

This proves the assertion.

6. The generalized Stokes problem on $\mathbb{R}^n_+$. In this section we consider the generalized Stokes problem in $\mathbb{R}^n_+ = \mathbb{R}^{n-1} \times \mathbb{R}_+$ with any one of the three boundary conditions mentioned above. Thus we consider the problem

$$\begin{aligned} \partial_t u + \mathcal{A}(D)u + \nabla\pi &= f(t,x), \quad t \in J,\ x \in \mathbb{R}^n_+, \\ \operatorname{div} u &= g(t,x), \quad t \in J,\ x \in \mathbb{R}^n_+, \\ u(0,x) &= u_0(x), \quad x \in \mathbb{R}^n_+. \end{aligned} \tag{6.1}$$

Here, as in section 5, $\mathcal{A}(D) = \sum_{k,l=1}^n \mathcal{A}^{kl} D_k D_l$ denotes a strongly elliptic differential operator with constant coefficients acting on $\mathbb{C}^n$-valued functions. The boundary conditions are either

$$u(t,x) = h_0(t,x), \quad t \in J,\ x \in \partial\mathbb{R}^n_+, \tag{6.2}$$

or

$$u_n = h_{sn}(t,x),\ \left[\sum_{l=1}^n \mathcal{A}^{nl} D_l u\right]_k = h_{sk}(t,x) \quad t \in J,\ x \in \partial\mathbb{R}^n_+,\ k = 1,\dots,n-1, \tag{6.3}$$

or

$$\sum_{l=1}^n \mathcal{A}^{nl} D_l u + i\pi e_n = h_n(t,x), \quad t \in J,\ x \in \partial\mathbb{R}^n_+. \tag{6.4}$$

Of course, appropriate compatibility conditions have to be satisfied. Assuming strong normal ellipticity, we already verified that the parabolic problem without pressure and divergence condition satisfies the Lopatinskii–Shapiro condition for these boundary conditions, and hence is well-posed and has maximal L_p-regularity for $1 < p < \infty$. The main result of this section states that these properties carry over to the generalized Stokes problem. But before we state the theorem let us specify conditions (D) for the half-space case.

Using the notation introduced in section 4, we define ${}_0\dot{H}_p^{-1}(\mathbb{R}^n_+)$ as the dual of $\dot{H}^1_{p,\partial\mathbb{R}^n_+}(\mathbb{R}^n_+)$, and $\dot{H}_p^{-1}(\mathbb{R}^n_+)$ as the dual of $\dot{H}^1_{p,\emptyset}(\mathbb{R}^n_+)/constants$. Observe that the space ${}_0\dot{H}_p^{-1}(\mathbb{R}^n_+)$ consists solely of distributions in $\mathbb{R}^n_+$, but $\dot{H}_p^{-1}(\mathbb{R}^n_+)$ does not have this property. It should also be observed that $(0,h) \in \dot{H}_p^{-1}(\mathbb{R}^n_+)$ is equivalent to $h \in \dot{W}_p^{-1/p}(\mathbb{R}^{n-1})$.

The conditions for right-hand side f and for the initial value u_0 do not change; they become the following:

(a) $f \in L_p(J \times \mathbb{R}^n_+; \mathbb{C}^n)$, $u_0 \in W_p^{2-2/p}(\mathbb{R}^n_+; \mathbb{C}^n)$.

For g, part of the conditions are

(b) $g \in H^1_p(J; {}_0\dot{H}_p^{-1}(\mathbb{R}^n_+)) \cap L_p(J; H^1_p(\mathbb{R}^n_+))$, $\operatorname{div} u_0 = g|_{t=0}$.

The main part of the Dirichlet boundary condition is the same as well.

(d0) for Dirichlet boundary conditions:
$h_0 \in W_p^{1-1/2p}(J; L_p(\partial\mathbb{R}^n_+; \mathbb{C}^n)) \cap L_p(J; W_p^{2-1/p}(\partial\mathbb{R}^n_+; \mathbb{C}^n))$ and for $p > 3/2$ in addition $h(0,x) = u_0(x)$;

Similarly, we have

(ds) for slip boundary conditions:
$h_{sn} \in W_p^{1-1/2p}(J; L_p(\partial\mathbb{R}^n_+; \mathbb{C}^n)) \cap L_p(J; W_p^{2-1/p}(\partial\mathbb{R}^n_+; \mathbb{C}^n))$;
$h_{sk} \in W_p^{1/2-1/2p}(J; L_p(\partial\mathbb{R}^n_+)) \cap L_p(J; W_p^{1-1/p}(\partial\mathbb{R}^n_+))$ and, if $p > 3$,

$[\sum_{l=1}^n \mathcal{A}^{nl} D_l u_0]_k = h_{sk}(0,x)$ for $x \in \partial\mathbb{R}^n_+$, $k = 1, \dots, n-1$;

and

(dn) for Neumann boundary conditions:
$h_n \in W_p^{1/2-1/2p}(J; L_p(\partial\mathbb{R}^n_+; \mathbb{C}^n)) \cap L_p(J; W_p^{1-1/p}(\partial\mathbb{R}^n_+; \mathbb{C}^n))$ and, if $p > 3$,
$[h_n(0,x)]_k = [\sum_{l=1}^n \mathcal{A}^{nl} D_l u_0(x)]_k$ for $x \in \partial\mathbb{R}^n_+$, $k = 1, \dots, n-1$.

In case of Neumann conditions these are all requirements. In case of slip or Dirichlet conditions we have the additional property

(e) $(g, h_{jn}) \in H^1_p(J; \dot{H}_p^{-1}(\mathbb{R}^n_+))$ and $h_{jn}(0,x) = u_{0n}(x,0)$, $x \in \mathbb{R}^{n-1}$, $j = 0, s$.

After these preliminaries we can state the main result of this section.

THEOREM 6.1. *Let* $1 < p < \infty$, *and assume that* $\mathcal{A}(D) = \sum_{k,l=1}^n \mathcal{A}^{kl} D_k D_l$ *is strongly normally elliptic. Then* (6.1) *with boundary conditions* (6.2), (6.3), *or* (6.4) *has maximal* L_p*-regularity in the following sense. There is a unique solution* (u, π) *of* (6.1) *in the class*

$$u \in H^1_p(J; L_p(\mathbb{R}^n_+; \mathbb{C}^n)) \cap L_p(J; H^2_p(\mathbb{R}^n_+; \mathbb{C}^n)), \quad \pi \in L_p(J; \dot{H}^1_p(\mathbb{R}^n_+)),$$

satisfying the corresponding boundary condition with $\pi \in W_p^{1/2-1/2p}(J; L_p(\partial\mathbb{R}^n_+))$ *in case of boundary condition* (6.4), *if and only if the data* f, g, u_0, h *satisfy the conditions* (D). *The solution* u *depends continuously on the data in the corresponding spaces.*

Proof. According to the discussion above, we need to show only the sufficiency part. Let data f, g, u_0 and boundary data h with the corresponding regularity be given. Without loss of generality we may assume $f \equiv g \equiv u_0 \equiv 0$ and trace 0 of h at $t = 0$ in case it exists. This can be seen as follows. First we solve the parabolic initial-boundary value problem without pressure and divergence condition; this gives a function u_1 in the right regularity class. Then $u_2 := u - u_1$, and π should solve the full problem with $f \equiv h \equiv u_0 \equiv 0$ and g replaced by $g_1 := g - \operatorname{div} u_1$, which belongs to the same regularity class but has trace 0 at $t = 0$. Extend g_1 evenly in x_n to all of $J \times \mathbb{R}^n$, and solve the full-space generalized Stokes problem (5.1) with $f = u_0 = 0$ to obtain a pair (u_3, π_3) in the right regularity class. Then the pair (u_4, π_4) defined by $u_4 := u_2 - u_3$, $\pi_4 := \pi - \pi_3$ should solve (6.1) with the boundary condition in question, where $f \equiv g \equiv u_0 \equiv 0$ and $h = -\mathcal{B}(D)(u_3, \pi_3)$; here $\mathcal{B}(D)$ denotes the boundary operator under consideration. Note that the new boundary data h belongs to the right regularity class and has trace 0 at $t = 0$ whenever it exists.

So we have to solve the homogeneous problem (6.1) with one of the inhomogeneous boundary conditions. It is convenient to split the spatial variables as $x = (x', y)$, where $x' \in \mathbb{R}^{n-1}$ and $y > 0$; recall that $\nu = -e_n$. Similarly we decompose $u = (v, w)$, with $v \in \mathbb{R}^{n-1}$ the tangential and $w \in \mathbb{R}$ the normal velocity. Taking Fourier transform in the tangential space directions, Laplace transform in t, we obtain the parameter-dependent ode-problem

$$\begin{aligned}
(\lambda + \mathcal{A}_{11}(\xi + e_n D_y))\hat{v} + \mathcal{A}_{12}(\xi + e_n D_y)\hat{w} + i\xi\hat{\pi} = 0, \quad y > 0,\\
\mathcal{A}_{21}(\xi + e_n D_y)\hat{v} + (\lambda + \mathcal{A}_{22}(\xi + e_n D_y))\hat{w} + \partial_y\hat{\pi} = 0, \quad y > 0,\\
i\xi^{\mathsf{T}}\hat{v} + iD_y\hat{w} = 0, \quad y > 0,\\
\mathcal{B}_{11}(\xi + e_n D_y)v(0) + \mathcal{B}_{12}(\xi + e_n D_y)w(0) = \hat{h}_v,\\
\mathcal{B}_{21}(\xi + e_n D_y)v(0) + \mathcal{B}_{22}(\xi + e_n D_y)w(0) + \mathcal{B}_{23}\pi(0) = \hat{h}_w,
\end{aligned} \tag{6.5}$$

where $\mathcal{B}$ is defined by one of the boundary conditions (6.2), (6.3), or (6.4). The parameters ξ and λ satisfy $(\xi, \lambda) \in \mathbb{R}^n \times \Sigma_\phi$ for some $\phi > \pi/2$ and $\xi_n = 0$. Here and below we identify $\xi \in \mathbb{R}^{n-1}$ with $(\xi, 0) \in \mathbb{R}^n$. Introducing the vector

$$x = [\hat{v}, \hat{w}, \partial_y \hat{v}, \partial_y \hat{w}, \hat{\pi}]^{\mathsf{T}},$$

we rewrite this problem as the first order system

$$E\partial_y x + Ax = 0, \quad y > 0, \quad Bx(0) = \hat{h}, \tag{6.6}$$

where the dependence on (λ, ξ) has been dropped. Here the $(2n+1)$-dimensional square matrix E is defined as

$$E = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & A^0_{11} & A^0_{12} & 0 \\ 0 & 0 & A^0_{21} & A^0_{22} & -1 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

and A by

$$A = \begin{bmatrix} 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ -(\lambda + A^2_{11}) & -A^2_{12} & A^1_{11} & A^1_{12} & -i\xi \\ -A^2_{21} & -(\lambda + A^2_{22}) & A^1_{21} & A^1_{22} & 0 \\ i\xi^{\mathsf{T}} & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Here we have used the abbreviations

$$A^2 = (\mathcal{A}^{kl}\xi_k\xi_l), \quad A^1 = i(\mathcal{A}^{kl}\nu_k\xi_l + \mathcal{A}^{kl}\nu_l\xi_k), \quad A^0 = (\mathcal{A}^{kl}\nu_k\nu_l);$$

remember the summation convention. Observe that A^k are homogeneous in ξ of order k; in particular, A^0 is constant and invertible by ellipticity. Also note that E depends neither on λ nor on ξ. The boundary matrices B are

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

in the case of Dirichlet conditions,

$$B = \begin{bmatrix} B^1_{11} & B^1_{12} & B^0_{11} & B^0_{12} & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

for slip conditions, and

$$B = \begin{bmatrix} B^1_{11} & B^1_{12} & B^0_{11} & B^0_{12} & 0 \\ B^1_{21} & B^1_{22} & B^0_{21} & B^0_{22} & -1 \end{bmatrix}$$

in the case of Neumann conditions. Here B^k are homogeneous of order k in ξ, and $B^0 = A^0$. Recall that the Lopatinskii–Shapiro condition means that system (6.6) admits at most one solution $x \in C_0(\mathbb{R}_+; \mathbb{C}^{2n+1})$, for each $\hat{h} \in \mathbb{C}^n$ and $\xi \in \mathbb{R}^{n-1}$, $\operatorname{Re}\lambda \geq 0$, $\xi \neq 0$.

It is our purpose to derive a representation formula of the function x in terms of the given data $\hat{h}$, which is accessible to inversion of the Fourier and Laplace transform.

(i) Assume $x \in C_0(\mathbb{R}_+; \mathbb{C}^{2n+1})$ to be a solution of (6.6). Taking Laplace transform $\mathcal{L}$ in y, this yields

$$(zE + A)\mathcal{L}x(z) = Ex^0, \quad \operatorname{Re} z > 0, \quad Bx^0 = \hat{h},$$

where $x^0 = x(0)$ denotes the initial value of x. To obtain a representation of x we have to study the operator pencil $zE + A$. To this end note that E is not invertible but its kernel $N(E)$ is one-dimensional, and $N(E^2) = N(E)$; hence $N(E) \oplus R(E) = \mathbb{C}^{2n+1}$. Therefore, (6.6) is a differential-algebraic system of index ≥ 1. This implies that the characteristic polynomial $p(z) = \det(zE + A)$ has at most order $2n$. Let us show that it is precisely of order $2n$, i.e., that the index is 1. This can be seen as follows. Expand $\det(zE + A)$ first w.r.t. the last column and the last row and then w.r.t. the second row. This yields up to a sign

$$p(z) = z^2 \det \begin{bmatrix} z & -1 \\ -(\lambda + A_{11}^2) & zA_{11}^0 + A_{11}^1 \end{bmatrix} + q(z),$$

where $q(z)$ is of order less than $2n$. Asymptotically this yields for large z

$$p(z) \sim z^2 \det \begin{bmatrix} z & 0 \\ 0 & zA_{11}^0 \end{bmatrix} = z^{2n} \det A_{11}^0,$$

and $\det A_{11}^0 \neq 0$ by strong ellipticity. Therefore $p(z)$ is of order $2n$. Ellipticity shows also that $p(z)$ has no zeros on the imaginary axis for $\xi \neq 0$. Now consider the case $\xi = 0$. Then we see by the same procedure that $p(z)$ is in fact a function of z^2; i.e., if z_0 is a zero of p, then $-z_0$ is one as well. Unfortunately, $z = 0$ is a solution in case $\xi = 0$; this is the degeneracy of the Stokes problem. We have to look at this zero more closely.

The eigenvalue problem for these small zeros $z(\xi)$ for small ξ (or large λ) becomes

$$(A(z,\xi) - \lambda) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} i\xi \\ z \end{bmatrix}, \quad (i\xi | x_1) + z x_2 = 0,$$

where

$$A(z,\xi) = z^2 A^0 + zA^1(\xi) - A^2(\xi).$$

Since by $\lambda \neq 0$ we have invertibility of $A(z,\xi) - \lambda$, this implies the condition

$$\left(\begin{bmatrix} i\xi \\ z \end{bmatrix} \middle| (A(z,\xi) - \lambda)^{-1} \begin{bmatrix} i\xi \\ z \end{bmatrix} \right) = 0$$

for the small eigenvalues. Writing $(A(z,\xi) - \lambda)^{-1}$ as a Neumann series, this condition becomes

$$z^2 - |\xi|^2 + O((|\xi| + |z|)^4) = 0,$$

which shows that $z = \pm|\xi| + O(|\xi|^2)$ near $\xi = 0$. Therefore the double zero $z(0) = 0$ for $\xi = 0$ splits into two simple real zeros, which behave like $z_1^{\pm}(\xi) \sim \pm|\xi|$ near $\xi = 0$.

Now varying ξ, we may conclude that $p(z)$ has exactly n roots with positive real parts, counting with multiplicity, for each $\xi \in \mathbb{R}^{n-1}$, $\operatorname{Re} \lambda > 0$, $\xi \neq 0$, since none of them can cross the imaginary axis by ellipticity.

We may now write

$$\mathcal{L}x(z) = (zE + A)^{-1}Ex^0, \quad Bx^0 = \hat{h},$$

for the Laplace transform of x. The initial value x^0 thus must be chosen in such a way that $\mathcal{L}x(z)$ has no poles in the right half-plane, and $Bx^0 = \hat{h}$ holds.

Define the projection P^+ by means of

$$P^+ = \frac{1}{2\pi i}\int_{\Gamma_+} (zE + A)^{-1}Edz,$$

where Γ_+ denotes a closed simple contour in the right half-plane surrounding the poles of $(zE + A)^{-1}$, i.e., the zeros of $p(z)$ in the right half-plane. Let z_k, $k = 1, \dots, m^+$ denote the zeros of $p(z)$ in the right and for $k = -m^-, \dots, -1$ in the left half-plane. Set

$$P_k = \frac{1}{2\pi i}\int_{|z-z_k|=r} (zE + A)^{-1}Edz.$$

These operators are mutually disjoint projections, and by Cauchy's theorem we have

$$P^+ = \sum_{k=1}^{m^+} P_k.$$

It can be seen, e.g., by Cramer's rule, that $(zE + A)^{-1}$ is a rational function which is bounded at ∞ and hence admits a limit as $|z| \to \infty$. Therefore

$$z(zE + A)^{-1}E = I - (zE + A)^{-1}A$$

is bounded at ∞ as well and admits the limit

$$Q_0 = \lim_{z\to\infty} z(zE + A)^{-1}E,$$

which is a projection, too. We set $P_0 = I - Q_0$. Obviously, $Q_0x = 0$ for each $x \in N(E)$, and, on the other hand, we have

$$EQ_0 = \lim_{z\to\infty} zE(zE + A)^{-1}E = \lim_{z\to\infty}(E - A(zE + A)^{-1}E) = E.$$

This implies that P_0 projects onto the kernel of E. Moreover,

$$\sum_k P_k = P_0 + \lim_{R\to\infty}\frac{1}{2\pi i}\int_{|z|=R}(zE + A)^{-1}Edz = P_0 + Q_0 = I,$$

which also shows that $P_0P_k = P_kP_0 = 0$ for all $k \neq 0$. Linear algebra implies further that the dimension of the range of P_k is m_k, and hence P^+ has dimension n. Since

$$x^0 = x(0) = \lim_{t\to 0+} x(t) = \lim_{\mathbb{R}\ni z\to\infty} z\mathcal{L}x(z) = \lim_{z\to\infty} z(zE + A)^{-1}Ex^0 = Q_0x^0,$$

we must have $P_0x^0 = 0$. It is not difficult to compute the projection P_0; it is given by

$$P_0x = \frac{x_4 + (i\xi|x_1)}{\alpha_0}\begin{bmatrix} 0 \\ A^{0^{-1}}\left[\begin{smallmatrix} 0 \\ -1 \end{smallmatrix}\right] \\ 1 \end{bmatrix},$$

where

$$\alpha_0 := \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix} | A^{0^{-1}} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right)$$

is nonzero by ellipticity. Observe that

$$P_0 x^0 = 0 \quad \Leftrightarrow \quad x_4^0 + (i\xi | x_1^0) = 0.$$

For later purposes we also compute the projection $P_1^{\pm}$ corresponding to the small eigenvalue $z_1^{\pm}(\xi) \sim \pm|\xi|$ for small ξ. The analysis of $z_1^{\pm}$ given above shows that an eigenvector is given by

$$e_1^{\pm} = \left[(A(z_1^{\pm}) - \lambda)^{-1} \begin{bmatrix} i\xi \\ z_1^{\pm} \end{bmatrix}, z_1^{\pm}(A(z_1^{\pm}) - \lambda)^{-1} \begin{bmatrix} i\xi \\ z_1^{\pm} \end{bmatrix}, 1 \right]^{\mathsf{T}} \sim \left[\frac{1}{\lambda} \begin{bmatrix} -i\xi \\ \mp|\xi| \end{bmatrix}, 0, 1 \right]^{\mathsf{T}}.$$

For a dual eigenvector we get similarly

$$e_1^{*\pm} = \left[(z_1^{\pm} A^0 + A^1)^{\mathsf{T}} (A(z_1^{\pm})^{\mathsf{T}} - \lambda)^{-1} \begin{bmatrix} i\xi \\ z_1^{\pm} \end{bmatrix}, (A(z_1^{\pm})^{\mathsf{T}} - \lambda)^{-1} \begin{bmatrix} i\xi \\ z_1^{\pm} \end{bmatrix}, -1 \right]^{\mathsf{T}},$$

and hence

$$e_1^{*\pm} \sim \left[0, \frac{1}{\lambda} \begin{bmatrix} -i\xi \\ \mp|\xi| \end{bmatrix}, -1 \right]^{\mathsf{T}}.$$

The projections are then $P_1^{\pm} x = \frac{(e_1^{*\pm}|Ex)}{(e_1^{*\pm}|Ee_1^{\pm})} e_1^{\pm}$. Note that $(e_1^{*\pm}|Ee_1^{\pm}) \sim \pm 2|\xi|/\lambda$ for small ξ, and the asymptotics of $z_1^{\pm}$, $e_1^{\pm}$, and $e_1^{*\pm}$ do not depend on the coefficients a_{ij}^{kl}. Note also that

$$P_1^{+} x^0 = 0 \quad \Leftrightarrow \quad (e_1^{*+}|Ex^0) = 0,$$

which asymptotically yields the condition

$$x_5^0 - \frac{\lambda}{|\xi|} x_2^0 \sim \left(\begin{bmatrix} i\xi/|\xi| \\ 1 \end{bmatrix} | A^0 \begin{bmatrix} x_3^0 \\ x_4^0 \end{bmatrix} \right).$$

(ii) To determine the initial value x^0 we therefore have to solve the linear system

$$Bx^0 = \hat{h}, \quad P^+ x^0 = 0, \quad P_0 x^0 = 0. \tag{6.7}$$

The Lopatinskii–Shapiro condition is equivalent to the uniqueness of the solution x^0 of this system, for $\xi \neq 0$. To see that it is solvable for each $\hat{h} \in \mathbb{C}^n$, observe that the kernel N of $P^+ + P_0$ has dimension n. $B : N \to \mathbb{C}^n$ is injective, and hence the rank theorem implies that it is also surjective. Thus there is a linear operator $M = M(\lambda, \xi)$ such that $x^0 = M(\lambda, \xi)\hat{h}$ gives the unique solution of (6.7). We have the explicit representation

$$x^0 = (B^* B + (P^+)^* P^+ + P_0^* P_0)^{-1} B^* \hat{h},$$

which shows that $M(\lambda, \xi)$ is holomorphic since B and P^+ have this property. By homogeneity, λ can even be taken from a sector Σ_ϕ for some $\phi > \pi/2$, but $\xi \neq 0$ in general.

Therefore, we have to look more closely at $\xi = 0$. Note that the projections $P_1^\pm$ are not holomorphic at $\xi = 0$. However, $P_1^0 := P_1^+ + P_1^-$ does have this property. A simple calculation shows that for $\xi = 0$ we have

$$P_1^0 x = x_2 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + (x_5 - A_{21}^0 x_3 - A_{22}^0 x_4) \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Therefore it is convenient to decompose $x^0 = y^0 + \alpha e_1^-$, with $\alpha \in \mathbb{C}$ and $P_1^- y^0 = 0$. Setting $P = P_0 + P^+ + P_1^-$, we therefore have to solve the system

$$By^0 + \alpha B e_1^- = \hat{h}, \quad Py^0 = 0.$$

From $Py^0 = 0$ we obtain $y_2^0 = 0$, $y_4^0 = 0$, and $y_5^0 = A_{21}^0 y_3^0$. Solving the system $(zE + A)x = Ex^0$, we obtain with $e_1^- = [0,0,0,0,1]^\mathsf{T}$ and $x_2^0 = y_2^0 = x_4^0 = y_4^0 = 0$ the relations $x_2 = x_4 = 0$ and

$$(z^2 A_{11}^0 - \lambda) x_1 = A_{11}^0 (x_3^0 + z x_1^0), \quad x_3 = z x_1 - x_1^0, \quad x_5 = A_{21}^0 x_3 + \frac{\alpha}{z},$$

since $x_5^0 - A_{21}^0 x_3^0 = \alpha + y_5^0 - A_{21}^0 y_3^0 = \alpha$. By strong ellipticity, A_{11}^0 is invertible and has spectrum in the open right half-plane, and hence we may compute further

$$\begin{aligned} x_1(z) = {}& \frac{1}{2}(z + \sqrt{\lambda}(A_{11}^0)^{-1/2})^{-1} \left(y_1^0 + (A_{11}^0)^{1/2} \frac{y_3^0}{\sqrt{\lambda}} \right) \\ & + \frac{1}{2}(z - \sqrt{\lambda}(A_{11}^0)^{-1/2})^{-1} \left(y_1^0 - (A_{11}^0)^{1/2} \frac{y_3^0}{\sqrt{\lambda}} \right). \end{aligned}$$

Now, $x_1(z)$ must be holomorphic in the right half-plane, which means that necessarily we have $y_3^0 = -\sqrt{\lambda}(A_{11}^0)^{-1/2} y_1^0$. The boundary condition yields in the Dirichlet case $x_1^0 = y_1^0 = \hat{h}_1$, and in the slip or Neumann case $x_3^0 = y_3^0 = (A_{11}^0)^{-1}\hat{h}_3$. Note that in the Neumann case, $\alpha = -\hat{h}_4$ is uniquely determined, in contrast to the Dirichlet or slip case, where α is not unique. In fact, the function $\alpha(\lambda, \xi)$ is discontinuous at $\xi = 0$ for the latter, but holomorphic in the Neumann case.

Now, for $\xi \neq 0$ small, we may parametrize the kernel of P by a holomorphic map

$$y \mapsto R(\lambda,\xi)y := [y, 0, -\sqrt{\lambda}(A_{11}^0)^{-1/2} y, 0, -A_{21}^0 \sqrt{\lambda}(A_{11}^0)^{-1/2} y]^\mathsf{T} + R^1(\lambda,\xi)y,$$

where $R^1 = O(|\xi|)$ near $\xi = 0$, with $y \in \mathbb{C}^{n-1}$. Then we have to solve the equation $BRy + \alpha B e_1^- = \hat{h}$. For the pure Neumann case it then follows that y and α are uniquely determined and holomorphic near $\xi = 0$; hence $M(\lambda, \xi)$ is holomorphic also at $\xi = 0$. However, in the other cases things are more involved. We begin with the Dirichlet case. Then the system becomes

$$y - \frac{i\alpha\xi}{\lambda} = \hat{h}_1 + O(|\xi|)y + O(|\xi|^2)\alpha, \quad \frac{\alpha|\xi|}{\lambda} = \hat{h}_2 + O(|\xi|)y + O(|\xi|^2)\alpha,$$

and hence

$$\alpha \sim \frac{\lambda}{|\xi|}\hat{h}_2, \quad y \sim \hat{h}_1 + \frac{i\xi}{|\xi|}\hat{h}_2.$$

In the case of slip conditions we have similarly

$$-\sqrt{\lambda}{A_{11}^0}^{1/2}y - \alpha A_{11}^0\frac{i\xi}{\sqrt{\lambda}} = \hat{h}_3 + O(|\xi|)y + O(|\xi|^2)\alpha, \quad \frac{\alpha|\xi|}{\lambda} = \hat{h}_2 + O(|\xi|)y + O(|\xi|^2)\alpha,$$

and so

$$\alpha \sim \frac{\lambda}{|\xi|}\hat{h}_2, \quad y \sim -{A_{11}^0}^{1/2}\frac{{A_{11}^0}^{-1}\hat{h}_3 + \frac{i\xi}{|\xi|}\hat{h}_2}{\sqrt{\lambda}}.$$

Thus there are holomorphic functions $M_0(\lambda,\xi)$ and $\alpha_0(\lambda,\xi)$ such that

$$M(\lambda,\xi)\hat{h} = M_0(\lambda,\xi)\hat{h} + \left[\frac{\lambda}{|\xi|}\hat{h}_2 + (\alpha_0(\lambda,\xi)|\hat{h})\right]e_1^-,$$

where $\hat{h}_2$ denotes the normal component of u at the boundary $\partial\mathbb{R}^n_+$.

(iii) We may now write the following representation of the solution $x(y) = x(y,\lambda,\xi)$ of (6.6):

$$x(y,\lambda,\xi) = \frac{1}{2\pi i}\int_{\Gamma_-} e^{zy}(zE + A(\lambda,\xi))^{-1}EM(\lambda,\xi)\hat{h}(\lambda,\xi)dz, \tag{6.8}$$

where Γ_- denotes a closed simple contour in the open left half-plane surrounding the zeros of $p(z) = p(z,\lambda,\xi)$ in the left half-plane. Employing residue calculus, this representation can be rewritten as

$$x(y,\lambda,\xi) = \sum_{\mathrm{Re} z_k<0} \mathrm{Res}_{z=z_k(\lambda,\xi)}[e^{zy}(zE + A(\lambda,\xi))^{-1}E]M(\lambda,\xi)\hat{h}(\lambda,\xi);$$

hence it is an exponential polynomial in y.

Note that the zeros z_k of $p(z) = p(z,\lambda,\xi)$ depend on ξ and λ, and hence the integration path in (6.8) cannot be chosen independently of ξ and λ. To remove this defect a scaling argument will help. With $\rho = \sqrt{\lambda + |\xi|^2}$, the standard parabolic symbol, and $\sigma = \lambda/\rho^2$, $\zeta = \xi/\rho$, the pair (σ,ζ) belongs to a compact subset of $\mathbb{C}^n \setminus \{0\}$. Replace $\hat{\pi}(y)$ by $\hat{\pi}(\rho y)/\rho$, $x(y)$ by $x(\rho y)$, Neumann data $\hat{h}_k$ by $\hat{h}_k/\rho$, and leave Dirichlet data unchanged. Then homogeneity of $\mathcal{A}$ and $\mathcal{B}$ yield the modified representation formula

$$x(y,\lambda,\xi) = \frac{1}{2\pi i}\int_{\Gamma_-} e^{\rho zy}(zE + A(\sigma,\zeta))^{-1}EM(\sigma,\zeta)\hat{h}(\lambda,\xi)dz. \tag{6.9}$$

Since the poles of $(zE + A(\sigma,\zeta))^{-1}$ stay in a compact set in the left half-plane we may now choose the contour Γ_- independently of (σ,ζ). This argument parallels the scaling employed in section 6 of Denk, Hieber, and Prüss [4] for the parabolic case.

Observe that the scaling of h induces

$$h \in Y := {}_0W_p^{1-1/2p}(J; L_p(\mathbb{R}^{n-1};\mathbb{C}^n)) \cap L_p(J; W_p^{2-1/p}(\mathbb{R}^{n-1};\mathbb{C}^n)),$$

which is independent of the choice of the boundary conditions. Let $L := (\partial_t - \Delta_x)^{1/2}$ with natural domain

$$D(L) = {}_0H_p^{1/2}(J; L_p(\mathbb{R}^{n-1};\mathbb{C}^n)) \cap L_p(J; H_p^1(\mathbb{R}^{n-1};\mathbb{C}^n)).$$

Then $Y = D_L(2-1/p,p)$; hence if $h \in Y$, then $e^{-L\cdot}h \in D(L^2)$; i.e.,

$$g(\cdot) = L^2 e^{-L\cdot} h \in L_p(J \times \mathbb{R}^n_+; \mathbb{C}^n).$$

The symbol of L is $\sqrt{\lambda + |\xi|^2}$, which is precisely ρ. By means of the identity

$$\hat{h} = \int_0^\infty 2\rho e^{-2\rho\bar{y}} \hat{h} d\bar{y} = \frac{2}{\rho} \int_0^\infty e^{-\rho\bar{y}} \hat{g}(\bar{y}) d\bar{y},$$

we may rewrite the representation of $x(y)$ in the following way:

$$x(y,\lambda,\xi) = \operatorname{diag}\left[\frac{1}{\rho^2}, \frac{1}{\rho^2}, \frac{1}{\rho^2}, \frac{1}{\rho^2}, \frac{1}{\rho|\xi|}\right] \int_0^\infty \hat{k}(y,\bar{y},\lambda,\xi)\hat{g}(\bar{y},\lambda,\xi)d\bar{y}, \tag{6.10}$$

where the Fourier–Laplace transform of k is given by

$$\hat{k}(y,\bar{y},\lambda,\xi) = \frac{1}{i\pi}\int_{\Gamma_-} e^{\rho(yz-\bar{y})} D(\rho,|\xi|)(zE + A(\sigma,\zeta))^{-1} EM(\sigma,\zeta)dz, \tag{6.11}$$

where $D(\rho,|\xi|) = \operatorname{diag}[\rho,\rho,\rho,\rho,|\xi|]$.

It remains to be shown that the integral operator with operator-valued kernel $K(y,\bar{y})$ is bounded from $L_p(J \times \mathbb{R}^n_+; \mathbb{C}^n)$ to $L_p(J \times \mathbb{R}^n_+; \mathbb{C}^{2n+1})$, where the symbol of $K(y,\bar{y})$ is $\hat{k}(y,\bar{y},\lambda,\xi)$ from (6.11). This then implies that u belongs to the maximal regularity space, and the remaining regularity statements concerning the pressure π follow from the equations.

(iv) However, due to the presence of the small eigenvalues $z_1^\pm(\xi)$ introduced above, there are problems at $\zeta = 0$. We have to deal with the cases $|\zeta| \le \eta$ and $|\zeta| > \eta$ for some small $\eta > 0$ separately. For this purpose we introduce a cut-off function $\chi(|\zeta|^2)$, where χ belongs to C^∞ and is 1 in $B_\eta(0)$, 0 outside of $B_{2\eta}(0)$, and between 0 and 1 elsewhere. Then we may decompose $\hat{k}(y,\bar{y},\lambda,\xi)$ as $\hat{k} = \hat{k}_S + \hat{k}_R$, where

(6.12)
$$\hat{k}_R(y,\bar{y},\lambda,\xi) = \frac{1}{2\pi i}\int_{\Gamma_-} (1-\chi(\zeta))D(\rho,|\xi|)e^{\rho(zy-\bar{y})}(zE + A(\sigma,\zeta))^{-1} EM(\sigma,\zeta)dz.$$

Let us first deal with $\hat{k}_R$ and invert the Fourier transform via Mikhlin's theorem. Since Γ_- is compact and contained in the open left half-plane, for $|\zeta| > \eta$, (σ,ζ) runs through a compact subset of $\mathbb{C}^n$, and

$$\operatorname{Re}\rho \le |\rho| \le c_\phi \operatorname{Re}\rho,$$

we obtain

$$|\hat{k}_R(y,\bar{y},\lambda,\xi)| \le C|\rho|e^{-c|\rho|(y+\bar{y})} \le \frac{C}{y+\bar{y}}, \quad y,\bar{y} > 0,$$

where $C, c > 0$ are independent of $y, \bar{y}$, λ, and ξ. This is already sufficient in case $p = 2$, by Plancherel's theorem. For the case of general $p \in (1,\infty)$, note first that

$$|\xi|\left|\frac{1}{\rho}\partial_{\xi_k}\rho\right| = |\xi|\left|\frac{\xi_k}{\rho^2}\right| \le \frac{|\xi|^2}{\rho^2} \le 1,$$

and similarly we have by induction $|\xi|^{|\alpha|}|D^\alpha_\xi \rho| \le M_\alpha$, for each multi-index $\alpha \in \mathbb{N}_0^{n-1}$. Next,

$$|\xi||\partial_{\xi_k}\zeta_j| = |\xi|\left|\frac{\delta_{kj}}{\rho} - \frac{\zeta_j \partial_{\xi_k}\rho}{\rho^2}\right| \le M_1,$$

and similarly for higher derivatives, by induction. The relation $\sigma = 1 - |\xi|^2/\rho^2$ shows that also $|\xi|^{|\alpha|}|D^\alpha_\xi \sigma|$ is uniformly bounded for each α. Next

$$|\xi||\partial_{\xi_k} e^{\rho(yz-\bar{y})}| \le |\xi|\left|\frac{\partial_{\xi_k}\rho}{\rho^2}\right| |\rho^2(yz-\bar{y})e^{\rho(yz-\bar{y})}| \le C|\rho|e^{-c|\rho|(y+\bar{y})} \le \frac{C}{y+\bar{y}},$$

and similarly by induction also for all higher derivatives. Therefore we may conclude that

$$|\xi|^{|\alpha|}|D^\alpha_\xi \hat{k}_R(y,\bar{y},\lambda,\xi)| \le \frac{M_\alpha}{y+\bar{y}}, \quad y,\bar{y} > 0,$$

for each multi-index α, where M_α is independent of $y, \bar{y}$, λ, and ξ. Mikhlin's theorem implies that there is a family of operators $k_R(y,\bar{y},\lambda)$ from $L_p(\mathbb{R}^{n-1};\mathbb{C}^n)$ to $L_p(\mathbb{R}^{n-1};\mathbb{C}^{2n+1})$ with norms bounded by

$$||k_R(y,\bar{y},\lambda)|| \le \frac{c}{y+\bar{y}}, \quad y,\bar{y} > 0,\ \lambda \in \Sigma_\phi.$$

Because of uniformity of the "Mikhlin bounds," Theorem 3.2 of Girardi and Weis [10] implies that the family $\{(y+\bar{y})k(y,\bar{y},\lambda) :\ y,\bar{y} > 0,\ \lambda \in \Sigma_\phi\}$ is also $\mathcal{R}$-bounded in $\mathcal{B}(L_p(J \times \mathbb{R}^n;\mathbb{C}^{(2n+1)\times n}))$.

(v) Now we deal with the other part of $\hat{k}$. Since we have enough information about the small eigenvalue $z_1(\xi)$ we may use residue calculus to decompose $\hat{k}_S = \hat{k}_{S0} + \hat{k}_{S1}$, where

$$\hat{k}_{S1}(y,\bar{y},\lambda,\xi) = \frac{1}{i\pi}\int_{\Gamma_-} \chi(\zeta)e^{\rho(yz-\bar{y})}D(\rho,|\xi|)(zE + A(\sigma,\zeta))^{-1}E(I-P_1^-)M(\sigma,\zeta)dz,$$

with a fixed contour Γ_- contained in the open left half-plane. The part $\hat{k}_{S1}$ can then be treated as above.

The essential part is $\hat{k}_{S0}$, which is given by

$$\hat{k}_{S0}(y,\bar{y},\lambda,\xi) = \chi(\zeta)e^{\rho(z_1^-(\sigma,\zeta)y-\bar{y})}D(\rho,|\xi|)P_1^-(\sigma,\zeta)M(\sigma,\zeta).$$

Using the decomposition $x^0 = y^0 + \alpha e_1^-$ as above, this yields

$$\hat{k}_{S0}(y,\bar{y},\lambda,\xi) = \chi(\zeta)|\xi|e^{\rho(z_1^-(\sigma,\zeta)y-\bar{y})}D(\rho/|\xi|,1)e_1^-(\lambda,\xi)\otimes\alpha(\lambda,\xi).$$

In the Neumann case α is holomorphic, and

$$D(\rho/|\xi|,1)e_1^-(\lambda,\xi) = \left[0,0,-i\xi^{\mathsf{T}}\frac{\rho}{\lambda}, \frac{-|\xi|\rho}{\lambda}, 1\right]^{\mathsf{T}}$$

is bounded and satisfies the Mikhlin condition. Since $z_1^- \sim -|\xi|$ we obtain as above an estimate of the form

$$|\xi|^{|\alpha|}|D^\alpha_\xi \hat{k}_{S0}(y,\bar{y},\lambda,\xi)| \le \frac{M_\alpha}{y+\bar{y}},$$

where M_α is independent of $y, \bar{y}$, ξ and λ. Therefore we may argue as above to obtain

$$||k_{S0}(y,\bar{y},\lambda)|| \le \frac{c}{y+\bar{y}}, \quad y,\bar{y}>0,\ \lambda\in\Sigma_\phi.$$

The argument is more involved in the case of Dirichlet or slip conditions; it is here where the extra time regularity of the normal velocity h_2 comes in. As shown above, α decomposes as

$$\alpha(\lambda,\xi) = \alpha_0(\lambda,\xi) + \frac{\lambda}{|\xi|}\begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

where $\alpha_0(\lambda,\xi)$ is holomorphic. Since the term containing α_0 can be treated as before, we concentrate on the extra term. This yields the kernel k_{S00}, defined by

$$\hat{k}_{S00}(y,\bar{y},\lambda,\xi) = \chi(\zeta)|\xi|e^{\rho(z_1^-(\sigma,\zeta)y-\bar{y})}D(\rho/|\xi|,1)e_1^-(\lambda,\xi)\frac{\lambda}{|\xi|}\begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Since, by assumption, $\hat{h}_2$ is the Fourier–Laplace transform of a function of the class ${}_0H^1_p(J;\dot{W}_p^{-1/p}(\mathbb{R}^{n-1}))$, we see that $\lambda\hat{h}_2/|\xi|$ is the Fourier–Laplace transform of a function in $L_p(J;\dot{W}_p^{1-1/p}(\mathbb{R}^{n-1}))$. Thus we obtain $g\in L_p(J\times\mathbb{R}^n_+)$ such that

$$\hat{g}(\bar{y},\lambda,\xi) = e^{-|\xi|\bar{y}}\hat{h}_2(\lambda,\xi)\frac{\lambda}{|\xi|}.$$

Writing

$$\frac{\lambda}{|\xi|}\hat{h}_2 = 2\int_0^\infty |\xi|e^{-2|\xi|\bar{y}}\frac{\lambda}{|\xi|}\hat{h}_2 d\bar{y} = 2\int_0^\infty e^{-|\xi|\bar{y}}g(\bar{y})d\bar{y},$$

we have

$$|\xi|e^{\rho z_1^- y}D(\rho/|\xi|,1)e_1^-\frac{\lambda}{|\xi|}\hat{h}_2 = \int_0^\infty |\xi|e^{\rho z_1^- y-|\xi|\bar{y}}D(\rho/|\xi|,1)e_1^-\hat{g}(\bar{y},\lambda,\xi)d\bar{y},$$

and the kernel of this representation can be treated as before.

(vi) Summarizing, we have obtained a family

$$k(y,\bar{y},\lambda)\subset\mathcal{B}(L_p(J\times\mathbb{R}^{n-1};\mathbb{C}^{(2n+1)\times n}))$$

such that the family $\{(y+\bar{y})k(y,\bar{y},\lambda) : y,\bar{y}>0,\lambda\in\Sigma_\phi\}$ is also $\mathcal{R}$-bounded in $\mathcal{B}(L_p(J\times\mathbb{R}^{n-1};\mathbb{C}^{(2n+1)\times n}))$. In addition, $\hat{k}$ is holomorphic in $\lambda\in\Sigma_\phi$, and so the operator family $\{k(y,\bar{y},\lambda)\}_{\lambda\in\Sigma_\phi}$ also has this property; i.e.,

$$k(y,\bar{y},\cdot)\in\mathcal{H}^\infty(\Sigma_\phi;\mathcal{B}(L_p(J\times\mathbb{R}^{n-1};\mathbb{C}^{(2n+1)\times n}))),$$

with $\mathcal{H}^\infty$-norm less than $c/(y+\bar{y})$.

The operator ∂_t in $L_p(J;X)$ with natural domain ${}_0H^1_p(J;X)$ admits a bounded $\mathcal{H}^\infty$-calculus with $\mathcal{H}^\infty$-angle $\pi/2$, where $X=L_p(\mathbb{R}^{n-1};\mathbb{C}^{(2n+1)\times n})$. Therefore, by the Kalton–Weis theorem, $K(y,\bar{y}):=k_R(y,\bar{y},\partial_t)$ is bounded in $L_p(J;X)$ with bound

$$|K(y,\bar{y})|_{\mathcal{B}(L_p(J;X))}\le\frac{M}{y+\bar{y}}, \quad y,\bar{y}>0.$$

This shows that the integral operator in $L_p(\mathbb{R}_+; L_p(J;X))$ with operator-valued kernel $K(y,\bar{y})$ is dominated pointwise by the kernel $\kappa(y,\bar{y}) = \frac{1}{y+\bar{y}}$ of the one-sided Hilbert transform, which is well known to be bounded in $L_p(\mathbb{R}_+; Z)$, for any Banach space Z, provided $1 < p < \infty$.

This completes the proof of Theorem 6.1. □

We can easily extend Theorem 6.1 to the case of variable coefficients with small deviation from constant ones. To see this, let $\mathcal{A}(t,x,D) = \mathcal{A}_0(D) + \mathcal{A}_1(t,x,D)$, where

$$\sup\{|a_{1,ij}^{kl}| : k,l,i,j = 1,\ldots,n,\ t \in J,\ x \in \mathbb{R}^n\} \le \eta$$

and

$$a_{1,ij}^{kl} \in W_s^{1/2-1/2p}(J; L_r(\Gamma_s \cup \Gamma_n)) \cap L_s(J; W_r^{1-1/p}(\Gamma_s \cup \Gamma_n)),$$

for some $r, s \ge p$ such that $\frac{1}{s} + \frac{n-1}{2r} < 1 - \frac{1}{p}$. Let S denote the solution operator of the generalized Stokes problem (4.1) from Theorem 6.1 for $\mathcal{A}_0(D)$, and T that of the perturbed problem. Then we obtain the identity

$$T = S + SBT, \qquad \text{where} \qquad B = \begin{bmatrix} -\mathcal{A}_1(t,x,D) & 0 \\ 0 & 0 \\ -\mathcal{B}_1(t,x,D) & 0 \end{bmatrix},$$

in the case of slip or Neumann condition. Here $\mathcal{B}_1$ has the obvious meaning of the corresponding boundary operator generated by the perturbation $\mathcal{A}_1$. The norm of the first component of B as an operator from the maximal regularity space Z into $L_p(J \times \mathbb{R}^n; \mathbb{C}^n)$ is bounded by $C\eta$, where $C > 0$ denotes a constant independent of η, and the norm of its third component in the boundary space X_∂ is estimated by

$$|B_3 v|_{X_\partial} \le C\eta |v|_Z + C \max_{i,j,k,l} |a_{1,ij}^{kl}|_{W_s^{1/2-1/2p}(L_r) \cap L_s(W_r^{1-1/p})} |v|_Z^{\alpha} |v|_{L_p}^{1-\alpha}$$

for some $\alpha \in (0,1)$, due to the Gagliardo–Nirenberg inequality and the restriction on r and s. Then, as in the previous section, a Neumann series argument shows that $T = (I - SB)^{-1} S$ in fact exists and is bounded as a map from the data space to the maximal regularity space as well. Let us state this as follows.

Corollary 6.2. *The assertions of Theorem* 6.1 *remain valid in the case of variable coefficients* $\mathcal{A}(t,x,D) = \mathcal{A}_0(D) + \mathcal{A}_1(t,x,D)$, *provided*

$$\sup\{|a_{1,ij}^{kl}(t,x)| : k,l,i,j = 1,\ldots,n,\ t \in J,\ x \in \mathbb{R}^n\} \le \eta,$$

for some sufficiently small $\eta > 0$, *and*

$$a_{1,ij}^{kl} \in W_s^{1/2-1/2p}(J; L_r(\Gamma_s \cup \Gamma_n)) \cap L_s(J; W_r^{1-1/p}(\Gamma_s \cup \Gamma_n)),$$

for some $r, s \ge p$ *such that* $\frac{1}{s} + \frac{n-1}{2r} < 1 - \frac{1}{p}$.

7. Proof of Theorem 4.1: Localization.

Step 1. We first deal with the case $G = \mathbb{R}^n$, i.e., the problem without boundary conditions. Note that the generalized Stokes problems with coefficients frozen at any $t \in J$, $x \in \mathbb{R}^n \cup \{\infty\}$ have maximal L_p-regularity according to Theorem 4.1. Because of uniform continuity of the coefficients on $J \times (\mathbb{R}^n \cup \{\infty\})$ the maximal regularity constant is uniform as well, as are the norms of the solution operators S_j. Given

$\eta > 0$, cut the interval $J = [0, a]$ into pieces $J_i = [i\delta, i\delta + \delta]$ of equal length $\delta = a/l$ such that the coefficients $a(t, x)$ appearing in $\mathcal{A}(t, x, D)$ satisfy

$$\sup\{|a(t,x) - a(s,x)| : x \in \mathbb{R}^n,\ t, s \in J,\ |t-s| \le \delta\} \le \eta.$$

We then solve the problem successively on the intervals J_i, $i = 0, \ldots, l$, so w.l.o.g. it is enough to treat the first interval $J_0 = [0, \delta]$. The number $\delta > 0$ will be chosen later.

By Theorem 10.1, we may assume $v_0 = f = 0$. Choose a large ball $B_R(0)$ such that

$$\sup\{|a(t,x) - a(0,\infty)| : |x| \ge R,\ t \in J_0\} \le \eta.$$

Cover the ball $\bar{B}_R(0)$ by finitely many balls $B_r(x_k)$, $k = 1, \ldots, N$, such that

$$\sup\{|a(t,x) - a(0,x_k)| : x \in B_r(x_k),\ t \in J_0\} \le \eta.$$

Fix a C^∞-partition of unity ϕ_k, which is subordinate to the covering $\bigcup_{k=1}^N B_r(x_k) \cup \bar{B}_R(0)^c$ of $\mathbb{R}^n$. The index $k = 0$ corresponds to the chart at infinity. Define local operators $\mathcal{A}_k(D) = \mathcal{A}(t, x, D)$ for each chart $B_r(x_k)$, $k = 1, \ldots, N$, and $\mathcal{A}_0(D) = \mathcal{A}(t, x, D)$; extend these coefficients to all of $J \times \mathbb{R}^n$ by, say, reflection at the boundary of the corresponding ball. Corollary 5.2 shows that each of these operators has maximal regularity, provided that $\eta > 0$ and $\delta > 0$ are sufficiently small.

Suppose (v, q) is a solution of (4.1) with $G = \mathbb{R}^n$. In what follows we normalize the pressure by $\int_{B_{2R}(0)} q(t,x)dx = 0$. Setting $v_k = \phi_k v$, $q_k = \phi_k q$, $g_k = \phi_k g$, we obtain the following problem for the functions v_k and q_k:

$$\begin{aligned} \partial_t v_k + \mathcal{A}_k(D)v_k + \nabla q_k &= (\nabla\phi_k)q + [\mathcal{A}, \phi_k]v, \\ \operatorname{div} v_k &= g_k + (\nabla\phi_k|v), \quad t \in J_0,\ x \in \mathbb{R}^n, \\ v_k|_{t=0} &= 0, \end{aligned} \tag{7.1}$$

where $[\mathcal{A}, \phi_k]v = \mathcal{A}(\phi_k v) - \phi_k \mathcal{A}v$ means the commutator of $\mathcal{A}$ and ϕ_k. Denote the solution operator of the generalized Stokes problem for $\mathcal{A}_k$ by S^k. Then we have the representation

$$\begin{bmatrix} v_k \\ q_k \end{bmatrix} = S^k \begin{bmatrix} (\nabla\phi_k)q + [\mathcal{A}, \phi_k]v \\ g_k + (\nabla\phi_k|v) \end{bmatrix}.$$

Summing over all charts k, we deduce

$$\begin{bmatrix} v \\ q \end{bmatrix} = \sum_{k=0}^N \begin{bmatrix} v_k \\ q_k \end{bmatrix} = \sum_{k=0}^N S^k \begin{bmatrix} (\nabla\phi_k)q + [\mathcal{A}, \phi_k]v \\ g_k + (\nabla\phi_k|v) \end{bmatrix}.$$

We decompose this representation of the solution as

$$\begin{bmatrix} v \\ q \end{bmatrix} = \sum_{k=0}^N S^k \begin{bmatrix} 0 \\ g_k \end{bmatrix} + T \begin{bmatrix} q \\ v \end{bmatrix} + Rv,$$

where

$$T = \sum_{k=0}^N S^k \nabla\phi_k \quad \text{and} \quad R = \sum_{k=0}^N S^k \begin{bmatrix} [\mathcal{A}, \phi_k] \\ 0 \end{bmatrix}.$$

We estimate T and R separately. For this purpose, we define the spaces

$$Z = [H_p^1(J_0; L_p(\mathbb{R}^n;\mathbb{C}^n)) \cap L_p(J_0; H_p^2(\mathbb{R}^n;\mathbb{C}^n))] \times L_p(J_0; \dot{H}_p^1(\mathbb{R}^n))$$

and

$$X = L_p(J_0 \times \mathbb{R}^n;\mathbb{C}^n), \quad X_p = W_p^{2-2/p}(\mathbb{R}^n;\mathbb{C}^n),$$

as well as

$$Y = H_p^1(J_0; \dot{H}_p^{-1}(\mathbb{R}^n)) \cap L_p(J_0; H_p^1(\mathbb{R}^n)).$$

To begin with T, recall that each S^k splits into $S^k = S_0 + S_1^k$, with the same S_0 for each k, since the latter does not depend on the coefficients of $\mathcal{A}_k$. Hence

$$T = \sum_{k=0}^N S^k \nabla\phi_k = \sum_{k=0}^N S_1^k \nabla\phi_k + S_0 \nabla \sum_{k=0}^N \phi_k = \sum_{k=0}^N S_1^k \nabla\phi_k,$$

since ϕ_k forms a partition of unity. Let us decompose T into its components, employing the factorization of S_1 obtained in section 5. We have

$$\begin{aligned}
T_{11}q &= (\partial_t - \Delta)^{-1} \sum_k S_{11}^k (-\Delta)(\partial_t - \Delta)^{-1}(q\nabla\phi_k),\\
T_{21}q &= (-\Delta)^{-1/2} \sum_k S_{21}^k (-\Delta)(\partial_t - \Delta)^{-1}(q\nabla\phi_k),\\
T_{12}v &= (\partial_t - \Delta)^{-1} \sum_k S_{12}^k (-\Delta)^{1/2}(\nabla\phi_k|v),\\
T_{22}v &= (-\Delta)^{-1/2} \sum_k S_{22}^k (-\Delta)^{1/2}(\nabla\phi_k|v).
\end{aligned}$$

Since $\nabla\phi_k$ has compact support also for $k = 0$, we see that $(\nabla\phi_k)q$ belongs to $L_p(J_0; H_p^1(\mathbb{R}^n))$, and

$$|(-\Delta)^{1/2}(q\nabla\phi_k)|_{L_p(J_0\times\mathbb{R}^n)} \le C|\nabla q|_{L_p(J_0\times\mathbb{R}^n)}$$

holds with some constant $C > 0$; recall the normalization of the pressure

$$\int_{B_{2R}(0)} q(t,x)dx = 0;$$

hence Poincaré's inequality is valid. Therefore

$$|(-\Delta)(\partial_t - \Delta)^{-1}(\nabla\phi_k q)|_{H_p^{1/2}(J_0;L_p(\mathbb{R}^n))} \le C|\nabla q|_{L_p(J_0\times\mathbb{R}^n)}.$$

Since $H_p^{1/2} \hookrightarrow L_{2p}$, this implies that the norm of $(-\Delta)(\partial_t - \Delta)^{-1}(\nabla\phi_k q)$ in $L_p(J_0\times\mathbb{R}^n)$ is dominated by $C\delta^{1/2p}|\nabla q|_{L_p(J_0\times\mathbb{R}^n)}$, and the constant $C\delta^{1/2p}$ can be made small by choice of $\delta > 0$.

Similarly, $\nabla\phi_k$ is a multiplier for the space $H_p^s(\mathbb{R}^n)$, and hence there is a constant $C > 0$ such that

$$|(-\Delta)^{1/2}(\nabla\phi_k|v)|_{H_p^{1/2}(J_0;L_p(\mathbb{R}^n))} \le C|(\partial_t - \Delta)v|_{L_p(J_0\times\mathbb{R}^n)}.$$

By this estimate we may conclude as before that the norm of $(-\Delta)^{1/2}(\nabla\phi_k v)$ in $L_p(J_0\times\mathbb{R}^n;\mathbb{C}^n)$ is dominated by $C\delta^{1/2p}|(\partial_t-\Delta)v|_{L_p(J_0\times\mathbb{R}^n;\mathbb{C}^n)}$. As a consequence, the operator T satisfies

$$\left|T\begin{bmatrix} q\\ v\end{bmatrix}\right|_Z \le C\delta^{1/2p}\left|\begin{bmatrix} v\\ q\end{bmatrix}\right|_Z.$$

Next, R is given by

$$R=\sum_k S^k\begin{bmatrix} [\mathcal{A},\phi_k]v\\ 0\end{bmatrix}.$$

The commutator $[\mathcal{A},\phi_k]$ is a differential operator of first order, and hence we see that

$$[\mathcal{A},\phi_k]v\in H^{1/2}_p(J_0;L_p(\mathbb{R}^n;\mathbb{C}^n))\cap L_p(J_0;H^1_p(\mathbb{R}^n;\mathbb{C}^n)).$$

Therefore, we obtain

$$\begin{aligned}|Rv|_Z&\le C\max_k|[\mathcal{A},\phi_k]v|_{L_p(J_0\times\mathbb{R}^n;\mathbb{C}^n)}\\ &\le C\delta^{1/2p}\max_k|[\mathcal{A},\phi_k]v|_{L_{2p}(J;L_p(\mathbb{R}^n;\mathbb{C}^n))}\le C\delta^{1/2p}|v|_Z;\end{aligned}$$

i.e., the norm of R in Z is also bounded by $C\delta^{1/2p}$, which becomes small by the smallness of $\delta>0$.

The above arguments show that, choosing first $\eta>0$ and then $\delta>0$ small enough, there is a constant $C>0$ such that the estimate

$$|(v,q)|_Z\le C[|f|_X+|g|_Y+|v_0|_{X_p}] \tag{7.2}$$

holds for each solution (v,q) on J_0 for given data $v_0\in X_p$, $f\in X$, $g\in Y$. Therefore, the generalized Stokes operator $L\in\mathcal{B}(Z;X\times Y\times X_p)$ defined by the first lines of the left-hand side of (4.1) is injective and has closed range; hence it is semi-Fredholm for each set of coefficients which are continuous on $J_0\times\mathbb{R}^n$, admit uniform limits as $|x|\to\infty$, and are strongly elliptic, uniformly on $J_0\times(\mathbb{R}^n\cup\{\infty\})$. Define the family $\mathcal{A}_\tau=\tau\mathcal{A}+(1-\tau)(-\Delta)$; we then may conclude that for each $\tau\in[0,1]$, the corresponding generalized Stokes operator L_τ is injective and has closed range. By the continuity of the Fredholm index, it must be constant; i.e., the index is zero for all $\tau\in[0,1]$ since L_0 is bijective by Theorem 4.1. This shows that $L=L_1$ is also surjective, and hence the proof is complete in case $G=\mathbb{R}^n$.

Step 2. Next we consider the case of a half-space, i.e., $G=\mathbb{R}^n_+$. We extend the coefficients $a^{kl}_{ij}(t,x)$ of $\mathcal{A}(t,x,D)$ by symmetry to $J\times\mathbb{R}^n$. According to Proposition 4.2 we may assume $v_0=g=0$, and also $h_\nu=0$ in the case of Dirichlet or slip conditions. We localize as in Step 1 to obtain

$$\begin{aligned}\partial_t v_k+\mathcal{A}_k(D)v_k+\nabla q_k&=f_k+(\nabla\phi_k)q+[\mathcal{A},\phi_k]v,\\ \operatorname{div}v_k&=(\nabla\phi_k|v),\quad t\in J_0,\ x\in\mathbb{R}^n,\\ v_k|_{t=0}&=0,\end{aligned}\tag{7.3}$$

for charts which are interior to $\mathbb{R}^n_+$ (according to Step 1 only one such chart is sufficient), and

$$\begin{aligned}\partial_t v_k+\mathcal{A}_k(D)v_k+\nabla q_k&=f_k+(\nabla\phi_k)q+[\mathcal{A},\phi_k]v,\\ \operatorname{div}v_k&=(\nabla\phi_k|v),\quad t\in J_0,\ x\in\mathbb{R}^n_+,\\ v_k|_{t=0}&=0,\\ \mathcal{B}_k(D)(v_k,q_k)&=h_k+[\mathcal{B},\phi_k]v\end{aligned}\tag{7.4}$$

for charts intersecting the boundary. We concentrate now on the charts intersecting the boundary; the interior chart is treated as in Step 1. Decompose $v_k = \nabla\psi_k + v_k^2 + v_k^3$, $q_k = q_k^2 + q_k^3$ as follows. For Dirichlet or slip conditions, ψ_k is taken as the solution of the problem

$$\Delta\psi_k = (\nabla\phi_k|v),\ x \in \mathbb{R}^n_+, \quad \partial_\nu\psi_k = 0,\ x \in \partial\mathbb{R}^n_+.$$

Thus, $\psi_k \in {}_0H^1_p(J; \dot{H}^2_p(\mathbb{R}^n_+)) \cap L_p(J; \dot{H}^4_p(\mathbb{R}^n_+))$; hence $\nabla\psi_k = R_0(\nabla\phi_j v)$ belongs to ${}_0H^1_p(J; \dot{H}^1_p(\mathbb{R}^n_+)) \cap L_p(J; \dot{H}^3_p(\mathbb{R}^n_+))$, and $\sum_k \nabla\psi_k = R_0 \sum_k \nabla\phi_k v = 0$. On the other hand, for $\theta \in \dot{H}^1_{p'}(\mathbb{R}^n_+)$ we have

$$\int_{\mathbb{R}^n_+} (\nabla\phi_k) v\theta dx = -\int_{\mathbb{R}^n_+} \phi_k v \nabla\theta dx,$$

since div $v = (v|\nu) = 0$; hence $\nabla\phi_k v \in {}_0H^1_p(J; \dot{H}^{-1}_p(\mathbb{R}^n_+))$, and consequently $\nabla\psi_k \in {}_0H^1_p(J; L_p(\mathbb{R}^n_+))$. Therefore, we see $\nabla\psi_k \in {}_0H^1_p(J; H^1_p(\mathbb{R}^n_+)) \cap L_p(J; H^3_p(\mathbb{R}^n_+))$. In case of Neumann boundary conditions we use instead

$$\Delta\psi_k = (\nabla\phi_k|v),\ x \in \mathbb{R}^n_+, \quad \psi_k = 0,\ x \in \partial\mathbb{R}^n_+.$$

Next (v_k^2, q_k^2) denotes the full-space solution of the problem

$$\begin{aligned} \partial_t v_k^2 + \mathcal{A}_k(D) v_k^2 + \nabla q_k^2 &= f_k + \nabla\phi_k q + [\mathcal{A}, \phi_k] v - (\partial_t + \mathcal{A}_k(D))\nabla\psi_j, \\ \operatorname{div} v_k &= 0, \quad t \in J_0,\ x \in \mathbb{R}^n, \\ v_k|_{t=0} &= 0. \end{aligned} \tag{7.5}$$

With the same notation as in Step 1 we then obtain

$$\begin{bmatrix} v^2 \\ q^2 \end{bmatrix} = \sum_{k=0}^N \begin{bmatrix} v_k^2 \\ q_k^2 \end{bmatrix} = \sum_{k=0}^N S^k \begin{bmatrix} f_k - (\partial_t + \mathcal{A}_k(D))\nabla\psi_k + \nabla\phi_k q + [\mathcal{A}, \phi_k] v \\ 0 \end{bmatrix}.$$

Since $(\partial_t + \mathcal{A}_k(D))\nabla\psi_k \in L_p(J; H^1_p(\mathbb{R}^n_+))$ we may argue as in Step 1 to estimate (v^2, q^2).

Thus it remains to estimate (v^3, q^3), which is a solution of the problem

$$\begin{aligned} \partial_t v_k^3 + \mathcal{A}_k(D) v_k^3 + \nabla q_k^3 &= 0, \\ \operatorname{div} v_k^3 &= 0, \quad t \in J_0,\ x \in \mathbb{R}^n_+, \\ v_k^3|_{t=0} &= 0, \\ \mathcal{B}_k(D)(v_k^3, q_k^3) &= h_k + [\mathcal{B}, \phi_k] v - \mathcal{B}_k(D)(\nabla\psi_k + v_k^2, q_k^2). \end{aligned} \tag{7.6}$$

To achieve this, note that the commutator $[\mathcal{B}, \phi_k]$ is either zero in the case of Dirichlet conditions or an operator of order zero and hence gains time regularity. Since $\nabla\psi_k$ belongs to the space ${}_0H^1_p(J; H^1_p(\mathbb{R}^n_+)) \cap L_p(J; H^3_p(\mathbb{R}^n_+))$, the term $\mathcal{B}_k(D)\nabla\psi_k$ has also more time regularity than required by the boundary space for h. Next (v_k^2, q_k^2) splits into

$$(v_k^2, q_k^2) = S^k f_k + S_0 \bar{f}_k + S_1^k \bar{f}_k,$$

where

$$\bar{f}_k = [\mathcal{A}, \phi_k] v + (\nabla\phi_k) q - (\partial_t + \mathcal{A}_k(D))\nabla\psi_k$$

belongs to $L_p(J;H^1_p(\mathbb{R}^n_+))$. S_0 maps this space into ${}_0H^1_p(J;H^1_p(\mathbb{R}^n))\cap L_p(J;H^3_p(\mathbb{R}^n))$, and hence $\mathcal{B}_k(D)S_0f_k$ also gains extra regularity in time. Due to the factorization of S_1^j, the term $\mathcal{B}_k(D)S_1^kf_k$ will be small, as in Step 1.

Thus, as in Step 1, we obtain an a priori estimate of the form

$$|(v,q)|_Z \le C[|f|_X + |g|_Y + |v_0|_{X_p} + |h|_{Y_\partial}], \tag{7.7}$$

by choosing first $\eta>0$ and then $\delta>0$ small enough. Here Y_∂ denotes the relevant boundary space. The Fredholm argument at the end of Step 1 again completes the proof for the half-space case.

Step 3. The localization procedure for general domains with compact boundary of class C^{3-} follows the lines of Step 2. The only difference concerns that the charts intersecting the boundary cannot be handled by Theorem 6.1 directly, since they lead to problems in perturbed half-spaces. Therefore we show in this step how to transfer Theorem 6.1 to perturbed half spaces.

For this purpose, consider a graph ρ over $\mathbb{R}^{n-1}$, and let G be the corresponding epigraph

$$G=\{x=(x',x_n)\in\mathbb{R}^n:\, x_n>\rho(x')\}.$$

We may assume that ρ is of class C^{3-} and has compact support, and $|\nabla\rho|_{L^\infty}\le\eta$, where $\eta>0$ is sufficiently small. Suppose that (u,π) is a solution in the maximal regularity class of the generalized Stokes problem (4.1) with given data f, g, h, u_0 subject to (D). According to Proposition 4.2 we may assume $u_0=g=h_\nu=0$. Introduce the pull back (v,q) defined on $\mathbb{R}^n_+$ by

$$v(t,x)=u(t,x+\rho(x')e_n),\quad q(t,x)=\pi(t,x+\rho(x')e_n),$$

where $t\in J$, $x=(x',x_n)$, $x'\in\mathbb{R}^{n-1}$, and $x_n>0$. Then (v,q) satisfies the perturbed problem

$$\begin{aligned}\partial_t v+\mathcal{A}(D)v+\nabla q &= f+\mathcal{A}_\rho(D)v+(\nabla\rho)\partial_n q,\\ \operatorname{div} v &= (\nabla\rho|\partial_n v),\quad t\in J,\ x\in\mathbb{R}^n_+,\\ \mathcal{B}(D)(v,q) &= \mathcal{B}_\rho(D)v+h,\quad t\in J,\ x=(x',0),\ x'\in\mathbb{R}^{n-1},\\ v_{|t=0} &= 0,\quad x\in\mathbb{R}^n_+.\end{aligned} \tag{7.8}$$

Here, f and h also denote the transformed data, which belong to the right regularity class, since $\rho\in C^{3-}$. The operator $\mathcal{A}_\rho(D)$ is given by

$$\mathcal{A}_\rho(D)=a^{kl}_{ij}[\partial_k\rho D_lD_n+\partial_l\rho D_kD_n+\partial_k\rho\partial_l\rho D_n^2+i\partial_k\partial_l\rho D_n].$$

Since the second order derivatives in $\mathcal{A}_\rho(D)$ carry the factor $\nabla\rho$, they are small by the condition $|\nabla\rho|_{L^\infty}\le\eta$, as is $(\nabla\rho)\partial_n q$. Next, for the same reason, the term $(\nabla\rho|\partial_n v)$ in the divergence equation is also small in $L_p(J;H^1_p(\mathbb{R}^n_+))$.

The normal ν at ∂G is given by

$$\nu(x+\rho(x')e_n)=\begin{bmatrix}\nabla_{x'}\rho\\ -1\end{bmatrix}\frac{1}{\sqrt{1+|\nabla_{x'}\rho|^2}};$$

therefore the condition $(u|\nu)=0$ transforms to $(v|-e_n)=-(\nabla\rho|v)$. Thus we have to verify that the pair $((\nabla\rho|\partial_n v),-(\nabla\rho|v)_{|x_n=0})$ is small in $H^1_p(J;\dot H^{-1}_p(\mathbb{R}^n_+))$. For this purpose, let $\phi\in\dot H^1_{p'}(\mathbb{R}^n_+)$; then

$$\int_{\mathbb{R}^n_+}\phi(\partial_n v|\nabla\rho)dx-\int_{\partial\mathbb{R}^n_+}-\phi(\nabla\rho|v)dx'=-\int_{\mathbb{R}^n_+}(v|\nabla\rho)\partial_n\phi dx,$$

and hence

$$|((\nabla\rho|\partial_n v), -(\nabla\rho|v)_{|x_n=0})|_{H^1_p(\dot H^{-1}_p)} \le |\nabla\rho|_{L^\infty}|v|_{H^1_p(L_p)},$$

thereby showing the smallness of this term.

Finally, we show smallness of the boundary perturbation $\mathcal{B}_\rho(D)$. We have

$$a^{kl}_{ij}\nu_k D_l u_j = \frac{1}{\sqrt{1+|\nabla\rho|^2}} a^{kl}_{ij}(\partial_k\rho - \delta_{kn})(D_l v - \partial_l\rho D_n v);$$

hence

$$\mathcal{B}_\rho(D) = \left(1 - \frac{1}{\sqrt{1+|\nabla\rho|^2}}\right) a^{nl}_{ij} D_l + \frac{1}{\sqrt{1+|\nabla\rho|^2}}[a^{nl}_{ij}\partial_l\rho D_n + a^{kl}_{ij}\partial_k\rho(D_l - \partial_l D_n)]$$

in the case of Neumann conditions, and a similar expression in the slip case. Since

$$1 - \frac{1}{\sqrt{1+|\nabla\rho|^2}} = \frac{|\nabla\rho|^2}{(1+\sqrt{1+|\nabla\rho|^2})\sqrt{1+|\nabla\rho|^2}},$$

each term in $\mathcal{B}_\rho(D)$ carries a factor $\nabla\rho$; hence, as at the end of section 6, we have an estimate of the form

$$|\mathcal{B}_\rho(D)v|_{W^{1/2-1/2p}_p(L_p)\cap L_p(W^{1-1/p}_p)} \le C|\nabla\rho|_{L_\infty}|v|_Z + C|\nabla^2\rho|_{L_\infty}|v|^\alpha_Z|v|^{1-\alpha}_{L_p},$$

for some $\alpha \in (0,1)$. This shows that the boundary operator $\mathcal{B}_\rho(D)$ is also small, provided that $\eta > 0$ is small enough.

Therefore, we may apply once more a Neumann series argument to conclude from Theorem 6.1 that this result is also valid in perturbed half-spaces. This completes the proof of Theorem 4.1. □

8. The nonlinear problem with homogeneous slip and nonslip conditions. The nonlinear problem with $\Gamma_n = \emptyset$ will be solved by means of an abstract result which is essentially due to Clément and Li [3]. This is possible since in this case the involved boundary conditions are actually linear and homogeneous. We describe a version of the abstract result proved in Prüss [21].

Let X_0, X_1 be Banach spaces with norms $|\cdot|_0$, $|\cdot|_1$, $X_1 \hookrightarrow X_0$ densely, $J = [0,a]$, $J_0 = [0,a_0]$, and let $1 < p < \infty$. Consider the quasi-linear problem

$$\dot u(t) + A(u(t))u(t) = F(u(t)), \quad t \in J,\ u(0) = u_0. \tag{8.1}$$

Here $u_0 \in X_p := (X_0, X_1)_{1-1/p,p}$, $A : X_p \to \mathcal{B}(X_1, X_0)$ is continuous, and $F : X_p \to X_0$ is continuous. Moreover, we assume the following conditions on the Lipschitz continuity of A and F:

(A) For each $R > 0$ there is a constant $L(R) > 0$ such that

$$|A(u)v - A(\bar u)v|_0 \le L(R)|u-\bar u|_p|v|_1, \qquad u,\bar u \in X_p, |u|_p, |\bar u|_p \le R,\ v \in X_1.$$

(F) For each $R > 0$ there is a constant $l(R)$ such that

$$|F(u) - F(\bar u)|_0 \le l(R)|u-\bar u|_p, \quad u,\bar u \in X_p, |u|_p, |\bar u|_p \le R.$$

In the situation described above we have the following theorem.

THEOREM 8.1. *Suppose that the assumptions* (A) *and* (F) *are satisfied, and assume that $A(u)$ has the property of maximal L_p-regularity for each $u \in X_p$.*

Then (8.1) *admits a unique solution u on a maximal time interval $J(u_0) = [0, t_+(u_0))$ in the maximal regularity class*

$$u \in H^1_p(J; X_0) \cap L_p(J; X_1) \quad \textit{for each } a < t_+(u_0).$$

In case $t_+(u_0) < a_0$, the maximal time $t_+(u_0)$ is characterized by the equivalent conditions

$$\int_{J(u_0)} [|u(t)|_1^p + |\dot{u}(t)|_0^p] dt = \infty$$

and

$$\lim_{t \to t^+(u_0)-} u(t) \quad \textit{does not exist in } X_p.$$

The map $u_0 \mapsto u(t)$ defines a local semiflow on the natural phase space X_p.

Now we apply this result directly to the problem (2.2) under consideration, for a proof of Theorem 2.1 in case $\Gamma_n = \emptyset$. For this purpose we let

$$X_0 = \{u \in L_p(G, \mathbb{R}^n) :\ \operatorname{div} u = 0\},$$

$$X_1 = \{u \in H^2_p(G) \cap X_0 :\ u|_{\Gamma_0} = (u|\nu)|_{\Gamma_s} = 0,\ \mathcal{PE}(u)|_{\Gamma_s} = 0\}.$$

The operator family $A(u)$ will be defined by $A(u) = P\mathcal{A}(u, D)$ with $\mathcal{A}(u, D)$ as in section 2; here P denotes the standard Helmholtz projection. Theorem 4.1 implies that for each $u \in X_p \hookrightarrow W^{2-2/p}_p(G)$ the operator $A(u)$ has maximal L_p-regularity. Note that the involved boundary conditions are actually linear and homogeneous. This is obvious for the Dirichlet conditions $u|_{\Gamma_0} = (u|\nu)|_{\Gamma_s} = 0$, but is also true for the seemingly nonlinear condition $\mathcal{S}\nu - (\mathcal{S}\nu|\nu)\nu = 0$. In fact, the latter is equivalent to $\mathcal{E}\nu - (\mathcal{E}\nu|\nu)\nu = 0$, which is linear.

It is easy to check by Sobolev embedding that condition (A) holds, provided $p > n + 2$. Finally, we set $F(u) = -P(u \cdot \nabla u)$ and check, once more by Sobolev embedding, that condition (F) is satisfied as well. Thus we may apply Theorem 8.1 to prove Theorem 2.1 in the case $\Gamma_n = \emptyset$. We want to emphasize that this approach works as long as the boundary conditions are homogeneous and linear, so it does not apply for the Neumann problem since this one involves a truly nonlinear boundary condition.

9. The nonlinear problem with general boundary conditions. We now consider the nonlinear problem with general inhomogeneous boundary conditions, i.e.,

$$\begin{aligned}
&\partial_t u + \operatorname{div}\,(u \otimes u) = \operatorname{div}\, S + f,\\
&S = \mu[\nabla u + (\nabla u)^{\mathsf{T}}] - \pi I,\\
&\operatorname{div}\, u = g, \quad t > 0,\ x \in G,\\
&u(t,x) = h_0, \quad t > 0,\ x \in \Gamma_0,\\
&S\nu - (S\nu|\nu)\nu = h_s,\ (u|\nu) = h_\nu, \quad t > 0,\ x \in \Gamma_s,\\
&S\nu = h_n, \quad x \in \Gamma_n,\\
&u(0,x) = u_0(x),\ x \in G.
\end{aligned} \tag{9.1}$$

Assume $p > n+2$ in what follows, and let $\mu \in C^{2-}(\mathbb{R}_+)$ be a function subject to (2.3). Fix a time interval $J_0 = [0, a_0]$ and let $J_a = [0, a]$ for $a \le a_0$. We define the maximal regularity spaces

$$Z(a) := H_p^1(J_a; L_p(G; \mathbb{R}^n)) \cap L_p(J_a; H_p^2(G; \mathbb{R}^n))$$

and set $Z(0) = Z(a_0)$. We further set

$$X(a) := L_p(J_a; L_p(G; \mathbb{R}^n)), \quad X_p := W_p^{2-2/p}(G; \mathbb{R}^n),$$
$$Y_j(a) := W_p^{1/2-1/2p}(J_a; L_p(\Gamma_j; \mathbb{R}^n)) \cap L_p(J_a; W_p^{1-1/p}(\Gamma_j; \mathbb{R}^n)),$$

and $X_0 = X(a_0)$, $Y_{j0} = Y(j(a_0)$. Note that we have the embedding

$$Z(a) \hookrightarrow C(J_a; X_p)$$

since X_p is the time-trace space of Z_a. However, the embedding constant blows up as $a \to 0+$. This unpleasant fact can be removed if one restricts attention to functions with time trace 0 at $t = 0$. Therefore we let

$${}_0Z(a) = \{u \in Z(a) : u|_{t=0} = 0\}, \quad {}_0Y_j(a) = \{u \in Y_j(a) : u|_{t=0} = 0\}.$$

We also have the embedding

$$Y_j(a) \hookrightarrow C(J_a; W_p^{1-3/p}(\Gamma_j; \mathbb{R}^n)), \quad j = s, n,$$

where the embedding constant is uniform in a if the embedding is restricted to ${}_0Y_j(a)$. In particular, since $p > n+2$ by assumption, we have

$$Z_0 \hookrightarrow C(J_0; C^1(\overline{G}; \mathbb{R}^n)).$$

Now suppose that $u_0 \in X_p$, $f \in X_0$, $g \in L_p(J_0; H_p^1(G))$, $h_j \in Y_{j0}$, and $(g, h_\nu) \in H_p^1(J_0; H_{p,\Gamma_n}^{-1}(G))$ are given such that the compatibility conditions

$$\operatorname{div} u_0 = g|_{t=0}, \quad u|_{\Gamma_0} = h_0, \quad (u_0|\nu)|_{\Gamma_s} = h_\nu,$$

and

$$(h_s|\nu) = 0, \quad 2\mu(|\mathcal{E}(u_0)|_2^2)\mathcal{P}\mathcal{E}(u_0)\nu = \mathcal{P}h_j|_{t=0}, \quad j = s, n,$$

are valid, where as before $\mathcal{E}(u) = \frac{1}{2}(\nabla u + \nabla u)^{\mathsf{T}}$. To achieve reduction to time traces 0 at $t = 0$, first choose extensions h_{*j} of $2(1 - \mu(|\mathcal{E}(u_0)|_2^2))\mathcal{E}(u_0)\nu$ in Y_{j0}, which is possible since $u_0 \in W_p^{1-3/p}(\Gamma; \mathbb{R}^n)$. Then solve the classical Stokes equation with $\mu = 1$, with initial value $u_*|_{t=0} = u_0$, right-hand sides f, $\operatorname{div} u_* = g$, Dirichlet conditions $u_*|_{\Gamma_0} = h_0$ on Γ_0, slip conditions $(u_*|\nu) = h_\nu$, $2\mathcal{P}\mathcal{E}(u_*)\nu = h_s + \mathcal{P}h_{*s}$ on Γ_s, and Neumann condition $2\mathcal{E}(u_*)\nu - \pi_*\nu = h_n + h_{*n}$. Note that the map

$$(u_0, f, g, h_0, h_\nu, h_s, h_n) \mapsto (u_*, \pi_*)$$

is bounded linear; i.e., (u_*, π_*) depends continuously on the data in the correct spaces, by Theorem 4.1.

Suppose that (u,π) is a solution of (1.1) in the maximal regularity class of type L_p on an interval J_a. Set $u_1 = u - u_*$ and $\pi_1 = \pi - \pi_*$. Then the time-traces of u_1 and $\pi|_{\Gamma_n}$ are zero, and the pair (u_1,π_1) is a solution of the quasi-linear problem

$$
\begin{aligned}
&\partial_t u + \mathcal{A}(t,x,D)u + \nabla\pi = \bar f(t,x) + F(u),\\
&\operatorname{div} u = 0, \quad t \in J_a,\ x \in G,\\
&u|_{\Gamma_0} = (u|\nu)|_{\Gamma_s} = u|_{t=0} = 0,\\
&\mathcal{P}[\nu_k \mathcal{A}^{kl}(t,x)D_l u] = \mathcal{P}\bar h_s + \mathcal{P}H(u) \quad \text{ on } J_a \times \Gamma_s,\\
&\nu_k \mathcal{A}^{kl}(t,x)D_l u - \pi\nu = \bar h_n + H(u) \quad \text{ on } J_a \times \Gamma_n,
\end{aligned}
\tag{9.2}
$$

where we have once more used the sum convention and $\mathcal{A}(t,x,D) = \mathcal{A}^{kl}(t,x)D_k D_l$. Here $\mathcal{A}^{kl}(t,x) = \mathcal{A}^{kl}(u_*(t,x)) = a^{kl}_{ij}(u_*(t,x))$ denote the matrices

$$\mathcal{A}^{kl}(u) = \mu(|\mathcal{E}(u)|_2^2)(\delta_{kl}\delta_{ij} + \delta_{il}\delta_{jk}) + 4\mu'(|\mathcal{E}(u)|_2^2)\varepsilon_{ik}(u)\varepsilon_{jl}(u),$$

as introduced in (2.1). Since $\nabla u_* \in H_p^{1/2}(J_0; L_p(G;\mathbb{R}^n)) \cap L_p(J_0; H_p^1(G;\mathbb{R}^n))$ and this space embeds into $C(J_0 \times \overline{G})$, we see that the coefficients are continuous on $J_0 \times \overline{G}$ and strongly normally elliptic; see section 3. Moreover, $\nabla u_* \in Y_0$, and since $\mu \in C^2(\mathbb{R}_+)$ we also see that the coefficients of the first order boundary operators belong to $W_p^{1/2-1/2p}(J_0; L_p(\Gamma)) \cap L_p(J_0; W_p^{1-1/p}(\Gamma))$. Therefore Theorem 4.1 implies maximal regularity of the linear initial-boundary value problem defined by (4.1).

Next the function $\bar f$ is defined by

$$\bar f = -\Delta u_* - \sum_{k,l=1}^n \mathcal{A}^{kl} D_k D_l u_* - u_* \cdot \nabla u_*,$$

which belongs to $X_0 = L_p(J_0 \times G)$, and $F : {}_0Z_a \to X_a$ is given by

$$F(v) = [\mathcal{A}(u_*, D) - \mathcal{A}(u_* + v, D)](u_* + v) - [u_* \cdot \nabla v + v \cdot \nabla u_* + v \cdot \nabla v].$$

We decompose $F(v) = F_1(v) - F_2(v)$, indicated by the bracketing. The functions $\bar h_j$ read

$$\bar h_j = h_j + p_*\nu - 2\mu(|\mathcal{E}(u_*)|_2^2)\mathcal{E}(u_*)\nu, \quad j = s, n,$$

which belong to ${}_0Y_{j0}$ by the regularity of u_* and the choice of h_*. Finally, $H(v)$ is given by

$$H(v) = H_1(v) - H_2(v),$$

with

$$H_1(v) = 2[\mu(|\mathcal{E}(u_*)|_2^2) - \mu(|\mathcal{E}(u_* + v)|_2^2)]\mathcal{E}(v)\nu$$

and

$$H_2(v) = 2[\mu(|\mathcal{E}(u_* + v)|_2^2) - \mu(|\mathcal{E}(u_*)|_2^2) + 2\mu'(|\mathcal{E}(u_*)|_2^2)((\mathcal{E}(u_*)|\mathcal{E}(v)))]\mathcal{E}(u_*)\nu.$$

As a consequence, by maximal regularity we may rewrite (9.2) as the fixed point problem in ${}_0Z(a)$,

$$v = Tv := L(\bar f + F(v)) + L_s(\bar h_s + \mathcal{P}H(v)) + L_n(\bar h_n + H(v)), \quad v \in {}_0Z(a), \tag{9.3}$$

where $L \in \mathcal{B}(X(a); {}_0Z(a))$ and $L_j \in \mathcal{B}({}_0Y_j(a); {}_0Z(a))$ are causal bounded linear operators with bounds independent of $a \in (0, a_0]$. We denote this common bound by C_M in what follows.

To carry out the contraction argument, fix $r \in (0, 1]$ and consider the closed ball $\mathbb{B} := B_r(0) \subset {}_0Z(a)$. The numbers $a > 0$ and r will be chosen later small enough so that T becomes a contraction in ${}_0Z(a)$. We have $F_k(0) = H_k(0) = 0$ by definition. By uniform embedding, there is a constant $C_E > 0$ such that

$$|v|_\infty + |\nabla v|_\infty \le C_E |v|_{Z(a)}, \quad v \in {}_0Z(a).$$

Let $m := \sup\{|\mu(s)| + |\mu'(s)| + |\mu''(s)| : s \in [0, R]\}$. We first choose $a > 0$ so small that $|L\bar{f} + L_s\bar{h}_s + L_n\bar{h}_n|_{Z_a} \le r/4$.

For arbitrary $v, w \in \mathbb{B}$ we obtain

$$\begin{aligned} F_1(v) - F_1(w) = {} & [\mathcal{A}(u_*, D) - \mathcal{A}(u_* + v, D)](v - w) \\ & + [\mathcal{A}(u_* + v, D) - \mathcal{A}(u_* + w, D)](u_* + w), \end{aligned}$$

and hence

$$\begin{aligned} |F_1(v) - F_1(w)|_{X_a} \le {} & \sup_{J_a \times G} |\mathcal{A}(u_*, D) - \mathcal{A}(u_* + v, D)||v - w|_{Z(a)} \\ & + \sup_{J_a \times G} |\mathcal{A}(u_* + v, D) - \mathcal{A}(u_* + w, D)||u_* + w|_{Z(a)} \\ \le {} & C|v - w|_{Z(a)}[|v|_\infty + |\nabla v|_\infty] \\ & + C[|u_*|_{Z(a)} + r][|v - w|_\infty + |\nabla (v - w)|_\infty] \\ \le {} & C|v - w|_{Z(a)}[|u_*|_{Z(a)} + r] \le \frac{1}{8C_M}|v - w|_{Z(a)}, \end{aligned}$$

provided that $a > 0$ and $r \in (0, 1]$ are chosen small enough. Here the constants $C > 0$ differ from line to line but depend only linearly on C_E and m.

In a similar, even simpler, way we estimate F_2. Since

$$Z(a) \hookrightarrow L_p(J_a; H^1_{2p}(G; \mathbb{R}^n)),$$

with embedding constant uniform in a, we obtain

$$|F_2(v) - F_2(w)|_{X(a)} \le C(|u^*|_{Z(a)} + r)|v - w|_{Z(a)} \le \frac{1}{8C_M}|v - w|_{Z(a)},$$

provided that $a > 0$ and $r \in (0, 1]$ are small enough. This takes care of the nonlinear term containing F in the definition of T.

The estimates for H are more complicated since they involve fractional Sobolev spaces, but they are still elementary. While we do not claim that the arguments below are new, we include them for the sake of completeness. Recall that a norm for $W^s_p(\Sigma)$, Σ a compact C^1-manifold in $\mathbb{R}^n$, is given by

$$|a|_{W^s_p(\Sigma)} = |a|_p + \left[\int_\Sigma \int_\Sigma \frac{|a(x) - a(y)|^p}{|x - y|^{sp+n}} d\sigma(x) d\sigma(y)\right]^{1/p},$$

where $d\sigma$ denotes the surface measure on Σ.

There are two fundamental estimates for fractional Sobolev spaces that one should keep in mind. The first concerns products and reads as

$$|ab|_{W^s_p} \le |a|_\infty |b|_{W^s_p} + |b|_\infty |a|_{W^s_p},$$

valid for all functions $a, b \in W_p^s \cap L_\infty,\ s \in (0,1)$. The second concerns substitution operators in W_p^s of the form $\phi(a)$, where $\phi \in C^2$. Based on the identity

$$\begin{aligned}
&[\phi(a(x)) - \phi(b(x))] - [\phi(a(y)) - \phi(b(y))] \\
&= \int_0^1 \int_0^1 \partial_t \partial_s \phi(s[ta(x) + (1-t)b(x)] + (1-s)[ta(y) + (1-t)b(y)]) dsdt \\
&= \int_0^1 \int_0^1 \phi'(\xi(t,s))([a(x) - b(x)] - [a(y) - b(y)]) dsdt \\
&+ \int_0^1 \int_0^1 \phi''(\xi(t,s))([ta(x) + (1-t)b(x)] - [ta(y) + (1-t)b(y)]) \\
&\qquad\qquad \cdot (s[a(x) - b(x)] + (1-s)[a(y) - b(y)]) dtds,
\end{aligned}$$

we obtain

$$\begin{aligned}
&|[\phi(a(x)) - \phi(b(x))] - [\phi(a(y) - \phi(b(y)]| \le |\phi'|_\infty |(a(x) - b(x)) - (a(y) - b(y))| \\
&+ |\phi''|_\infty \{ |(a(x) - b(x)) - (a(y) - b(y))| \\
&\qquad \cdot (3|a - b|_\infty + |b(x) - b(y)|) + |a - b|_\infty |b(x) - b(y)| \}.
\end{aligned}$$

This implies

$$|\phi(a) - \phi(b)|_{W_p^s(\Sigma)} \le |\phi|_{BUC^2} [|a - b|_{W_p^s(\Sigma)} (1 + 3|a - b|_\infty + |b|_{W_p^s}) + |a - b|_\infty |b|_{W_p^s}].$$

Using these two inequalities, we obtain for H_1 the Lipschitz estimate

$$|H_1(v) - H_1(w)|_{Y_j(a)} \le Cr|v - w|_{Z(a)},$$

with some constant C which depends only on L and embedding constants that are uniform in a. In a similar way, skipping the details, we also get

$$|H_2(v) - H_2(w)|_{Y_j(a)} \le C(r + |u_*|_{Z(a)})|v - w|_{Z(a)}.$$

Thus, choosing $a > 0$ and $r \in (0,1]$ small enough, we obtain

$$|H(v) - H(w)|_{Y_j(a)} \le \frac{1}{8C_M} |v - w|_{Z(a)}.$$

Therefore $T : \mathbb{B} \to \mathbb{B}$ is a strict contraction and hence, by the contraction mapping principle, admits a unique fixed point in $Z(a)$.

Thus we have shown unique solvability of (9.1) on a probably small interval $J_a = [0, a]$ in the maximal regularity class. Now, we may repeat the above arguments to obtain successively solutions in the maximal regularity class on intervals $[t_i, t_{i+1}]$. Either after finitely many steps we reach a_0, or we have an infinite strictly increasing sequence which converges to some $t_* = t_*(u_0) \le a_0$. In case $\lim_{i\to\infty} u(t_i) =: u(t_*)$ exists in X_p, we may continue the process, which shows that the maximal time is characterized by the property that this limit does not exist in X_p. The continuous dependence of the solution on the data is obvious, and therefore the semiflow property claimed in Theorem 2.1 follows as well. This completes the proof of Theorem 2.1. ∎

10. Appendix: Normally elliptic boundary value problems. Above, we employed results on maximal regularity of completely inhomogeneous vector-valued parabolic initial value problems of the form

$$\begin{aligned}
\partial_t u + \mathcal{A}(t,x,D)u &= f(t,x), \quad t \in J,\ x \in G, \\
\mathcal{B}_j(t,x,D)u &= g_j(t,x), \quad t \in J,\ x \in \partial G,\ j = 1, \dots, m, \\
u(0,x) &= u_0(x), \quad x \in G.
\end{aligned} \tag{10.1}$$

Here $J = [0, T]$ for some $T > 0$, and $G \subset \mathbb{R}^n$ is an open connected set with compact boundary ∂G. The operator $\mathcal{A}(t, x, D)$ is a partial differential operator of order $2m$, and $\mathcal{B}_j(t, x, D)$ are partial boundary differential operators of order $m_j < 2m$. More precisely, let E be a Banach space and $m, m_1, \ldots, m_m$ be natural numbers with $m_j < 2m$ for $j = 1, \ldots, m$, and let

$$\mathcal{A}(t,x,D) = \sum_{|\alpha| \le 2m} a_\alpha(t,x) D^\alpha,$$
$$\mathcal{B}_j(t,x,D) = \sum_{|\beta| \le m_j} b_{j\beta}(t,x) D^\beta,$$

where a_α and $b_{j\beta}$ are $\mathcal{B}(E)$-valued variable coefficients. Here and in the following, we use the standard multi-index notation $D^\alpha = (-i)^{|\alpha|} \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n}$ with $\partial_{x_i} = \frac{\partial}{\partial x_i}$ and, later on, $\xi^\alpha = \xi_1^{\alpha_1} \cdots \xi_n^{\alpha_n}$. Note that the boundary operators have to be interpreted as $\mathcal{B}_j(t,x,D)u = \sum_{|\beta| \le m_j} b_{j\beta}(t,x) \gamma_0 D^\beta u$, where $\gamma_0 v$ denotes the trace of the function v on the boundary ∂G (given sufficient smoothness of v and ∂G); we will often omit γ_0 in the notation. In particular, the coefficients $b_{j\beta}$ are defined on ∂G only.

We are interested in maximal L_p-regularity of (10.1) for $1 < p < \infty$, which means that we are looking for solutions in the class

$$u \in H_p^1(J; L_p(G; E)) \cap L_p(J; H_p^{2m}(G; E)). \tag{10.2}$$

Trace theorems show that, for this regularity of the solution, the given data have to satisfy the following conditions:

(D) *Assumptions on the data.*

(i) $f \in L_p(J \times G; E)$,

(ii) $g_j \in W_p^{\kappa_j}(J; L_p(\partial G; E)) \cap L_p(J; W_p^{2m\kappa_j}(\partial G; E))$ with $\kappa_j := (2m - m_j - 1/p)/(2m)$,

(iii) $u_0 \in W_p^{2m(1-1/p)}(G; E)$,

(iv) if $\kappa_j > 1/p$, then $B_j(0,x)u_0(x) = g_j(0,x)$ for $x \in \partial G$.

The assumptions on the coefficients needed are of two types, namely smoothness of the coefficients and ellipticity. We start with ellipticity. For this, we denote the principal part of a partial differential operator $\mathcal{A}$ by $\mathcal{A}_\#$. The outer normal of ∂G at $x \in \partial G$ is as before $\nu(x)$, and we set $\mathbb{C}_+ := \{z \in \mathbb{C} : \Re z > 0\}$.

(E) *Ellipticity of the interior symbol.* For all $t \in J$, $x \in \overline{G}$ and $\xi \in \mathbb{R}^n$, $|\xi| = 1$ we have

$$\sigma(\mathcal{A}_\#(t,x,\xi)) \subset \mathbb{C}_+;$$

i.e., $\mathcal{A}(t,x,D)$ is *normally elliptic.*

(LS) *Lopatinskii–Shapiro condition.* For all $t \in J$, $x \in \partial G$, all $\xi \in \mathbb{R}^n$ with $(\xi|\nu(x)) = 0$, all $h \in E^m$, and all $\lambda \in \overline{\mathbb{C}}_+$ with $|\xi| + |\lambda| \neq 0$, the ordinary differential equation system in $\mathbb{R}_+$

$$\begin{aligned} \lambda v(y) + \mathcal{A}_\#(t,x,\xi - \nu(x)D_y)v(y) &= 0, \quad y > 0, \\ \mathcal{B}_{j,\#}(t,x,\xi - \nu(x)D_y)v(0) &= h_j, \quad j = 1,\ldots,m, \end{aligned} \tag{10.3}$$

admits a unique solution $v \in C_0(\mathbb{R}_+; E)$.

Now we turn to smoothness assumptions on the coefficients of $\mathcal{A}$ and $\mathcal{B}_j$.

(SD) There are $r_k, s_k \geq p$ with $\frac{1}{s_k} + \frac{n}{2mr_k} < 1 - \frac{k}{2m}$ such that

$$a_\alpha \in L_{s_k}(J; (L_{r_k} + L_\infty)(G; \mathcal{B}(E))), \quad |\alpha| = k < 2m,$$
$$a_\alpha \in C(J \times \overline{G}; \mathcal{B}(E)), \quad |\alpha| = 2m.$$

If G is unbounded, the limits $a_\alpha(t, \infty) := \lim_{|x| \to \infty, x \in G} a_\alpha(t, x)$ exist uniformly in $t \in J$, $|\alpha| = 2m$.

(SB) There are $s_{jk}, r_{jk} \geq p$ with $\frac{1}{s_{jk}} + \frac{n-1}{2mr_{jk}} < \kappa_j + \frac{m_j - k}{2m}$ such that

$$b_{j\beta} \in W^{\kappa_j}_{s_{jk}}(J; L_{r_{jk}}(\partial G; \mathcal{B}(E))) \cap L_{s_{jk}}(J; W^{2m\kappa_j}_{r_{jk}}(\partial G; E)), \qquad |\beta| = k \leq m_j.$$

The main result of this appendix states that, under the assumptions made so far, the initial boundary value problem (10.1) admits maximal regularity. This result is taken from Denk, Hieber, and Prüss [5].

THEOREM 10.1. *Let $G \subset \mathbb{R}^n$ be open and connected with compact boundary ∂G of class C^{2m}. Let the Banach space E be of class $\mathcal{HT}$, suppose assumptions* (E), (LS), (SD), *and* (SB) *are satisfied, and let $1 < p < \infty$. Then problem* (10.1) *admits a unique solution*

$$u \in H^1_p(J; L_p(G; E)) \cap L_p(J; H^{2m}_p(G; E))$$

if and only if the data are subject to conditions (D).

It should be observed that, assuming the regularity assumptions (SD) and (SB), normal ellipticity (E) and the Lopatinskii–Shapiro condition (LS) are *necessary* for this result. These facts are also proved in Denk, Hieber, and Prüss [5].

Also important, as used in this paper in the slip case, it is possible to extend this result to situations where the boundary conditions do not have fixed order. So, for example, we may split boundary condition j as follows. Suppose $E = E_1 \oplus E_2$, and let

$$\mathcal{B}_j(t, x, D)u = (\mathcal{B}_{j1}(t, x, D)u, \mathcal{B}_{j2}(t, x, D)u),$$

where the coefficients are in $\mathcal{B}(E, E_k)$, $k = 1, 2$. Then the same results are valid in case the orders m_{jk} of $\mathcal{B}_{jk}$ differ; of course, the boundary spaces have to be adjusted accordingly.

REFERENCES

[1] H. AMANN, *Stability of the rest state of a viscous incompressible fluid*, Arch. Ration. Mech. Anal., 126 (1994), pp. 231–242.
[2] R. B. BIRD, R. C. ARMSTRONG, AND O. HASSAGER, *Dynamics of Polymeric Liquids, Vol.* 1*, Fluid Mechanics*, John Wiley & Sons, New York, 1977.
[3] PH. CLÉMENT AND S. LI, *Abstract parabolic quasilinear equations and applications to a groundwater flow problem*, Adv. Math. Sci. Appl., 3 (1994), pp. 17–32.
[4] R. DENK, M. HIEBER, AND J. PRÜSS, *$\mathcal{R}$-boundedness, Fourier multipliers and problems of elliptic and parabolic type*, Mem. Amer. Math. Soc. 166, 788 (2003).
[5] R. DENK, M. HIEBER, AND J. PRÜSS, *Optimal L_p-L_q-regularity for parabolic problems with inhomogeneous boundary data*, Math. Z., to appear.
[6] L. DIENING AND M. RŮŽIČKA, *Strong solutions for generalized Newtonian fluids*, J. Math. Fluid Mech., 7 (2005), pp. 413–450.
[7] J. ESCHER, *Analyticity of Viscous Incompressible Stokesian Fluids in Layers*, preprint.
[8] J. FREHSE, J. MÁLEK, AND M. STEINHAUER, *On existence results for fluids with shear-dependent viscosity—Unsteady flows*, in Partial Differential Equations: Theory and Numerical Solution, W. Jäger, J. Nečas, O. John, and J. Stará, eds., Res. Notes Math. 406, Chapman & Hall/CRC Press, Boca Raton, FL, 2000, pp. 121–129.

[9] J. Frehse, J. Málek, and M. Steinhauer, *On analysis of steady flows of fluids with shear-dependent viscosity based on the Lipschitz truncation method*, SIAM J. Math. Anal., 34 (2003), pp. 1064–1083.

[10] M. Girardi and L. Weis, *Criteria for $\mathcal{R}$-boundedness of families of operators*, in Evolution Equations: Proceedings in Honor of J. A. Goldstein's 60th Birthday, G. R. Goldstein, R. Nagel, and S. Romanelli, eds., Marcel Dekker, New York, 2003, pp. 203–221.

[11] N. J. Kalton and L. Weis, *The H^∞-calculus and sums of closed operators*, Math. Ann., 321 (2001), pp. 319–345.

[12] P. Kaplický, J. Málek, and J. Stará, *Full regularity of weak solutions to a class of nonlinear fluids in two dimensions—Stationary, periodic problem*, Comment. Math. Univ. Carolinae, 38 (1997), pp. 681–695.

[13] P. Kaplický, J. Málek, and J. Stará, *Global-in-time Hölder continuity of the velocity gradients for fluids with shear-dependent viscosities*, Nonlinear Differential Equations Appl., 9 (2002), pp. 175–195.

[14] O. A. Ladyzhenskaya, *Modification of the Navier–Stokes equations for large velocity gradients*, in Seminars in Mathematics 7, V. A. Steklov Math. Inst., Leningrad, 1968, pp. 57–69.

[15] E. Lauga, M. P. Brenner, and H. A. Stone, *Microfluids*: *The no-slip boundary condition*, in Handbook of Experimental Fluid Dynamics, J. Foss, C. Tropea, and Y. Yarin, eds., Springer, New York, 2007, Chapter 15, to appear.

[16] J. Málek, J. Nečas, and M. Růžička, *On the non-Newtonian incompressible fluids*, Math. Models Methods Appl. Sci., 3 (1993), pp. 35–63.

[17] J. Málek and K. R. Rajagopal, *Mathematical issues concerning the Navier-Stokes equations and some of their generalizations*, in Handbook of Differential Equations, Evolutionary Equations, Vol. 2, C. Dafermos and E. Feireisl, eds., Elsevier, New York, 2005, pp. 371–459.

[18] J. Málek, K. R. Rajagopal, and M. Růžička, *Existence and regularity of solutions and the stability of the rest state for fluids with shear dependent viscosity*, Math. Models Methods Appl. Sci., 6 (1995), pp. 789–812.

[19] J. Málek, J. Nečas, M. Rokyta, and M. Růžička, *Weak and Measure-Valued Solutions to Evolutionary PDEs*, Appl. Math. Math. Comput. 13, Chapman & Hall, London, 1996.

[20] J. Málek, J. Nečas, and M. Růžička, *On weak solutions to a class of non-Newtonian incompressible fluids in bounded domains*: *The case $p \geq 2$*, Adv. Differential Equations, 6 (2001), pp. 257–302.

[21] J. Prüss, *Maximal regularity for evolution equations in L_p-spaces*, Conf. Semin. Mat. Univ. Bari, 285 (2003), pp. 1–39.

[22] C. G. Simader and H. Sohr, *The Dirichlet Problem for the Laplacian in Bounded and Unbounded Domains*, Longman, Edinburgh, 1996.

[23] V. A. Solonnikov, *L_p-estimates for solutions to the initial-boundary-value problem for the generalized Stokes system in a bounded domain*, J. Math. Sci., 105 (2001), pp. 2448–2484.

[24] R. I. Tanner and K. Walters, *Rheology*: *An Historical Perspective*, Elsevier, Amsterdam, 1998.

[25] K. Wielage, *Analysis of Non-Newtonian Two-Phase Flows*, Ph.D. Dissertation, Department of Mathematics, University of Paderborn, Germany, 2005.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 422–436

VORTEX DYNAMICS IN A TWO-DIMENSIONAL DOMAIN WITH HOLES AND THE SMALL OBSTACLE LIMIT*

M. C. LOPES FILHO†

Abstract. In this work we examine the asymptotic behavior of solutions of the incompressible two-dimensional Euler equations on a domain with several holes, when one of the holes becomes small. We show that the limit flow satisfies a modified Euler system in the domain with the small hole removed. In vorticity form, the limit system is the usual equation for transport of vorticity, coupled with a modified Biot–Savart law which includes a point vortex at the point where the small hole disappears, together with the appropriate correction for the harmonic part of the flow. This work extends results by Iftimie, Lopes Filho, and Nussenzveig Lopes, obtained in the context of the exterior of a single small obstacle in the plane; see [*Comm. Partial Differential Equations*, 28 (2003), pp. 349–379]. The main difficulty in the present situation lies in controlling the behavior of the harmonic part of the flow, which is not an exact conserved quantity. As part of our analysis we develop a new description of two-dimensional vortex dynamics in a general domain with holes.

Key words. incompressible flow, Euler equations, vorticity

AMS subject classifications. Primary, 76B03; Secondary, 35Q35, 76B47

DOI. 10.1137/050647967

1. Introduction.

In this article we consider incompressible two-dimensional (2D) ideal flow in a bounded smooth domain with k holes and study its asymptotic behavior when one of the holes becomes small. This research is a natural outgrowth of the work done in [3], where the authors studied the asymptotic behavior for ideal 2D flow in the exterior of a single small obstacle in the plane.

Although treating a similar problem, the present work requires an entirely new approach. In [3], a family of conformal transformations between the exterior of the small obstacle and the exterior of the unit disk played a crucial role, allowing the use of explicit formulas for basic harmonic fields and the Biot–Savart law. This approach makes it convenient to assume that the obstacle is vanishing in a self-similar fashion. For the present work, it is not possible to work with explicit formulas, and the conformal mapping technique has to be substituted by qualitative analysis using elliptic techniques, including variational methods and the maximum principle. One consequence of this new approach is that the self-similarity hypothesis, used in [3], could be removed altogether.

One important distinction between 2D ideal flow in the exterior of a single obstacle and in a bounded domain with several holes is the behavior of the harmonic part of the flow. In the single obstacle case, the harmonic part is constant in time and determined by the initial data, so no topological difficulties in reducing 2D ideal flow to vortex dynamics exist. In contrast, there was no reduction available of flow in domains with multiple holes to vortex dynamics (see, for example, [5] for an ad hoc formulation of vortex dynamics in the exterior of two identical disks). Our point of departure is a new formulation of 2D ideal flow using vorticity alone as the dynamical variable.

*Received by the editors December 19, 2005; accepted for publication (in revised form) December 4, 2006, published electronically June 15, 2007. Research supported in part by CNPq grant 300.962/91-6.
http://www.siam.org/journals/sima/39-2/64796.html

†Department of Mathematics, IMECC-UNICAMP, Campinas, SP 13083-250, Brazil (mlopes@ime.unicamp.br).

We refer to the system obtained by rewriting the 2D Euler equations using vorticity as the unknown as the vortex dynamics formulation of the 2D Euler equations. The main ingredient in this new formulation is an expression for the harmonic part of the flow in terms of vorticity.

The remainder of this work is divided into five sections. In section 2 we derive an expression for the harmonic part of 2D ideal flow on a bounded domain with holes in terms of vorticity, and deduce a vortex dynamics formulation for the incompressible 2D Euler equations. In section 3 we formulate the small obstacle problem and study limits for quantities that depend on the domain alone. In section 4 we introduce the dynamical problem and obtain uniform estimates for the solution of the 2D Euler system on the domain with one small hole. In section 5 we put together the estimates in sections 3 and 4 to obtain a limit equation. We draw conclusions and discuss some extensions and open problems in section 6.

2. Vortex dynamics on a domain with holes. Let Ω be a bounded connected domain in the plane whose boundary $\partial\Omega$ is the disjoint union of a finite number of smooth Jordan curves Γ_i, $i = 0, 1, \ldots, k$. We assume that Ω is the bounded region with boundary Γ_0, with the regions bounded by Γ_i, $i = 1, \ldots, k$, removed. We require basic terminology and results from DeRham cohomology and Hodge theory, in the particularly easy setting of bounded domains with holes in the plane. We refer the reader to [11] for this material.

A vector field in Ω is called *harmonic* if it is both divergence- and curl-free. We identify one-forms with vector fields by $adx_1 + bdx_2 \mapsto (a, b)$. The one-dimensional (1D) deRham cohomology of Ω is the k-dimensional real vector space $H^1_{DR}(\Omega)$. It is a consequence of Hodge's theorem that the vector space $H^1_{DR}(\Omega)$ has a basis consisting of harmonic vector fields. There is a unique family of harmonic vector fields $\{X_1, X_2, \ldots, X_k\}$ satisfying the period conditions

$$\oint_{\Gamma_j} X_i \cdot ds = \delta_{ij}, \qquad \text{with } i = 1, \ldots, k \text{ and } j = 0, 1, \ldots, k, \tag{1}$$

which are a basis for $H^1_{DR}(\Omega)$.

Let $V(\Omega)$ denote the vector space of smooth divergence-free vector fields tangent to $\partial\Omega$, and $H(\Omega)$ be the subspace of $V(\Omega)$ consisting of irrotational vector fields. For each $i = 0, 1, \ldots, k$, we introduce the function ϕ_i as the unique solution to the Dirichlet problem

$$\begin{cases} \Delta\phi_i = 0, \\ \phi_i = 0 \quad \text{on } \Gamma_j, j \neq i, \\ \phi_i = 1 \quad \text{on } \Gamma_i. \end{cases} \tag{2}$$

Given $\omega \in C^\infty(\overline{\Omega})$, let ψ be the solution of the Poisson problem

$$\begin{cases} \Delta\psi = \omega, \\ \psi = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{3}$$

Denote $(a, b)^\perp = (-b, a)$. We define the Biot–Savart operator

$$K_\Omega[\omega] \equiv \nabla^\perp\psi.$$

Clearly, $K_\Omega[\omega] \in V(\Omega)$ for any $\omega \in C^\infty(\overline{\Omega})$. The Biot–Savart operator is an integral operator with kernel given by $\nabla_x^\perp G_\Omega(x, y)$, where G_Ω is the Green function of the Dirichlet Laplacian in Ω.

Fix $u \in V(\Omega)$ and let $\omega \equiv$ curl u. The Hodge–Kodaïra decomposition theorem (see [2, section 7]) implies that u can be decomposed in a unique manner in the form $u = \nabla^\perp \psi + H$, with ψ smooth, vanishing on $\partial\Omega$ and $H \in H(\Omega)$. This decomposition is orthogonal with respect to the L^2 inner product.

Taking the curl of this decomposition, using the uniqueness of solutions to (3), and expanding H, we get

$$u = K_\Omega[\omega] + \sum_{j=1}^{k} \alpha_j X^j, \tag{4}$$

for some $\alpha_j \in \mathbb{R}$.

Our first result is an identity that expresses the components of the irrotational part of u in a useful manner.

PROPOSITION 1. *For each $i = 1, \dots, k$ we have*

$$\alpha_i = \int_\Omega \phi_i \omega dx + \oint_{\Gamma_i} u \cdot ds.$$

Proof. First we integrate by parts to get

$$\mathcal{I}_i \equiv \int_\Omega \phi_i \omega dx = \int_\Omega \phi_i \text{ curl } u dx = -\int_\Omega \phi_i \text{ div } u^\perp dx$$

$$= \int_\Omega \nabla\phi_i \cdot u^\perp dx - \int_{\partial\Omega} \phi_i u^\perp \cdot \widehat{n} ds = -\int_\Omega \nabla^\perp \phi_i \cdot u dx - \int_{\Gamma_i} u^\perp \cdot \widehat{n} ds.$$

We note that $\widehat{n}$ is the outward normal of Ω at Γ_i, and hence

$$\int_{\Gamma_i} u^\perp \cdot \widehat{n} ds = -\int_{\Gamma_i} u \cdot \widehat{n}^\perp ds = \int_{\Gamma_i} u \cdot \widehat{\tau} ds,$$

where $\widehat{\tau}$ is the unit counterclockwise tangent vector to Γ_i. Therefore,

$$\int_{\Gamma_i} u^\perp \cdot \widehat{n} ds = \oint_{\Gamma_i} u \cdot ds. \tag{5}$$

Next we use (4) to obtain

$$\int_\Omega \nabla^\perp \phi_i \cdot u dx = \int_\Omega \nabla^\perp \phi_i \left(K_\Omega[\omega] + \sum_{j=1}^{k} \alpha_i X^i \right) dx$$

$$= \int_\Omega \nabla \phi_i \cdot \nabla \psi dx + \sum_{j=1}^{k} \alpha_j \int_\Omega \nabla^\perp \phi_i \cdot X^j dx$$

$$= -\int_\Omega \psi \Delta \phi_i dx + \int_{\partial\Omega} \psi \nabla \phi_i \cdot \widehat{n} ds + \sum_{j=1}^{k} \alpha_j \int_\Omega \nabla^\perp \phi_i \cdot X^j dx.$$

The two first terms above vanish because $\Delta\phi_i = 0$ in Ω and $\psi = 0$ on $\partial\Omega$. Therefore,

$$\int_\Omega \nabla^\perp \phi_i \cdot u dx = \sum_{j=1}^k \alpha_j \int_\Omega \nabla^\perp \phi_i \cdot X^j dx. \tag{6}$$

Putting together the calculation done on I_i, (5), and (6), we get

$$I_i = -\sum_{j=1}^k \alpha_j \int_\Omega \nabla^\perp \phi_i \cdot X^j dx - \oint_{\Gamma_i} u \cdot ds. \tag{7}$$

Next, we use (4) on (7) to isolate and collect terms with α_i to obtain

$$\int_\Omega \phi_i \omega dx + \oint_{\Gamma_i} K_\Omega[\omega] \cdot ds = \sum_{j=1}^k \alpha_j \left\{ -\int_\Omega \nabla^\perp \phi_i \cdot X^j dx - \oint_{\Gamma_i} X^j \right\}. \tag{8}$$

We note that ω and each α_j can be chosen independently when one chooses u. The left-hand side of (8) and each of the expressions in brackets on the right-hand side of (8) do not depend on the α_j's, and therefore identity (8) implies both that

$$\int_\Omega \phi_i \omega dx + \oint_{\Gamma_i} K_\Omega[\omega] \cdot ds = 0, \tag{9}$$

and that for each $j = 1, \ldots, k$,

$$\int_\Omega \nabla^\perp \phi_i \cdot X^j dx = -\oint_{\Gamma_i} X^j = -\delta_{ij}, \tag{10}$$

by (1). Using (10) back on (7), we get

$$I_i = \alpha_i - \oint_{\Gamma_i} u \cdot ds,$$

which concludes the proof. $\square$

The initial-boundary value problem for the incompressible 2D Euler equations in Ω and in the time interval $[0, T)$ has the form

$$\begin{cases} u_t + u \cdot \nabla u = -\nabla p & \text{in } \Omega \times (0, T), \\ \text{div } u = 0 & \text{in } \Omega \times \{t\}, \text{ for } 0 \le t < T, \\ u \cdot \widehat{n} = 0 & \text{on } \partial\Omega \times [0, T), \\ u(x, 0) = u_0(x) & \text{in } \Omega \times \{t = 0\}, \end{cases} \tag{11}$$

where the velocity u and the scalar pressure p are unknowns, u_0 is the initial data, and $\widehat{n}$ is the unit exterior normal to $\partial\Omega$. Well-posedness for this problem when u_0 is sufficiently smooth is due to Kato (see [6]), and existence of weak solutions for $u_0 \in L^2$, with $\omega_0 \equiv$ curl $u_0 \in L^1 + \mathcal{BM}^+$, is part of Delort's existence result; see [1, 10]. Vortex dynamics is an approach to the problem above where the dynamic variables u and p are replaced by the vorticity $\omega =$ curl u. Formulating problem (11) in terms of vorticity alone becomes complicated by the topology of the domain and can be accomplished only by using Proposition 1 as follows.

Let $u = u(x,t)$ be a smooth solution of (11) with $\omega = \omega(x,t) \equiv \operatorname{curl} u(x,t)$. We can write

$$u = K_\Omega[\omega] + \sum_{j=1}^{k} \alpha_j(t) X_j,$$

and Proposition 1 gives that for each $j = 1, \dots, k$

$$\alpha_j(t) = \int_\Omega \phi_j(x)\omega(x,t)dx + \oint_{\Gamma_j} u(x,t) \cdot ds.$$

By Kelvin's circulation theorem, the circulations $\oint_{\Gamma_j} u(x,t) \cdot ds$ are conserved in time, so that, if $\omega_0 = \operatorname{curl} u_0$, we have

$$\oint_{\Gamma_j} u(x,t) \cdot ds = \oint_{\Gamma_j} u(x,0) \cdot ds = \alpha_j(0) - \int_\Omega \phi_j \omega_0 dx.$$

Therefore, taking the curl of the momentum equations in (11), we can write the following initial value problem, with only ω as unknown, equivalent to (11) for classical solutions:

$$\text{(12)} \qquad \begin{cases} \omega_t + u \cdot \nabla\omega = 0 \quad \text{in } \Omega \times (0,T), \\ u = K_\Omega[\omega] + \sum_{j=1}^{k} \alpha_j(t) X_j \quad \text{in } \Omega \times \{t\}, \text{ for } 0 \le t < T, \\ \alpha_j(t) = \int_\Omega \phi_j \omega dx + \alpha_{j,0} - \int_\Omega \phi_j \omega_0 dx \quad \text{in } [0,T), \\ \omega(x,0) = \omega_0(x) \quad \text{and} \quad \alpha_j(0) = \alpha_{j,0} \quad \text{in } \Omega \times \{t = 0\}. \end{cases}$$

The initial data for problem (12) is the initial vorticity ω_0 and the initial coefficients for the harmonic part of the velocity, $\alpha_{j,0}$. Clearly, assigning this data is equivalent to choosing an initial velocity u_0. System (12) is the vortex dynamics formulation of (11).

3. Harmonic fields. Next we introduce the small obstacle problem. Let us consider a family of bounded connected domains Ω_δ, $0 \le \delta \le \delta_0$, which, for each fixed δ, has its boundary made up of curves Γ_i, $i = 0, 1, \dots, k$, arranged as in section 2. We assume that for each $i = 0, 1, \dots, k-1$, Γ_i is independent of δ, and that $\Gamma_k = \Gamma_k^\delta$ is contained on a ball of radius δ centered at some fixed point $P \in \Omega_\delta$. We declare Ω to be the original set Ω_{δ_0} union with the closure of the bounded side of $\Gamma_k^{\delta_0}$.

Fix $\omega_0 \in C_c^\infty(\Omega_{\delta_0})$, and assume that the support of ω_0 does not intercept $\{|x| \le \delta_0\}$. This implies that $\omega_0 \in C_c^\infty(\Omega_\delta)$ for any $0 \le \delta < \delta_0$. Let $X_\delta^1, \dots, X_\delta^k$ be the basis of $H(\Omega_\delta)$ satisfying condition (1). Define $\phi_{i,\delta}$ solutions of (2) in Ω_δ. Fix $\alpha_{1,0}, \dots, \alpha_{k,0}$ real numbers and define

$$u_{\delta,0} \equiv K_{\Omega_\delta}[\omega_0] + \sum_{j=1}^{k} \alpha_{j,0} X_\delta^j.$$

By Proposition 1 we have that

$$\alpha_{i,0} = \int_{\Omega_\delta} \omega_0 \phi_{i,\delta} + \oint_{\Gamma_i} u_{\delta,0} \cdot ds.$$

We begin our study of the small obstacle limit by exploring the behavior of $\phi_{i,\delta}$ as $\delta \to 0$. This is an exercise in elementary PDE, which we summarize in the following result.

LEMMA 1. *Let ϕ_i, $i = 1, \ldots, k-1$, be the solutions of* (2) *in the unperturbed domain Ω. For any $i = 1, \ldots, k$, $\phi_{i,\delta}$ is uniformly bounded. For $i = 1, \ldots, k-1$, $\phi_{i,\delta}$ converges uniformly to ϕ_i as $\delta \to 0$, uniformly outside any neighborhood of P, and $\phi_{k,\delta}$ converges uniformly to* 0 *outside any neighborhood of P when $\delta \to 0$. Furthermore, there exists a constant $C > 0$ such that $\|\nabla \phi_{i,\delta}\|_{L^2(\Omega_\delta)} \leq C$.*

Proof. First note that, by the maximum principle, $0 \leq \phi_{i,\delta}(x) \leq 1$ for all $x \in \Omega_\delta$, which proves uniform boundedness.

Next, we consider the behavior of $\phi_{k,\delta}$. Let $R > 0$ be such that $\Omega_{\delta_0} \subseteq \{|x - P| \leq R\}$. The function

$$\Psi_\delta(x) \equiv \frac{\log(R/|x-P|)}{\log(R/\delta)}$$

is the unique solution of

$$\begin{cases} \Delta \Psi_\delta = 0 & \text{in } \delta < |x-P| < R, \\ \Psi_\delta = 1 & \text{on } |x-P| = \delta, \\ \Psi_\delta = 0 & \text{on } |x-P| = R. \end{cases}$$

We consider the domain $\widetilde{\Omega}_\delta$ consisting of the domain Ω_δ with the disk $|x - P| < \delta$ removed. It is easy to see that $\phi_{k,\delta} \leq \Psi_\delta$ on $\partial\widetilde{\Omega}_\delta$, as this boundary consists of the union of the curves Γ_i, $i = 0, \ldots, k-1$ (where $\phi_{k,\delta}$ vanishes and $\Psi_\delta > 0$) with $\{|x - P| = \delta\}$, where $\Psi_\delta = 1$ and $\phi_{k,\delta} \leq 1$. Therefore, by the maximum principle, $0 \leq \phi_{k,\delta} \leq \Psi_\delta$ in $\widetilde{\Omega}_\delta$. Next we observe that if $|\log \delta|^{-1} \leq |x - P| \leq R$,

$$\Psi_\delta(x) \leq \frac{\log R + \log|\log \delta|}{\log R - \log \delta} \to 0 \quad \text{as } \delta \to 0,$$

which proves our assertion regarding $\phi_{k,\delta}$.

Next fix $i < k$ and consider $\Phi_\delta \equiv \phi_i - \phi_{i,\delta}$. The function Φ_δ is harmonic in Ω_δ and vanishes on all boundary components except Γ_k^δ, where it is bounded between 0 and 1. Therefore, the same argument used above for $\phi_{k,\delta}$ applies to Φ_δ as well, which proves the statement regarding $\phi_{k,\delta}$.

Finally, to show the L^2 estimate on derivatives, we begin with the case $j \neq k$. We claim that $\|\nabla \phi_{j,\delta}\|_{L^2(\Omega_\delta)}$ is monotonically decreasing as $\delta \to 0$. Indeed, let $0 < \delta_2 < \delta_1$ and consider $E\phi_{j,\delta_1}$ the extension of ϕ_{j,δ_1} to Ω_{δ_2} obtained by setting $E\phi_{j,\delta_1}(x) = 0$ for $x \in \Omega_{\delta_2} \setminus \Omega_{\delta_1}$. The function $E\phi_{j,\delta_1}$ is defined in Ω_{δ_2} and satisfies the same Dirichlet data as ϕ_{j,δ_2}. However, ϕ_{j,δ_2} is harmonic and hence minimizes energy among functions with the same boundary data. Therefore,

$$\|\phi_{j,\delta_2}\|_{L^2(\Omega_{\delta_2})} \leq \|E\phi_{j,\delta_1}\|_{L^2(\Omega_{\delta_2})} = \|\phi_{j,\delta_1}\|_{L^2(\Omega_{\delta_1})}.$$

This estimate implies the L^2 bound for $j = 1, \ldots, k-1$. Next, let $\phi_{0,\delta}$ be defined by 2 in Ω_δ. The variational argument above applies to this function as well, so that $\|\nabla \phi_{0,\delta}\|_{L^2(\Omega_\delta)}$ is also monotonically increasing in δ and therefore bounded as $\delta \to 0$. We observe that $\sum_{i=0}^k \phi_{i,\delta}$ is a harmonic function in Ω_δ which is identically equal to 1 on $\partial\Omega_\delta$. Therefore, $\sum_{i=0}^k \phi_{i,\delta} \equiv 1$ and

$$\nabla \phi_{k,\delta} = -\sum_{i=0}^{k-1} \nabla \phi_{i,\delta},$$

and since all summands are uniformly bounded in L^2 by the argument above, this identity concludes the proof. ◻

Remark. Note that the variational argument used in the proof of the L^2 estimate on derivatives given above assumed implicitly that $\Omega_{\delta_1} \subset \Omega_{\delta_2}$ if $\delta_2 < \delta_1$. It is easy to modify the argument to remove this implicit assumption, by comparing the L^2 norm of $\nabla\phi_{i,\delta}$ with the L^2 norm of $E\nabla\phi_{\delta_0}$, where ϕ_{δ_0} is harmonic in $\Omega \setminus \{|x-P| < \delta_0\}$, with $\phi_{\delta_0} = \phi_{i,\delta}$ on $\partial\Omega$ and $\phi_{\delta_0} = 0$ on $|x-P| = \delta_0$. If the obstacles are not nested, the quantity $\|\nabla\phi_{\delta,j}\|_{L^2}$ may not be monotonic in δ, but the L^2 bound is still valid.

Next we study the asymptotic behavior of X^j_δ and $\delta \to 0$. For each $j = 1, \ldots, k$, we denote by EX^j_δ the extension of the vector field X^j_δ from Ω_δ to Ω, obtained by setting it to vanish inside Γ^k_δ. Regarding EX^j_δ we have the following result.

LEMMA 2. *For each $j = 1, \ldots, k$ and $\delta > 0$, the vector field EX^j_δ is divergence-free, in the sense of distributions, and its curl $\mu^j_\delta \equiv$ curl EX^j_δ is a measure supported on Γ^k_δ with density given by the tangential component of X^j_δ. Moreover, $\mu^j_\delta \rightharpoonup 0$ for $j = 1, \ldots, k-1$ and $\mu^k_\delta \rightharpoonup \delta(x-P)$ weak-$*$ bounded measures when $\delta \to 0$.*

Proof. First we observe that EX^j_δ is divergence-free, in the sense of distributions. The vector field EX^j_δ is divergence-free both in Ω_δ, by construction, and in $\Omega \setminus \overline{\Omega_\delta}$ because EX^j_δ vanishes there. To show that a vector field is divergence-free as a distribution it suffices to show that its integral against the gradient of an arbitrary test function vanishes. Indeed, given that $\phi \in C^\infty_c(\Omega)$ is a test function, we can compute

$$\int_\Omega \nabla\phi \cdot EX^j_\delta dx = \int_{\Omega_\delta} \nabla\phi \cdot X^j_\delta dx = \int_{\Gamma^k_\delta} \phi(x) X^j_\delta(x) \cdot \widehat{n} dS(x) = 0,$$

since X^j_δ is tangent to Γ^k_δ.

A similar calculation, integrating EX^j_δ against $\nabla^\perp\phi$ and using the divergence theorem proves that the curl of EX^j_δ is a Dirac with smooth density supported on the curve Γ^k_δ. More precisely, let $\mu^j_\delta =$ curl EX^j_δ as a distribution. If $z_\delta = z_\delta(s)$ is a parametrization by arc length of Γ^k_δ, with $0 \le s \le L_\delta$ and $\phi \in C^\infty_c(\Omega)$, we have

$$\langle \mu_\delta, \phi \rangle = \int_0^{L_\delta} \phi(z_\delta(s))(X^j_\delta \cdot \widehat{\tau})(z_\delta(s) ds,$$

where $\widehat{\tau}$ is the counterclockwise unit tangent vector to Γ^k_δ. In other words, μ^j_δ is a Dirac supported on the curve Γ^k_δ with smooth density $X^j_\delta \cdot \widehat{\tau}$.

Next we use an adaptation of the variational argument used in Proposition 1 to show that $\|X^j_\delta\|_{L^2(\Omega_\delta)}$ is nondecreasing in δ and therefore bounded when $\delta \to 0$ for all $j = 1, \ldots, k-1$. Take $0 < \delta_2 < \delta_1$, and observe that $EX^j_{\delta_1}$ has the same circulation as $X^j_{\delta_2}$ in all boundary components. By the orthogonality of the Hodge decomposition, the harmonic vector field minimizes energy in its cohomology class, which implies that

$$\|X^j_{\delta_2}\|_{L^2(\Omega_{\delta_2})} \le \|EX^j_{\delta_1}\|_{L^2(\Omega_{\delta_2})} \|X^j_{\delta_1}\|_{L^2(\Omega_{\delta_1})}.$$

As observed in the remark following the proof of Lemma 1, the argument above makes the implicit assumption that $\Omega_{\delta_1} \subset \Omega_{\delta_2}$, but this assumption may easily be removed.

Repeating the argument done for EX^j_δ in the beginning of the proof, we see that $E\nabla^\perp\phi_{i,\delta}$ is divergence-free in Ω, and we define curl $E\nabla^\perp\phi_{i,\delta} = \nu^i_\delta$, which are also smooth measures supported on Γ_δ^k. The density of these measures is given by $\nabla^\perp \phi_{i,\delta} \cdot \widehat{\tau} = -\nabla \phi_{i,\delta} \cdot \widehat{n}$, in contrast with EX_δ^j. However, there is one important difference between the μ and the ν measures. We have that $0 \leq \phi_{i,\delta} \leq 1$ by the maximum principle, and therefore $\nu_\delta^i \leq 0$ for $i = 1, \ldots, k-1$ and $\nu_\delta^k \geq 0$. The distinguished sign obtained above, together with the fact that $\nabla^\perp \phi_{i,\delta}$ is bounded in L^2 and the fact that the support of ν_δ^i is contained in a ball of radius δ implies that ν_δ^i converges strongly as a bounded measure to zero for all $i = 1, \ldots, k$; see [1].

For each $i, j = 1, \ldots, k$, let a_δ^{ij} be defined by

$$EX_\delta^j = \sum_{i=1}^{k} a_\delta^{ij} E \nabla^\perp \phi_{i,\delta}. \tag{13}$$

Taking the scalar product with EX_δ^l, integrating in Ω, and using (10), it follows that

$$a_\delta^{ij} = -\int_{\Omega_\delta} X_\delta^i \cdot X_\delta^j dx.$$

By the L^2 estimate of X_δ^j for $j = 1, \ldots, k-1$ and using the Cauchy–Schwarz inequality, we conclude that $|a_\delta^{ij}|$ is bounded if $1 \leq i, j \leq k-1$. Fix $j = 1, \ldots, k-1$ and take the curl of (13) to obtain

$$\mu_\delta^j = \sum_{i=1}^{k-1} a_\delta^{ij} \nu_\delta^i + a_\delta^{kj} \nu_\delta^k.$$

This identity implies that μ_δ^j is the sum of a measure that converges strongly to zero, given by the summation above, with a distinguished sign measure, which may be positive or negative depending on the sign of a_δ^{kj}. This, together with the fact that $\mu_\delta^j(\Omega) = 0$, implies both that $|\mu_\delta^j| \to 0$ and that $|a_\delta^{kj}| \nu_\delta^k \to 0$. We use the fact that $\nu_\delta^k = -\sum_{i=0}^{k-1} \nu_\delta^i$ to conclude that $a_\delta^{kj} \nu_\delta^j = a_\delta^{jk} \nu_\delta^j$ also converges to zero in bounded measures for all $j = 1, \ldots, k-1$. Finally, we write

$$\mu_\delta^k = \sum_{j=1}^{k-1} a_\delta^{jk} \nu_\delta^j + a^k k_\delta \nu_\delta^k.$$

As before, the summation converges to zero in bounded measures, so μ_δ^k is asymptotically distinguished signed and therefore the total mass converges to the total variation, which concludes the proof. $\square$

For $j = 1, \ldots, k-1$, we denote by X^j the unique vector field in $H(\Omega)$ with vanishing circulation around Γ_i, $i \neq j$, and unit circulation around Γ_j.

LEMMA 3. *We have that for any* $1 \leq p < 2$,

$$EX_\delta^j \to X^j,$$

and

$$EX_\delta^k \to K_\Omega[\delta(x-P)] + \sum_{j=1}^{k-1} \phi_j(P) X^j,$$

strongly in $L^p(\Omega)$ *as* $\delta \to 0$.

Proof. We can write

$$X_\delta^j = K_\Omega[\mu_j^\delta] + \sum_{l=1}^{k-1} h_\delta^{lj} X^j.$$

By Lemma 2, we have that $\mu_k^\delta \to \delta(x-P)$ weak-$*$ bounded measures. By Proposition 1 and the fact that the circulation of X_δ^k vanishes around each Γ_l, $l = 1, \ldots, k-1$, we have

$$h_\delta^{lk} = \int_\Omega \phi_l \mu_k^\delta dx \to \phi_j(P) \quad \text{as } \delta \to 0.$$

For each $1 \le p < 2$, $\mathcal{BM} \hookrightarrow W^{-1,p}$ compactly, and therefore, $\mu^\delta \to \delta(x-P)$ strongly in $W^{-1,p}$. Elliptic regularity implies that $K_\Omega[\mu^\delta] \to K_\Omega[\delta(x-P)]$ strongly in L^p, which concludes the proof of the case $j = k$.

For $j = 1, \ldots, k-1$, we have that $\mu_j^\delta \to 0$ weak-$*$ bounded measures, and therefore, using Proposition 1 and repeating the argument above concludes the argument. □

4. The time-dependent problem. The results obtained thus far have been exclusively about divergence-free vector fields in a multiply connected domain, without dynamics. In this section we restrict our attention to solutions of the incompressible 2D Euler equations and derive the required a priori estimates.

For $0 \le \delta \le \delta_0$, let Ω_δ be as before, a multiply connected domain with one of the holes small. Fix $\omega_0 \in C_c^\infty(\Omega_{\delta_0})$ and $\alpha_0 = (\alpha_{1,0}, \ldots, \alpha_{k,0}) \in \mathbb{R}^k$ and let $u_\delta = u_\delta(x,t)$ be the solution of the problem (11) with initial velocity $u_{\delta,0} = K_{\Omega_\delta}[\omega_0] + \sum_{j=1}^k \alpha_{j,0} X_\delta^j$. As we noted before, existence and uniqueness of the smooth solution u_δ is due to Kato in [6]. Let $\omega_\delta = \operatorname{curl} u_\delta$ be the vorticity, and $\alpha_{\delta,j} = \alpha_{\delta,j}(t)$ be defined by the identity

$$u_\delta = K_{\Omega_\delta}[\omega_\delta] + \sum \alpha_{\delta,j} X_\delta^j. \tag{14}$$

System (12) takes the form

$$\begin{cases} \partial_t \omega_\delta + u_\delta \cdot \nabla \omega_\delta = 0 \quad \text{in } \Omega_\delta \times (0,T), \\[2ex] u_\delta = K_{\Omega_\delta}[\omega_\delta] + \sum_{j=1}^k \alpha_{\delta,j}(t) X_\delta^j \quad \text{in } \Omega_\delta \times \{t\}, \text{ for } 0 \le t < T, \\[2ex] \alpha_{\delta,j}(t) = \int_{\Omega_\delta} \phi_{\delta,j} \omega_\delta dx + \alpha_{j,0} - \int_{\Omega_\delta} \phi_{\delta,j} \omega_0 dx \quad \text{in } [0,T), \\[2ex] \omega_\delta(x,0) = \omega_0(x) \quad \text{and} \quad \alpha_{\delta,j}(0) = \alpha_{j,0} \quad \text{in } \Omega_\delta \times \{t=0\}, \end{cases} \tag{15}$$

It follows from the transport equation for vorticity in (15) and the fact that u_δ is divergence-free that

$$\|\omega_\delta(\cdot,t)\|_{L^p(\Omega_\delta)} = \|\omega_0\|_{L^p(\Omega_\delta)} = \|\omega_0\|_{L^p(\Omega)}. \tag{16}$$

Beyond the a priori estimate given by (16), we require velocity estimates in order to study the $\delta \to 0$ asymptotics. We have the following result.

LEMMA 4. *There exists a constant* $C = C(\omega_0, \Omega)$ *such that*

$$\|K_{\Omega_\delta}[\omega_\delta](\cdot,t)\|_{L^2(\Omega_\delta)} \le C.$$

Moreover, for any $1 \le p < 2$,

$$\|K_{\Omega_\delta}[\omega_\delta](\cdot,t) - K_\Omega[\omega_\delta](\cdot,t)\|_{L^p(\Omega_\delta)} \to 0,$$

as $\delta \to 0$.

Proof. We write $K_{\Omega_\delta}[\omega_\delta] = \nabla^\perp \psi_\delta$, with ψ_δ the unique solution of

$$\begin{cases} \Delta\psi_\delta = \omega_\delta & \text{in } \Omega_\delta, \\ \psi_\delta = 0 & \text{on } \partial\Omega_\delta. \end{cases}$$

Multiply this equation by ψ_δ and integrate by parts to obtain the energy identity

$$\int_{\Omega_\delta} |\nabla\psi_\delta|^2 dx = -\int_{\Omega_\delta} \psi_\delta \omega_\delta dx. \tag{17}$$

Next we consider ψ_0, the unique solution of

$$\begin{cases} \Delta\psi_0 = |\omega_\delta| & \text{in } \Omega, \\ \psi_0 = 0 & \text{on } \partial\Omega. \end{cases}$$

We note that ψ_0 is nonpositive in Ω, and, in particular, in Ω_δ as well. The functions $\psi_0 - \psi_\delta$ and $\psi_0 + \psi_\delta$ are both nonpositive on $\partial\Omega_\delta$ and have nonnegative laplacians in Ω_δ. By the maximum principle, they are both nonpositive in Ω_δ, which implies that $|\psi_\delta| \le -\psi_0$, which, by elliptic regularity and Sobolev imbeddings, implies that

$$\|\psi_\delta\|_{L^\infty(\Omega_\delta)} \le C\|\omega_0\|_{L^p}, \tag{18}$$

for any $p > 2$. We go back to (17) using (18) and Hölder's inequality to conclude the proof of the first assertion.

For the second assertion, let ϕ_δ be the unique solution of

$$\begin{cases} \Delta\phi_\delta = \omega_\delta & \text{in } \Omega, \\ \phi_\delta = 0 & \text{on } \partial\Omega. \end{cases}$$

We have that $K_\Omega[\omega_\delta] = \nabla^\perp\phi_\delta$. Let $W_\delta = \psi_\delta - \phi_\delta$, which satisfies the system

$$\begin{cases} \Delta W_\delta = 0 & \text{in } \Omega_\delta, \\ W_\delta = 0 & \text{on } \partial\Omega, \\ W_\delta = -\phi_\delta & \text{on } \Gamma_\delta^k. \end{cases}$$

By elliptic regularity, $\|\phi_\delta\|_{W^{2,p}} \le \|\omega_0\|_{L^p}$, and therefore ϕ_δ is bounded in Γ_δ^k. Using the argument in Lemma 1, we see that $W_\delta \to 0$ uniformly away from P. Let $\Phi \in C^\infty(\Omega)$ be such that $\Phi = 0$ in a neighborhood U_0 of P and $\Phi = 1$ outside a larger neighborhood of U_1 of P. We can write

$$\operatorname{div}(\Phi\nabla W_\delta) = \nabla\Phi \cdot \nabla W_\delta \quad \text{in } \Omega_\delta,$$

multiplying this identity by W_δ, integrating in Ω_δ, and integrating by parts leads to

$$\int_{\Omega_\delta} \Phi|\nabla W_\delta| dx = (1/2)\int_{\Omega_\delta} (\Delta\Phi)|W_\delta|^2 dx,$$

which in turn leads to the estimate:

$$\|\nabla W_\delta\|^2_{L^2(\Omega_\delta\setminus U_1)} \le C\|W_\delta\|_{L^2(\Omega_\delta\setminus U_0)} \to 0 \quad \text{as } \delta\to 0.$$

The family $\{\nabla W_\delta\}$ is bounded in $L^2(\Omega)$ since $K_\Omega[\omega_\delta]$ is trivially bounded in L^2, and we have established earlier in this proof that $K_\Omega[\omega_\delta]$ is bounded in L^2 as well. For any $p>2$, given $\varepsilon>0$, we can choose $r>0$ such that $\|\nabla W^\delta\|_{L^p((B(P,r))}<\varepsilon/2$, by using the boundedness of ∇W^ε in L^2 and the Hölder equality. Next we choose δ sufficiently small so that $\|\nabla W^\delta\|_{L^p(\Omega\setminus B(P;r))}\le\varepsilon/2$, which is possible by the convergence in L^2 outside the compacts proved above. This concludes the proof. ∎

Remark. We observe that the proof above also yields the estimate

$$\|K_{\Omega_\delta}[\omega]\|_{L^2} \le C\|\omega\|^2_{L^p},$$

for any $\omega\in L^p(\Omega_\delta)$, $p>2$.

5. Passage to the limit. We now state and prove the main result of this article.

THEOREM 1. *Fix $1\le p<2$. There exists a subsequence δ_k such that ω_{δ_k} converges weak-$*$ to ω in $L^\infty((0,T);L^{p'}(\Omega))$, which satisfies (in a weak sense) the system*

$$\left\{\begin{array}{l}\omega_t+v\cdot\nabla\omega=0,\\[2ex] v=K_\Omega[\omega+\alpha_{k,0}\delta(x-P)]+\sum_{j=1}^{k-1}(\beta_j(t)+\alpha_{k,0}\phi_j(P))X^j,\\[2ex] \beta_j(t)=\int_\Omega\phi_j(x)\omega(x,t)dx+\alpha_{j,0}-\int_\Omega\phi_j(x)\omega_0(x)dx,\quad j=1,\dots,k-1,\\[2ex] \omega(x,0)=\omega_0(x)\quad \text{and}\quad \beta_j(0)=\alpha_{j,0}.\end{array}\right. \tag{19}$$

Here, P marks the location where the small obstacle disappeared, $\alpha_{k,0}$ is the circulation of the harmonic part of the initial velocity around the small obstacle, and ϕ_j is the harmonic function in Ω such that $\phi_j=1$ on the boundary component Γ_j and $\phi_j=0$ on $\partial\Omega\setminus\Gamma_j$.

Proof. Since $\{\omega_\delta\}$ is bounded in $L^{p'}(\Omega)$ by (16), we use Alaoglu's theorem to extract a subsequence, which we will not rename, which converges weak-$*$ to $\omega\in L^\infty((0,T);L^{p'}(\Omega))$.

The bulk of the proof is to pass to the limit in u^δ. We have

$$u^\delta=K_{\Omega_\delta}[\omega_\delta]+\sum_{j=1}^k\alpha_{\delta,j}(t)X^j_\delta,$$

with

$$\alpha_{\delta,j}(t)=\int_{\Omega_\delta}\phi_{\delta,j}\omega_\delta dx+\alpha_{j,0}-\int_{\Omega_\delta}\phi_{\delta,j}\omega_0 dx.$$

First, by Lemma 4 we have

$$\begin{aligned}\|K_{\Omega_\delta}[\omega_\delta]-K_\Omega[\omega]\|_{L^p(\Omega)} &\le \|K_{\Omega_\delta}[\omega_\delta]-K_\Omega[\omega_\delta]\|_{L^p(\Omega)}+\|K_\Omega[\omega_\delta]-K_\Omega[\omega]\|_{L^p(\Omega)}\\ &\le \|K_\Omega[\omega_\delta]-K_\Omega[\omega]\|_{L^p(\Omega)}+o(1).\end{aligned}$$

Next we observe that the operator K_Ω maps $L^{p'}$ continuously into $W^{1,p'}$, which is compactly embedded in L^p for any $1 \le p < 2$. Therefore, $K_\Omega[\omega_\delta] \to K_\Omega[\omega]$ strongly in L^p, and hence

$$K_{\Omega_\delta}[\omega_\delta] \to K_\Omega[\omega] \quad \text{strongly in } L^p(\Omega) \text{ as } \delta \to 0 \tag{20}$$

Next we need to study the behavior of the harmonic part of u_δ. Extending ω_δ and $\phi_{\delta,j}$ to Ω by setting them to vanish in $\Omega \setminus \Omega_\delta$, we have

$$\int_{\Omega_\delta} \phi_{\delta,j}\omega_\delta dx = \int_\Omega \phi_{\delta,j}\omega_\delta dx.$$

By Lemma 1, $\phi_{\delta,j}$ is uniformly bounded in L^∞ and converges to ϕ_j uniformly in compacts which exclude P for $j = 1, \dots, k-1$, and converges to zero uniformly in compacts which exclude P for $j = k$. For each $\varepsilon > 0$ we have

$$\|\phi_{\delta,k}\|_{L^p(B(P,r))} \le Cr^{2/p} < \frac{\varepsilon}{2}$$

if we choose $r = r(\varepsilon)$ sufficiently small. By the uniform convergence of $\phi_{\delta,k}$ outside $B(P,r)$ we may choose $\delta = \delta(\varepsilon)$ such that $\|\phi_{\delta,k}\|_{L^p(\Omega\setminus B(P,r)} \le \varepsilon/2$. This implies that $\phi_{\delta,k} \to 0$ strongly in $L^p(\Omega)$. The same argument applied to $\phi_{\delta,j} - \phi_j$, for $j = 1, \dots, k-1$, implies that $\phi_{\delta,j} \to \phi_j$ strongly in $L^p(\Omega)$. Now, ω_δ and $\phi_{\delta,k}$ form a weak-strong pair, so that we have

$$\int_\Omega \phi_{\delta,k}\omega_\delta dx \to 0 \text{ as } \delta \to 0.$$

Furthermore, ω_δ and $\phi_{\delta,j} - \phi_j$ also form a weak-strong pair, so that

$$\int_\Omega \phi_{\delta,j}\omega_\delta dx \to \int_\Omega \phi_j \omega dx \quad \text{as } \delta \to 0 \quad \text{for } j = 1, \dots, k-1.$$

Clearly, the same convergence is true replacing ω_δ by ω_0, and therefore

$$\alpha_{\delta,j} \to \int_\Omega \phi_j \omega dx + \alpha_{j,0} - \int_\Omega \phi_j \omega_0 dx \quad \text{as } \delta \to 0, \text{ for } j = 1, \dots, k-1, \tag{21}$$

and

$$\alpha_{\delta,k} \to \alpha_{k,0} \quad \text{as } \delta \to 0. \tag{22}$$

For each $j = 1, \dots, k-1$ we define

$$\beta_j = \beta_j(t) \equiv \int_\Omega \phi_j \omega dx + \alpha_{j,0} - \int_\Omega \phi_j \omega_0 dx,$$

so that (21) reads simply $\alpha_{\delta,j} \to \beta_j$ as $\delta \to 0$.

By Lemma 3 and (21) we have

$$\sum_{j=1}^{k-1} \alpha_{\delta,j} X_\delta^j \to \sum_{j=1}^{k-1} \beta_j X^j \tag{23}$$

strongly in L^2, and therefore in L^p, $1 \leq p < 2$.

By Lemma 3 and (22) we have

$$\alpha_{\delta,k} X_\delta^k \to \alpha_{k,0} \left(K_\Omega[\delta(x-P)] + \sum_{j=1}^{k-1} \phi_j(P) X^j \right), \tag{24}$$

strongly in L^p, for $1 \leq p < 2$.

Putting together (20), (23), and (24), we conclude that $u^\delta \to v$ strongly in $L^\infty((0,T); L^p(\Omega))$, with v given by

$$v(x,t) = K_\Omega[\omega + \alpha_{k,0}] + \sum_{j=1}^{k-1} (\beta_j + \phi_j(P)) X^j.$$

With v and β_j defined above, the initial conditions and the modified Biot–Savart relation on the limit system (19) are satisfied, pointwise in time. The equation for the propagation of vorticity is satisfied in a weak sense, because the weak convergence of the vorticities is enough to pass to the limit in the time-derivative term, whereas, after integration by parts, the convergence of the nonlinear term is conditioned by the behavior of the product $u^\delta \omega^\delta$. Our argument above made this product into a weak-strong pair, which therefore converges to $v\omega$ in $L^\infty((0,T); L^1(\Omega))$. This concludes the proof. $\square$

Remark. Note that the asymptotic correction due to the presence of the small obstacle not only adds a point vortex at the limit point, but it also adds a correction to the remaining harmonic components of the flow. As an illustration, let us look at the special case $\omega_0 = 0$. Harmonic vector fields are stationary solutions of the Euler equations, and one may look specifically at the asymptotic behavior of the vector field X_δ^k. Its circulation is 1 around Γ_k and zero around Γ_j, $j = 1, \dots, k-1$. By Lemma 3, this vector field converges in L^p, $1 \leq p < 2$, to $H \equiv K_\Omega[\delta(x-P)] + \sum_{j=1}^{k-1} \phi_j(P) X^j$. If we formally apply Proposition 1 to H, we find

$$\phi_j(P) = \int_\Omega \delta(x-P)\phi_j(P)dx + \oint_{\Gamma_j} H \cdot ds,$$

which implies that

$$\oint H \cdot ds = 0,$$

preserving the circulation around Γ_j of the approximations X_δ^k, as would be desired. In other words, the correction to the harmonic part of the flow exists to compensate for an imbalance of circulation around the remaining components of the boundary that appears when the flow around the small boundary component becomes that induced by a point vortex.

Remark. Theorem 1 is a compactness result, and it implies convergence for the full family of approximations only if we have uniqueness for the limit problem. This is available only in the case $\alpha_{k,0} = 0$, because then the limit problem is the standard 2D Euler equations, and Kato's theorem includes uniqueness, even at the level of weak solutions, provided that the initial data is sufficiently smooth. As in [3], uniqueness for problem (19) with smooth initial data and $\alpha_{k,0} \neq 0$ is an interesting open problem.

6. Conclusion. It is natural to ask at this point whether our analysis can be extended to the exterior domain, i.e., the full plane with k-holes instead of a bounded domain with k-holes. This is specially desirable if we wish to compare the result obtained here with that in [3], where we studied the asymptotic behavior of ideal flow in the exterior of a single small hole in the plane. First, the exterior domain version of Kato's well-posedness result is available. It is due to Kikuchi; see [7]. Therefore, the small obstacle limit for the exterior flow around multiple holes is a reasonable problem to pose.

We chose to work on a bounded domain because this greatly simplifies the problem, allowing us to focus the analysis on the small obstacle and not on infinity. Our analysis relies on the boundedness of the domain in many ways, most notably on arguments using the maximum principle. To extend our argument to exterior domains would require detailed knowledge of the asymptotic behavior of the velocity at infinity, something which we had to work hard for in the case of a single obstacle - by conformally mapping the exterior domain to the exterior of a disk and working explicitly there. With flows on the exterior of three or more holes there is no convenient symmetric domain conformally equivalent to the original one, so the work would become even more complicated. Exterior flow with two holes is conformally equivalent to an annulus, a well-known fact (see [9] and references therein), so that perhaps that case could be done by adapting the argument on [3] more directly, using the estimates for the conformal map developed in [9] when one of the holes becomes small. In the end, extending the present work to the exterior domain is an arduous technical challenge, and it will probably lead to precisely the same result.

Another natural question is what happens if one wishes to make two or more holes small. This is a trivial extension of our work, because the treatment used here shows that the asymptotics when two holes become small independently commute with each other. Therefore, two holes may be taken small in either order or simultaneously, leading to the same behavior. Also, this work extends naturally to compactly supported initial vorticities in L^p, $p > 2$. Finally, we leave open the formulation of a viscous version of the present result. A small obstacle result for viscous flow in the exterior of a single obstacle was obtained in [4], and it is natural to ask whether such an extension could be obtained for several obstacles as well, since this is the physically meaningful case.

Finally, one natural extension of the small obstacle problem with many small obstacles is the homogenization problem, where one looks for an effective equation obtained from flow outside a large number of small holes. Such a problem has been recently addressed by Lions and Masmoudi in [8]. The problem itself, as formulated in [8], has very little to do with the problem we treated here, in large part due to scaling. In their problem, both the velocity and the viscosity scale in such a way as to enforce strict two-scale asymptotic behavior. Consequently, there is no limit flow, but a generalized limit flow depending, independently and simultaneously, on the slow and fast variables and satisfying an appropriate PDE in this extended domain. In particular, the topological complications live in the microscopic (fast) scale. It should be noted, however, that some of the difficulties involved in our analysis also appear in the homogenization problem studied by Lions and Masmoudi. In their two-scale analysis, the harmonic part of the flow becomes a new unknown, of the form $\sum_i a_i(x,t)H_i(y)$, where x is the slow (physical) variable, y is the fast variable, $\{H_i\}$ is a basis for the harmonic vector fields in a microscopic periodic cell, and a_i are the components of the harmonic part. Part of their problem is to formulate an appropriate evolution equation for a_i and study its limit. The equation for the evolution of $\widehat{v}$ in [8], equation (38), is related to the equation for β in (19), with one important distinction—without Proposition 1, Lions and Masmoudi have an equation for the temporal evolution of the harmonic part, instead of a formula for it in terms of present time vorticity. It would be interesting to reformulate the analysis in [8] in terms of vortex dynamics, using Proposition 1.

Acknowledgments. The author would like to thank D. Iftimie, N. Masmoudi, H. Nussenzveig Lopes, and M. O. de Souza for many fruitful discussions.

REFERENCES

[1] J.-M. Delort, *Existence de nappes de tourbillon en dimension deux*, J. Amer. Math. Soc., 4 (1991), pp. 553–586.

[2] Ebin and Marsden, *Groups of diffeomorphisms and the notion of an incompressible fluid*, Ann. of Math. (2), 92 (1970), pp. 102–163.

[3] D. Iftimie, M. C. Lopes Filho, and H. J. Nussenzveig Lopes, *Two dimensional incompressible ideal flow around a small obstacle*, Comm. Partial Differential Equations, 28 (2003), pp. 349–379.

[4] D. Iftimie, M. C. Lopes Filho, and H. J. Nussenzveig Lopes, *Two dimensional incompressible viscous flow around a small obstacle*, Math. Ann., 336 (2006), pp. 449–489.

[5] E. Johnson and N. Robb McDonald, *The motion of a vortex near two circular cylinders*, Proc. Roy. Soc. London A, 460 (2004), pp. 939–954.

[6] T. Kato, *On classical solutions of the two-dimensional nonstationary Euler equation*, Arch. Ration. Mech. Anal., 25 (1967), pp. 188–200.

[7] K. Kikuchi, *Exterior problem for the two-dimensional Euler equation*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 30 (1983), pp. 63–92.

[8] P.-L. Lions and N. Masmoudi, *Homogenization of the Euler system in a 2D porous medium*, J. Math. Pures Appl., 84 (2005), pp. 1–20.

[9] M. Lanza de Cristoforis, *Asymptotic behavior of the conformal representation of a Jordan domain with a small hole in Schauder space*, Comput. Methods Funct. Theory, 2 (2002), pp. 1–27.

[10] S. Schochet, *The weak vorticity formulation of the 2-D Euler equations and concentration-cancellation*, Comm. Partial Differential Equations, 20 (1995), pp. 1077–1104.

[11] F. Warner, *Foundations of differentiable manifolds and Lie groups*, Scott, Foresman, and Co., Glenville, IL, 1971.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 437–448

SHARP ESTIMATES ON MINIMUM TRAVELLING WAVE SPEED OF REACTION DIFFUSION SYSTEMS MODELLING AUTOCATALYSIS*

XINFU CHEN† AND YUANWEI QI‡

Abstract. This article studies propagating wave fronts in an isothermal chemical reaction $A + 2B \to 3B$ involving two chemical species, a reactant A and an autocatalyst B, whose diffusion coefficients, D_A and D_B, are unequal due to different molecular weights and/or sizes. Explicit bounds v_* and v^* that depend on D_B/D_A are derived such that there is a unique travelling wave of every speed $v \geq v^*$ and there does not exist any travelling wave of speed $v < v_*$. New to the literature, it is shown that $v_* \propto v^* \propto D_B/D_A$ when $D_B \leq D_A$. Furthermore, when $D_A \leq D_B$, it is shown rigorously that there exists a $v_{\min}$ such that there is a travelling wave of speed v if and only $v \geq v_{\min}$. Estimates on $v_{\min}$ significantly improve those of early works. The framework is built upon general isothermal autocatalytic chemical reactions $A + nB \to (n+1)B$ of arbitrary order $n \geq 1$.

Key words. cubic autocatalysis, travelling wave, minimum speed, reaction diffusion

AMS subject classifications. 34C20, 34C25, 92E20

DOI. 10.1137/060665749

1. Introduction.

In this paper we consider an isothermal autocatalytic chemical reaction step governed by the cubic reaction relation

$$A + 2B \to 3B \qquad \text{with rate} \quad kab^2.$$

Here, $k > 0$ is the reaction rate, and a and b are the concentrations of reactant A and autocatalyst B, respectively. Well documented in the literature, the cubic reaction relation has appeared in several important models of real chemical reactions, e.g., almost isothermal flames in the carbon-sulphide-oxygen reaction (Voronkov and Semenov [22]), iodate-arsenous acid reactions (Saul and Showwalter [19]), hydroxylamine-nitrate reactions (Gowland and Stedman [10]), as well as other applications (Aris, Gray, and Scott [1] and Sel'kov [20]).

Experimental observations demonstrate the existence of propagating chemical wave fronts in chemical systems for which cubic catalysis forms a key step [11, 12, 13, 24]. These wave fronts, or travelling waves, arise due to the interaction of reaction and diffusion. Quite often when a quantity of autocatalyst is added locally into an expanse of reactant which is initially at uniform concentration, the ensuing reaction is observed to generate wave fronts which propagate outward from the initial reaction zone, consuming fresh reactant ahead of the wave front as it propagates. This is the phenomenon to be addressed in this paper.

For this purpose, we consider a one-dimensional slab geometry and the following partial differential equations (PDEs) that govern mass concentration and molecular

*Received by the editors July 21, 2006; accepted for publication February 7, 2007; published electronically June 15, 2007. This work was supported by National Science Foundation grant DMS-0504691.
http://www.siam.org/journals/sima/39-2/66574.html
†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (xinfu@pitt.edu).
‡Department of Mathematics, University of Central Florida, Orlando, FL 32816 (yqi@pegasus.cc.ucf.edu).

diffusion for the cubic reaction scheme:

$$\frac{\partial a}{\partial t} = D_A \frac{\partial^2 a}{\partial x^2} - kab^2, \qquad \frac{\partial b}{\partial t} = D_B \frac{\partial^2 b}{\partial x^2} + kab^2,$$

where D_A and D_B are the constant diffusion rates of A and B, respectively. Initial conditions, in accordance with the observed experiments, are

$$a(x,0) = a_0, \quad b(x,0) = g(x) \qquad \forall\, x \in \mathbf{R},$$

where a_0 is a positive constant representing the initial uniform distribution of A and $g(x)$ is a nonnegative function with compact support. It is not very difficult to derive from the PDEs that the solution has the following behavior at $x = \pm\infty$:

$$a(x,t) \to a_0, \quad b(x,t) \to 0 \qquad \text{as } |x| \to \infty \qquad \forall\, t \geq 0.$$

Introducing dimensionless parameters, dependent and independent variables

$$D = \frac{D_B}{D_A}, \quad \bar{a} = \frac{a}{a_0}, \quad \bar{b} = \frac{b}{a_0}, \quad \bar{t} = ka_0^2 t, \quad \bar{x} = x\sqrt{\frac{ka_0^2}{D_A}}, \qquad \bar{g} := \frac{g}{a_0},$$

and dropping the bars, the initial value problem takes the dimensionless form

$$\tag{1.1} \begin{cases} \dfrac{\partial a}{\partial t} = \dfrac{\partial^2 a}{\partial x^2} - ab^2, & x \in \mathbf{R}, t > 0, \\[2ex] \dfrac{\partial b}{\partial t} = D\dfrac{\partial^2 b}{\partial x^2} + ab^2, & x \in \mathbf{R}, t > 0, \\[2ex] a(x,0) = 1, \quad b(x,0) = g(x), & x \in \mathbf{R}, t = 0. \end{cases}$$

Here D measures the rate of diffusion of the autocatalyst relative to that of the reactant.

In the special case $D = 1$, the function $a + b$ satisfies a linear heat equation and can be solved explicitly; thus the system is reduced to a single nonlinear equation. For scalar equations, significant results are established and rich theories are available; see, for example, the works of Aronson and Weinberger [2], Chen and Guo [7], Fife and McLeod [8], and Sattinger [18] and the excellent review paper by Xin [23] for detailed information on single equations. The primary concern of the present paper is the case $D \neq 1$, which arises when the chemical species involved have different molecular weights and/or sizes. In particular, enzyme reactions may involve large enzyme molecules and smaller substrate molecules, leading to significantly different rates of diffusion. The system (1.1) also arises in epidemiology (Bailey [3]), where a represents the population density of healthy individuals and b the population density of infected individuals; again, when healthy individuals are significantly more or less mobile than the infected, D is far away from unity.

The wave front propagating phenomenon corresponds to the following behavior of solutions to (1.1): After a certain time of initiation, there are two wave fronts expanding towards $x = \pm\infty$ at a certain speed v. In between the two fronts, the reactant is consumed and thus $a \approx 0$; since each unit of reactant consumed produces exactly one unit of autocatalyst, one can expect that $b \approx 1$ inside the wave front. Outside the wave front, the reactant mixture is basically unstirred; thus $a \approx 1$ and $b \approx$

0. Focusing on the right-hand front, one expects that $(a(x,t), b(x,t)) = (\alpha(z), \beta(z))$, where $z = x - vt$ and (α, β) solves the following system:

$$
(1.2) \qquad \begin{cases} \alpha_{zz} + v\alpha_z = \alpha\beta^2, & \alpha \geq 0 \qquad \forall\, z \in \mathbf{R}, \\ D\beta_{zz} + v\beta_z = -\alpha\beta^2, & \beta \geq 0 \qquad \forall\, z \in \mathbf{R}, \\ \lim_{z\to\infty}(\ \alpha(z), \beta(z)\ ) = (1, 0), \\ \lim_{z\to-\infty}(\ \alpha(z), \beta(z)\ ) = (0, 1). \end{cases}
$$

Here $v > 0$ is the constant travelling speed.

The Travelling Wave Problem. Given $v > 0$, find $(\alpha, \beta) \in [C^2(\mathbf{R})]^2$ that satisfies (1.2).

In this paper we study the existence and nonexistence of the travelling waves, which can be generated from the initial value problem (1.1) as just described. One of the most important questions in the study of (1.2) is the existence of a minimum speed travelling wave and the estimate of the minimum speed v_{min}. In particular, for what range of v, in relation to D, does a travelling wave solution exist?

For quadratic autocatalytic ($A + B \to 2B$), namely, for the travelling wave problem related to the dynamics

$$
\frac{\partial\alpha}{\partial t} = \frac{\partial^2\alpha}{\partial x^2} - \alpha\beta, \qquad \frac{\partial\beta}{\partial t} = D\frac{\partial^2\beta}{\partial x^2} + \alpha\beta,
$$

Billingham and Needham proved that there is a travelling wave if and only if $v \geq 2\sqrt{D}$; namely, the minimum wave speed is exactly $2\sqrt{D}$ (see also the works of Billingham, Merkin, and Needham [4, 5, 14, 15, 16]). Focant and Gallay [9] investigated the existence and stability of travelling waves when both the quadratic and cubic nonlinearities are present in the system.

For the cubic autocatalysis, i.e., for (1.2), the answer is far from complete. Based on an invariant region argument, it was shown in [4] that a travelling wave exists if its speed $v \geq 2\sqrt{D}$. A more recent work [17] by one of the authors improved the result of [4] to the following: for (1.2),

(a) there exists a solution if $\quad v \geq \begin{cases} \sqrt{2D-1} & \text{when } D \geq 1, \\ \sqrt{D} & \text{when } D < 1; \end{cases}$

(b) no solution exists if $\quad v \leq \begin{cases} \sqrt{D/6} & \text{when } D \geq 1, \\ D/\sqrt{6} & \text{when } D < 1. \end{cases}$

This result, which comes out of a much more delicate analysis than that in [4], supplied both upper and lower bounds on minimum wave speed; nevertheless, it is still far from answering the key question of providing a good estimate of minimum speed. In particular, it falls short of providing an accurate order of $v_{\min}$ in terms of small D. Numerical simulation by the authors of [4] suggests that $v_{\min} \propto D$ when $D \ll 1$. Furthermore, it is well known that $v_{min} = 1/\sqrt{2}$ when $D = 1$, but neither the results in [4] nor the results in [17] recover this special case from their results of the general case. In this paper we shall provide affirmative answers to these questions and fill in the gap between the general case and the special case of $D = 1$.

THEOREM 1. *Suppose $D < 1$. For the travelling wave problem* (1.2),

(i) *there exists a unique (up to translation) solution if* $\quad v \geq \dfrac{4D}{\sqrt{1+4D}}$;

(ii) *there does not exist any solution if* $\quad v < \dfrac{D}{\sqrt{2}}.$

Clearly the above result provides a pretty satisfying bound on the range of wave speeds. In particular, it shows that $v_{\min}(D) \propto D$ for small D.

One of the important issues in discussing existence and nonexistence of a travelling wave solution is whether the set of v of the speed for which existence holds is a single interval. While there are heuristic and numerical arguments in [4] demonstrating that the set of admissible wave speed is an interval $[v_{\min}, \infty)$, for the moment we can only supply a rigorous proof for the case $D \geq 1$.

THEOREM 2. *Suppose $D \geq 1$. There exists a positive constant $v_{\min}$ such that* (1.2) *admits a solution if and only if $v \geq v_{\min}$. In addition, $v_{\min}$ satisfies the estimate*

$$\sqrt{\frac{D}{2}} \leq v_{\min} \leq \sqrt{\frac{D}{1+1/D}}.$$

It is clear from Theorem 2 that in the special case of $D = 1$, (1.2) admits a solution if and only if $v \geq 1/\sqrt{2}$.

The general nth order isothermal autocatalytic chemical reaction step is governed by the chemical reaction relation

$$A + nB \rightarrow (n+1)B \qquad \text{with rate} \quad kab^n.$$

We can use the same idea developed in this paper to establish lower bounds for the existence of a travelling wave solution and upper bounds for nonexistence. For this general case, the governing equations are, after proper scaling,

$$\frac{\partial a}{\partial t} = \frac{\partial^2 \alpha}{\partial x^2} - ab^n, \qquad \frac{\partial b}{\partial t} = D\frac{\partial^2 b}{\partial x^2} + ab^n,$$

where $D = D_B/D_A$ and the initial value is the same as that in (1.1). The corresponding travelling wave problem is to solve

$$(1.3) \qquad \begin{cases} \alpha_{zz} + v\alpha_z = \alpha\beta^n, & \alpha \geq 0 \qquad \forall\, z \in \mathbf{R}, \\ D\beta_{zz} + v\beta_z = -\alpha\beta^n, & \beta \geq 0 \qquad \forall\, z \in \mathbf{R}, \\ \lim_{z\to\infty}(\ \alpha(z), \beta(z)\ ) = (1,0), \\ \lim_{z\to-\infty}(\ \alpha(z), \beta(z)\ ) = (0,1). \end{cases}$$

THEOREM 3. *Suppose $D < 1$ and $n \geq 2$. A unique (up to translation) travelling wave solution exists for* (1.3) *if $v \geq 4D/\sqrt{1+4D}$. On the other hand, there exists no solution for* (1.3) *if $v \leq D/\sqrt{K(n)}$, where $K(n)$ is a constant which increases with n. In particular, $K(1) = 1/4$, $K(2) = 2$.*

THEOREM 4. *Suppose $D \geq 1$ and $n \geq 1$. There exists a positive constant $v_{\min}$ such that* (1.3) *admits a travelling wave if and only if $v \geq v_{\min}$. In addition, $v_{\min}$ is bounded by*

$$\sqrt{\frac{D}{K(n)}} \leq v_{\min} \leq \sqrt{\frac{D}{K(n)}}\ \frac{1}{\sqrt{1-(1-\frac{1}{D})\frac{\sqrt{4K(n)+1}-1}{\sqrt{4K(n)+1}+1}}},$$

where $K(n)$ is the same constant as in Theorem 3.

We note in passing the recent works to study the spatiotemporal profiles of L^1 initial values by Bricmont, Kupiainen, and Xin [6] and the steady-state solutions by Shi and Wang [21].

The organization of this paper is as follows. Section 2 contains preliminary analysis and an outline of our approach. The case $D \geq 1$ is discussed in section 3, and the case $D < 1$ in section 4.

2. Preliminary.

2.1. A scalar equation. First we review the existence of a travelling wave solution of unit speed to the equation

$$u_{zz} + u_z = ku(1-u)^n, \quad 0 \leq u \leq 1 \quad \text{on } \mathbf{R}, \qquad u(-\infty) = 0, \quad u(\infty) = 1. \tag{2.1}$$

Here $n \geq 1$ is a parameter, and k is a positive constant. We seek upper bounds on k for the existence of a solution. Since a solution, if it exists, satisfies $u_z > 0$ on $\mathbf{R}$, we can write $u' = Q(u)$ and work on the (u, Q) phase plane. The resulting equation on the phase plane is

$$\begin{cases} QQ' + Q = ku(1-u)^n & \forall\, u \in [0,1], \\ Q(0) = 0, \quad Q > 0 \text{ on } (0,1). \end{cases} \tag{2.2}$$

There is a one-to-one correspondence between solutions to (2.1) and solutions to (2.2) satisfying the additional requirement $Q(1) = 0$.

LEMMA 2.1. *For each $n \geq 1$ and $k > 0$, there exists a unique solution $Q = Q(n, k; \cdot)$ to (2.2). In addition, there exists a positive constant $K(n)$ such that $Q(n, k; 1) = 0$ if $k \in (0, K(n)]$ and $Q(n, K; 1) > 0$ if $k \in (K(n), \infty)$. Consequently, (2.2) admits a solution if and only if $k \in (0, K(n)]$.*

In addition, $K(n)$ is a strictly increasing function of n and $K(1) = \frac{1}{4}, K(2) = 2$.

Proof. The existence of Q and K follows by the comparison principle. The exact value of $K(1)$ is calculated by a known fact that the function $K(1)u(1-u)$ is concave, and thus the minimum wave speed $v = 1$ satisfies $1 = 2\sqrt{K(1)}$; hence $K(1) = 1/4$. In the case $n = 2$, the exact solution is given by $Q = u(1-u)$; thus $K(2) = 2$. We omit details, because it is a standard argument. ∎

2.2. Basic properties of travelling waves. Suppose (v, α, β) solves (1.3). Then $[\alpha_z + v\alpha + D\beta_z + v\beta]_z = 0$, so that $\alpha_z + D\beta_z + v(\alpha + \beta)$ is a constant function. Using the boundary conditions, we find that

$$\alpha_z + D\beta_z + v(\alpha + \beta - 1) = 0 \qquad \text{on } \mathbf{R}.$$

With the new variable $w = \beta_z$, (1.3) is equivalent to the following third order ODE system

$$\begin{cases} \alpha_z = v(1 - \alpha - \beta) - Dw, \\ \beta_z = w, \\ w_z = -D^{-1}(\alpha\beta^n + vw), \\ \lim_{z\to\infty}(\alpha(z), \beta(z), w(z)) = (1, 0, 0), \\ \lim_{z\to-\infty}(\alpha(z), \beta(z), w(z)) = (0, 1, 0). \end{cases} \tag{2.3}$$

It is clear that in the (α, β, w) phase space, there are two equilibrium points: $(0, 1, 0)$ and $(1, 0, 0)$. The following are a few basic properties of travelling wave solutions.

PROPOSITION 1. *The systems* (1.3) *and* (2.3) *are equivalent. Any solution* (α, β) *to* (1.3) *or* (α, β, w) *to* (2.3) *has the following properties:*

(1) $\alpha_z > 0 > \beta_z$ *on* $\mathbf{R}$.
(2) $\alpha + \beta < 1$ *on* $\mathbf{R}$ *if* $D < 1$, $\alpha + \beta \equiv 1$ *if* $D = 1$, *and* $\alpha + \beta > 0$ *if* $D > 1$.
(3) $v = \int_{-\infty}^{\infty} \alpha(z)\beta^n(z)\,dz > 0$.
(4) *The equilibrium point* $(0, 1, 0)$ *of* (2.3) *is a saddle with a two-dimensional stable manifold and a one-dimensional unstable manifold. The eigenvalues and associated eigenvectors are*

$$\begin{aligned} \lambda_1 &= -vD^{-1}, & \mathbf{e}_{\lambda_1} &= (0, -1, -\lambda_1)^T, \\ \lambda_2 &= -\tfrac{1}{2}(\sqrt{v^2+4} + v), & \mathbf{e}_{\lambda_2} &= (\lambda_2(D\lambda_2 + v), -1, -\lambda_2)^T, \\ \lambda_3 &= \tfrac{1}{2}(\sqrt{v^2+4} - v), & \mathbf{e}_{\lambda_3} &= (\lambda_3(D\lambda_3 + v), -1, -\lambda_3)^T. \end{aligned}$$

(5) *When* $n > 1$, *the equilibrium point* $(1, 0, 0)$ *is degenerate; it has a two-dimensional stable manifold and a one-dimensional center manifold. The eigenvalues and associated eigenvectors are*

$$\begin{aligned} \mu_1 &= -v, & \mathbf{e}_{\nu_1} &= (1, 0, 0)^T, \\ \mu_2 &= -vD^{-1}, & \mathbf{e}_{\nu_2} &= (0, -1, -vD^{-1})^T, \\ \mu_3 &= 0, & \mathbf{e}_{\nu_3} &= (1, -1, 0)^T. \end{aligned}$$

All items except (3) were proved in [4]. The equation in property (3) is obtained by integrating the equation involving α_{zz} in (1.3) with the boundary conditions $\alpha(\infty) = 1$ and $\alpha(-\infty) = 0$.

The third property in the proposition demonstrates that $v > 0$. The fourth property clearly tells us that the travelling wave we are looking for is indeed the one-dimensional unstable manifold associated with the equilibrium $(0, 1, 0)$. Hence, given $v > 0$, a travelling wave of speed v, if it exists, is unique up to a translation.

2.3. New setting—A nonautonomous 2×2 system. Unlike in earlier works [4, 17], here we shall use a transformation to turn the third order autonomous system (2.3) into a second order nonautonomous system, using $u := 1 - \beta$ as the independent variable. This is allowed since for the solution of interest, $\beta_z < 0$, and thus $z \to 1 - \beta(z)$ has an inverse. To make the resulting system as simple as possible, we also scale the other variables. Hence, we introduce

$$u = 1 - \beta, \qquad A = \frac{D\alpha}{v^2}, \qquad y = \frac{vz}{D}, \qquad \kappa := \frac{D}{v}.$$

The system of differential equations (1.3) becomes

$$\begin{cases} u_{yy} + u_y = A(1-u)^n & \text{on } \mathbf{R}, \\ A_y = \kappa^2(u + u_y) - DA & \text{on } \mathbf{R}. \end{cases}$$

Since $u_y > 0$ for the solution of interest, we can use u as the independent variable. Introducing $P(u) = u_y$, we have an equivalent system of second order nonautonomous (singular) ODEs:

$$\begin{cases} PP' = A[1-u]^n - P & \forall u \in [0, 1], \\ PA' = \kappa^2[P + u] - DA & \forall u \in [0, 1], \\ P(u) > 0, \ A(u) > 0 & \forall\, u \in (0, 1), \\ P(0) = 0, \ A(0) = 0. \end{cases} \tag{2.4}$$

LEMMA 2.2. *For every $D > 0$ and $\kappa > 0$, (2.4) admits a unique solution. In addition,*

$$P(u) = \lambda\, u + O(u^2), \quad A(u) = \lambda(1+\lambda)u + O(u^2) \qquad \text{as } u \searrow 0, \tag{2.5}$$

where

$$\lambda := \tfrac{1}{2}(\sqrt{4\kappa^2 + D^2} - D) \qquad (\text{the only positive root to } \lambda(\lambda + D) = \kappa^2).$$

Furthermore, $A'(u) > 0$ for all $u \in [0, 1)$, and there are only two possible cases:

(a) $P(1) > 0$*: there does not exist any travelling wave solution to* (1.3)*.*

(b) $P(1) = 0$*: there exists a travelling wave solution to* (1.3)*, unique up to translation.*

Proof. We divide the proof into several steps.

1. A solution to (2.4) corresponds exactly to the part of the one-dimensional unstable manifold associated with the equilibrium point $(0, 1, 0)$ of the autonomous system (2.3) that has the property $\alpha,\ \beta > 0$ and $w < 0$. Hence, for some $\delta \in (0, 1]$, (2.4) admits a unique solution in $[0, \delta)$. The solution satisfies the asymptotic expansion (2.5) and can be extended as long as $P > 0$.

2. It is easy to see that A cannot hit zero before P does, since otherwise, $PA' = \kappa^2(p + u) > 0$ at $A = 0$ and $P > 0$, which is impossible. If $\lim_{u\nearrow\delta} P(u) = 0$ at some $\delta \in (0, 1)$, then $\liminf_{u\nearrow\delta} P(u)P'(u) \le 0$, and thus $A(\delta) := \lim_{u\nearrow\delta} A(u) = 0$. But the equation for A gives $PA' > \kappa^2\delta/2 > 0$ for all u sufficiently close to δ from below, which contradicts $A(\delta) = 0$. Since the system has at most a linear growth in P and A, the solution can be uniquely extended to $[0, 1)$ and $P > 0,\ A > 0$ in $(0, 1)$.

3. Now we show that $A' > 0$ on $[0, 1)$. From the asymptotic behavior (2.5), $A \in C^1([0, 1))$ and $A'(0) = \lambda(\lambda + 1) > 0$. Also a combination of the two equations in (2.4) yields

$$P[\kappa^2(P + u) - DA]' = -D[\kappa^2(P + u) - DA] + \kappa^2 A(1 - u)^n.$$

Gronwall's inequality then gives $\kappa^2(P + u) - DA > 0$ in $(0, 1)$. Thus,

$$0 < A < D^{-1}\kappa^2(P + u) \ \text{ in } \ (0, 1), \qquad A' > 0 \text{ in } \ [0, 1).$$

4. Since $(P+u)' = A(1-u)^n/P > 0$ in $(0, 1)$, $P+u$ is strictly increasing in $(0, 1)$ so that $\lim_{u\nearrow 1} P(u)$ exists. To show that it is finite, observe that when $P \ge 1$,

$$[P + u]' \le P[P + u]' = A(1 - u)^n \le D^{-1}\kappa^2(P + u)(1 - u)^n.$$

This implies that $P + u$ is bounded uniformly in $u \in [0, 1)$; thus $P(1) := \lim_{u\nearrow 1} P(u)$ exists and is finite. Consequently, $A(1) := \lim_{u\nearrow 1} A(u)$ also exists and is finite.

5. If $P(1) > 0$, we have a classical solution of (2.4) on $[0, 1]$. Since a travelling wave is required to have $u = 1-\beta \le 1$, we see that there is no travelling wave solution to (1.3).

6. Suppose $P(1) = 0$. Since $[P+u]' > 0$ in $(0, 1)$, we have $P(u)+u < P(1)+1 = 1$; i.e., $P(u) < 1 - u$ for all $u \in [0, 1)$. Since $\kappa^2(P + u) - DA > 0$ in $(0, 1)$, we see that

$$\begin{aligned} A(1) &= \int_0^1 \frac{\kappa^2[P(u) + u] - DA(u)}{P(u)} du \\ &\ge \int_0^1 \frac{\kappa^2[P(u) + u] - DA(u)}{1 - u} du \ge \int_0^1 \frac{\kappa^2[u - 1]}{1 - u} du + \int_0^1 \frac{\kappa^2 - DA(u)}{1 - u} du. \end{aligned}$$

Since $DA(u) < DA(1) \le \kappa^2(P(1)+1) = \kappa^2$ for all $u \in (0,1)$, for the last integral to be convergent, we must have $A(1) = \kappa^2/D$. It is then easy to see that $\alpha = DA/\kappa^2 \to 1$ as $u \to 1$. Transferring back to the original variable z, we then obtain a travelling wave solution to (1.3). $\square$

In what follows, we shall estimate upper and lower bounds of A/u; thus Lemma 2.1 can be applied to generate upper and lower bounds of $v_{\min}$.

3. The case $D \ge 1$. In this section we deal with the case of $D \ge 1$.

LEMMA 3.1. *Suppose $D \ge 1$. Then $DA(u) \ge \kappa^2 u$ for all $u \in [0,1]$. Consequently, there is no travelling wave solution to* (1.3) *when $\kappa^2 > DK(n)$, i.e., when $v < \sqrt{D/K(n)}$.*

Proof. If $D = 1$, $A(u) = \kappa^2 u$ for all $u \in [0,1]$. When $D > 1$, for every $u \in (0,1)$,

$$P[DA - \kappa^2 u]' = -D[DA - \kappa^2 u] + (D-1)\kappa^2 P > -D[DA - \kappa^2 u].$$

In addition, when u is sufficiently small, $DA(u) = D(1+\lambda)\lambda\, u + O(u^2) > [D+\lambda]\lambda\, u = \kappa^2 u$. Applying Gronwall's inequality, we derive that $DA > \kappa^2 u$ on $(0,1)$.

Now suppose $\kappa^2 > DK(n)$. Let $\hat{k} \in (K(n), k^2/D)$. Then $A(u) \ge \hat{k}u$ on $[0,1]$ so that

$$PP' + P = A(1-u)^n \ge \hat{k}u(1-u)^n \quad \forall\, u \in [0,1].$$

We compare $P(u)$ and the solution $Q(n,\hat{k};u)$ given in Lemma 2.1. Using a Taylor expansion, we can show that $P(u) > Q(n,\hat{k};u)$ for all $u \in (0,\epsilon]$ for some $\epsilon > 0$. In the interval $[\epsilon,1]$ we can use the regular comparison principle to show that $P(u) > Q(n,\hat{k};u)$ for all $u \in [\epsilon,1)$. In particular, $P(1) \ge Q(n,\hat{k};1) > 0$, so that there is no travelling wave solution to (1.3). Since $\kappa = D/v$, the condition $\kappa^2 > DK(n)$ is the same as $v < \sqrt{D/K(n)}$. $\square$

LEMMA 3.2. *Suppose $D > 1$. Then,*

$$A(u) < \lambda(1+\lambda)\, u, \qquad P(u) < \lambda\, u \quad \forall\, u \in (0,1).$$

Consequently, there exists a travelling wave solution to (1.3) *when $\lambda(\lambda+1) \le K(n)$, i.e., when*

$$v \ge \sqrt{\frac{D}{K(n)}}\,\frac{1}{\sqrt{1-(1-\frac{1}{D})\frac{\sqrt{4K(n)+1}-1}{\sqrt{4K(n)+1}+1}}}.$$

Proof. A higher order Taylor expansion near $u = 0$ shows that $A < \lambda(\lambda+1)\, u$ and $P < \lambda\, u$ for all sufficient small positive u. Set

$$\hat{B} = \sup\{b \in (0,1) \mid P(u) < \lambda\, u, \qquad A(u) < \lambda(1+\lambda)u \quad \forall u \in (0,b)\}.$$

We show that $\hat{B} = 1$. Suppose on the contrary that $\hat{B} < 1$. Then either $P(\hat{B}) - \lambda\hat{B} = 0$ or $A(\hat{B}) - \lambda(1+\lambda)\hat{B} = 0$. In $(0,\hat{B}]$,

$$\begin{aligned} P[A - \lambda(1+\lambda)u]' &= \kappa^2(P+u) - DA - \lambda(1+\lambda)P \\ &= \lambda(D+\lambda)(P+u) - DA - \lambda(1+\lambda)P \\ &= -D[A - \lambda(1+\lambda)u] + \lambda(D-1)(P - \lambda u) \\ &\le -D[A - \lambda(1+\lambda)u]. \end{aligned}$$

Gronwall's inequality then implies that $A < \lambda(\lambda+1)u$ on $(0, \hat{B}]$. Similarly, for all $u \in (0, \hat{B}]$,

$$\begin{aligned} P[P-\lambda u]' &= -(1+\lambda)P + A(1-u)^n \\ &= -(1+\lambda)(P-\lambda u) - \lambda(1+\lambda)u + A(1-u)^n \\ &< -(1+\lambda)(P-\lambda u). \end{aligned}$$

Gronwall's inequality shows that $P < \lambda\, u$ on $(0, \hat{B}]$. We reach a contradiction. This proves that $\hat{B} = 1$; i.e., $P(u) < \lambda\, u$ and $A(u) < \lambda(1+\lambda)u$ for all $u \in (0,1)$.

Suppose $\lambda(1+\lambda) \le K(n)$. We can use comparison to show that $P(u) \le Q(n, K(n); u)$ for all $u \in [0,1]$ so that $P(1) = 0$. Namely, there exists a travelling wave solution to (1.3). □

Proof of Theorem 2. This is a special case of Theorem 4, by setting $n = 2$ and acknowledging that $K(2) = 2$. □

Proof of Theorem 4. The estimate of $v_{\min}$, when it exists, follows from the above two lemmas.

We notice that the set of admissible speed is a closed set. Indeed, if there is no travelling wave of speed $\hat{v} > 0$, then the solution (P, A) to (2.4) with $v = \hat{v}$ has the property that $P(1) > 0$. It then follows by continuous dependence that for any v sufficiently close to $\hat{v}$, the solution to (2.4) also satisfies $P(1) > 0$. This implies that there is no travelling wave of speed v for any v sufficiently close to $\hat{v}$. Thus the complement of the set of admissible speed is open; that is, the set of admissible speed is closed.

Hence, to show the existence of $v_{\min}$, it suffices to show that if $v_1 > v_0$ and there exists a travelling wave of speed v_0, then there also exists a travelling wave of speed v_1. For this, we denote $\kappa_i = D/v_i$ and (P_i, A_i) the solution to (2.4) with $\kappa = D/v_i$, $i = 0, 1$. The existence of a travelling wave of speed v_0 implies that $P_0(1) = 0$. To show that there exists a travelling wave of speed v_1, it is necessary and sufficient to show that $P_1(1) = 0$. For this, it suffices to show that $P_1 < P_0$ in $(0,1)$.

Notice that $\kappa_1 < \kappa_0$. Denote by λ_i the positive root to $\lambda_i(\lambda_i + D) = \kappa_i^2$. Then $\lambda_1 < \lambda_0$. The asymptotic expansion for (P, A) then implies that there exists $\epsilon > 0$ such that for $u \in (0, \epsilon]$, $P_1(u) < P_0(u)$ and $A_1(u) < A_0(u)$. In addition, for small u, the functions $\alpha_i := DA_i/\kappa_i^2$ satisfy

$$\begin{aligned} \alpha_0 - \alpha_1 &= \left\{ \frac{D\lambda_0(\lambda_0+1)}{\kappa_0^2} - \frac{D\lambda_1(\lambda_1+1)}{\kappa_1^2} \right\} u + O(u^2) \\ &= D\left\{ \frac{\lambda_0+1}{\lambda_0+D} - \frac{\lambda_1+1}{\lambda_1+D} \right\} u + O(u^2) \\ ;6pt] &= \frac{D(D-1)(\lambda_0-\lambda_1)}{(\lambda_0+D)(\lambda_1+D)} u + O(u^2) > 0 \end{aligned}$$

since $D > 1$ and $\lambda_0 > \lambda_1$. Now let

$$\hat{B} = \sup\{b \in (0,1) \mid P_1(u) < P_0(u)\ \forall\, u \in (0, b)\}.$$

We claim that $\hat{B} = 1$. Suppose the contrary, $\hat{B} < 1$. Then $P_0(\hat{B}) = P_1(\hat{B}) > 0$.

First we claim that $A_0 > A_1$ on $(0, \hat{B}]$. Suppose it is not true; then there is a $u_1 \in (0, \hat{B}]$ at which $A_0(u_1) = A_1(u_1)$. Since $\kappa_0 > k_1$, there exists $u_2 \in (0, B_1)$ such that $\alpha_0(u_2) = \alpha_1(u_2)$ and $\alpha_0(u_2)' \le \alpha_1(u_2)'$. But, at $u = u_2$,

$$[\alpha_0 - \alpha_1]' = \frac{D(u-\alpha_0)}{P_0} - \frac{D(u-\alpha_1)}{P_1} = \frac{D(\alpha_0-u)(P_0-P_1)}{P_0P_1} > 0$$

since $\alpha_0 - u = [DA_0 - \kappa_0^2 u]/\kappa_0^2 > 0$ by Lemma 3.1 and $P_0 > P_1$ in $(0, \hat{B}) \ni u_2$. Thus, we must have $A_0 > A_1$ in $[0, \hat{B}]$. Consequently, we obtain from the equation for P_i that

$$\tfrac{1}{2}[P_1^2 - P_0^2]' = [P_0 - P_1] + (A_1 - A_0)[1-u]^n < [P_0 - P_1] = \frac{P_0^2 - P_1^2}{P_0 + P_1}.$$

Gronwall's inequality on $[\epsilon, \hat{B}]$ then gives $P_1^2 - P_0^2 < 0$ on $[\epsilon, \hat{B}]$, contradicting $P_0(\hat{B}) = P_1(\hat{B})$. Hence, $\hat{B} = 1$ and $P_1 < P_0$ on $(0,1)$. This completes the proof of Theorem 4. ∎

4. The case of $D < 1$. In this section, we establish the results on the case of $D < 1$.

LEMMA 4.1. *Suppose $D < 1$. Then $A > \kappa^2 u$ on $(0,1)$. Consequently, when $\kappa^2 > K(n)$, i.e., $v < D/\sqrt{K(n)}$, there is no travelling wave solution to* (1.3).

Proof. Direct calculation shows that

$$\begin{aligned} P[A - \kappa^2 u]' &= \kappa^2(P+u) - DA - \kappa^2 P = \kappa^2(1-D)u - D(A - \kappa^2 u) \\ &> -D(A - \kappa^2 u) \quad \forall u \in (0,1). \end{aligned}$$

Since $A = \lambda(1+\lambda)u + O(u^2) > \kappa^2 u$, for all sufficiently small positive u, Gronwall's inequality gives $A > \kappa^2 u$ on $[0,1)$.

One can show that $P(u) > Q(n, k^2; u)$ for all $u \in (0,1)$ by first using an asymptotic expansion at $u = 0$ for $0 < u \leq \epsilon$ and then a comparison principle for the differential equation in $(\epsilon, 1)$.

It then follows from Lemma 2.1 that when $\kappa^2 > K(n)$, we must have $P(1) \geq Q(n, k^2; 1) > 0$; i.e., there does not exist any solution to the travelling wave problem. ∎

To establish the existence of a solution, we need to find an upper bound of A. Although there is the estimate $A < \kappa^2(u+P)/D$ available for use, we are not satisfied with such an estimate since when D is very small, it is not sufficient to show that $v_{\min} = O(D)$. Hence, we seek another bound.

LEMMA 4.2. *Suppose $D < 1$. Then $A(u)(1-u)^{n/2} \leq \lambda[P(u) + u]$ $\forall\, u \in [0,1)$.*

Proof. When $u = 0$, the two sides are equal. Computation shows that, in $(0,1]$,

$$\begin{aligned} &P[(1-u)^{n/2}A - \lambda(P+u)]' \\ &= (1-u)^{n/2}[\kappa^2(P+u) - DA] - \tfrac{1}{2}nPA(1-u)^{n/2-1} - \lambda A(1-u)^n \\ &\leq -[D + \lambda(1-u)^{n/2}][A(1-u)^{n/2} - \lambda(P+u)] + (P+u)[(\kappa^2 - \lambda^2)(1-u)^{n/2} - \lambda D] \\ &= -[D + \lambda(1-u)^{n/2}][A(1-u)^{n/2} - \lambda(P+u)] - \lambda D(P+u)[1 - (1-u)^{n/2}] \\ &\leq -[D + \lambda(1-u)^{n/2}][A(1-u)^{n/2} - \lambda(P+u)]. \end{aligned}$$

Here we have dropped the term $\frac{1}{2}nP(1-u)^{n/2-1}$ in the first inequality and used $\kappa^2 = \lambda(\lambda + D)$ in the second inequality. The assertion of the lemma thus follows from Gronwall's inequality. ∎

Proof of Theorem 3. The nonexistence follows directly from Lemma 4.1. We now prove the existence. Simple computation shows that $v \leq 4D/\sqrt{1+4D}$ is equivalent to $\lambda \leq 1/4$. We proceed to show that $P - u(1-u)/2 \leq 0$ on $(0,1)$. It is easy to verify, using the result of Lemma 4.2, that

$$
\begin{aligned}
P[2P - u(1-u)]' &= P(2u-3) + 2A(1-u)^n \\
&\le P(2u-3) + 2\lambda(P+u)(1-u)^{n/2} \\
&= [u - 3/2 + \lambda(1-u)^{n/2}][2P - u(1-u)] \\
&\qquad + u(1-u)[2\lambda(1-u)^{n/2-1} + \lambda(1-u)^{n/2} + u - 3/2] \\
&< [u - 3/2 + \lambda(1-u)^{n/2}][2P - u(1-u)],
\end{aligned}
$$

since $\lambda \le 1/4$ and $n \ge 2$ yield

$$
\begin{aligned}
&2\lambda(1-u)^{n/2-1} + \lambda(1-u)^{n/2} + u - 3/2 \\
&\le 2\lambda + \lambda(1-u) + u - 3/2 \\
&= 2\lambda - 1/2 + (\lambda - 1)(1-u) \le 0.
\end{aligned}
$$

Because $2P < u(1+u)$ for small u, Gronwall's inequality shows that $P < u(1-u)/2$ on $(0,1)$. Thus $P(1) = 0$. This proves the existence and completes the proof of the theorem. ∎

Finally, Theorem 1 is a special case of Theorem 3 with $n = 2$.

Acknowledgments. The authors thank J. B. McLeod and Yi Li for stimulating discussions.

REFERENCES

[1] R. Aris, P. Gray, and S. K. Scott, *Modelling of cubic autocatalysis by successive biomolecular steps*, Chem. Eng. Sci., 43 (1988), pp. 207–211.

[2] D. G. Aronson and H. F. Weinberger, *Multidimensional diffusion arising in population genetics*, Adv. Math., 30 (1978), pp. 33–76.

[3] N. J. T. Bailey, *The Mathematical Theory of Infectious Diseases and its Applications*, Griffin, London, 1975.

[4] J. Billingham and D. J. Needham, *The development of travelling waves in quadratic and cubic autocatalysis with unequal diffusion rates.* I. *Permanent form travelling waves*, Philos. Trans. Roy. Soc. London Ser. A, 334 (1991), pp. 1–24.

[5] J. Billingham and D. J. Needham, *The development of travelling waves in quadratic and cubic autocatalysis with unequal diffusion rates.* II. *An initial value problem with an immobilized or nearly immobilized autocatalyst*, Philos. Trans. Roy. Soc. London Ser. A, 336 (1991), pp. 497–539.

[6] J. Bricmont, A. Kupiainen, and J. Xin, *Global large time self-similarity of a thermal-diffusive combustion system with critical nonlinearity*, J. Differential Equations, 130 (1996), pp. 9–35.

[7] X. Chen and J.-S. Guo, *Existence and asymptotic stability of traveling waves of discrete quasilinear monostable equations*, J. Differential Equations, 184 (2002), pp. 549–569.

[8] P. C. Fife and J. B. McLeod, *The approach of nonlinear diffusion equations to travelling wave front solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 335–361.

[9] S. Focant and Th. Gallay, *Existence and stability of propagating fronts for an autocatalytic reaction-diffusion system*, Phys. D, 120 (1998), pp. 346–368.

[10] R. J. Gowland and G. Stedman, *A novel moving boundary reaction involving hydoxylamine and nitric acid*, J. Chem. Soc. Chem. Commun., 10 (1983), pp. 1038–1039.

[11] P. Gray, J. F. Griffiths, and S. K. Scott, *Experimental studies of the ignition diagram and the effect of added hydrogen*, Proc. R. Soc. Lond. Ser. A, 397 (1984), pp. 21–44.

[12] P. Gray and S. K. Scott, *Chemical Oscillations and Instabilities*, Clarendon Press, Oxford University Press, Oxford, UK, 1990.

[13] A. Hanna, A. Saul, and K. Showalter, *Detailed studies of propagating fronts in the iodate oxidation of arsenous acid*, J. Am. Chem. Soc., 104 (1982), pp. 3838–3844.

[14] J. H. Merkin and D. J. Needham, *Propagating reaction-diffusion waves in a simple isothermal quadratic autocatalytic chemical system*, J. Engrg. Math., 23 (1989), pp. 343–356.

[15] J. H. Merkin and D. J. Needham, *The development of travelling waves in a simple isothermal chemical system.* II. *Cubic autocatalysis with quadratic and linear decay*, Proc. Roy. Soc. London Ser. A, 430 (1990), pp. 315–345.
[16] J. H. Merkin and D. J. Needham, *The development of travelling waves in a simple isothermal chemical system.* IV. *Cubic autocatalysis with quadratic decay*, Proc. Roy. Soc. London Ser. A, 434 (1991), pp. 531–554.
[17] Y. W. Qi, *The development of travelling waves in cubic auto-catalysis with different rates of diffusion*, Phys. D, 226 (2007), pp. 129–135.
[18] D. Sattinger, *On the stability of waves of nonlinear parabolic systems*, Adv. Math., 22 (1976), pp. 312–355.
[19] A. Saul and K. Showalter, *Propagating reaction-diffusion fronts*, in Oscillations and Travelling Waves in Chemical Systems, R. J. Field and M. Burgern, eds., Wiley, New York, 1984.
[20] V. G. E. E. Sel'kov, Euro. J. Biochem., 4 (1968), p. 79.
[21] J. Shi and X. Wang, *Hair-triggered instability of radial steady states, spread and extinction in semilinear heat equations*, J. Differential Equations, 231 (2006), pp. 235–251.
[22] V. G. Voronkov and N. N. Semenov, Zh. Fiz. Khim., 13 (1939), p. 1695.
[23] J. Xin, *Front propagation in heterogeneous media*, SIAM Rev., 42 (2000), pp 161–230.
[24] A. N. Zaikin and A. M. Zhabotinskii, *Concentration wave propagation in two-dimensional liquid-phase self-oscillating system*, Nature, 225 (1970), pp. 535–537.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 449–469

ELASTIC ENERGY STORED IN A CRYSTAL INDUCED BY SCREW DISLOCATIONS: FROM DISCRETE TO CONTINUOUS*

MARCELLO PONSIGLIONE†

Abstract. This paper deals with the passage from discrete to continuous in modeling the static elastic properties of vertical screw dislocations in a cylindrical crystal, in the setting of antiplanar linear elasticity. We study, in the framework of Γ-convergence, the asymptotic behavior of the elastic stored energy induced by dislocations as the atomic scale ε tends to zero, in the regime of dilute dislocations, i.e., rescaling the energy functionals by $1/\varepsilon^2|\log\varepsilon|$. First we consider a continuum model, where the atomic scale is introduced as an internal scale, usually called core radius. Then we focus on a purely discrete model. In both cases, we prove that the asymptotic elastic energy as $\varepsilon \to 0$ is essentially given by the number of dislocations present in the crystal. More precisely the energy per unit volume is proportional to the length of the dislocation lines, so that our result recovers in the limit as $\varepsilon \to 0$, a line tension model.

Key words. crystals, analysis of microstructure, stress concentration, calculus of variations

AMS subject classifications. 74N05, 74N15, 74G70, 74G65, 74C15, 74B15, 74B10

DOI. 10.1137/060657054

1. Introduction.

This paper deals with energy minimization methods to model static elastic properties of *dislocations* in crystals. We are interested in the asymptotic behavior of the elastic energy stored in a crystal, induced by a configuration of dislocations, as the atomic scale tends to zero. Our approach is completely variational and is based on Γ-*convergence*. First we consider a continuum model, where the atomic scale is introduced as an internal scale, usually called the *core radius*. Then we focus on a purely discrete model.

We consider the setting of *antiplanar* linear elasticity, so that all the physical quantities involved in our model will be defined on a domain $\Omega \subset \mathbb{R}^2$, which represents an horizontal section of an infinite cylindrical crystal. The elastic energy associated with a vertical *displacement* $u : \Omega \to \mathbb{R}$, in absence of dislocations, is given by[1]

$$E(\nabla u) := \int_\Omega |\nabla u(x)|^2\, dx.$$

Now we assume that vertical *screw dislocations* are present in the crystal. To model the presence of dislocations we follow the general theory of eigenstrains,[2] namely, to any dislocation corresponds a preexisting strain in the reference configuration. In this framework a configuration of screw dislocations in the crystal can be represented by a measure on Ω which is a finite sum of *Dirac masses* of the type $\mu := \sum_i z_i|\mathbf{b}|\,\delta_{x_i}$. Here x_i's represent the intersection of the dislocation lines with Ω, $\mathbf{b}$ is the so-called *Burgers vector*, which in this antiplanar setting is a vertical fixed vector whose modulus depends on the specific crystal lattice, and $z_i \in \mathbb{Z}$ represents the multiplicity of the dislocations. The class of admissible strains associated with a

*Received by the editors April 12, 2006; accepted for publication (in revised form) October 30, 2006; published electronically June 19, 2007.
http://www.siam.org/journals/sima/39-2/65705.html

†Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany (ponsigli@mat.uniroma1.it).

[1]For simplicity we will assume the *shear modulus* of the crystal is equal to 1.

[2]We refer the reader to [13], [14] for an exhaustive treatment of the subject.

dislocation μ is given by the fields whose circulation around the dislocations x_i are equal to $z_i|\mathbf{b}|$. These fields by definition have a singularity at each x_i and are not in $L^2(\Omega;\mathbb{R}^2)$. To set up a variational formulation it is then convenient to introduce an internal scale ε called the core radius, which is comparable to the atomic scale, and to remove balls of radius ε around each point of singularity x_i. More precisely to any admissible strain ψ we associate the elastic energy

$$E_\varepsilon(\psi) := \int_{\Omega_\varepsilon(\mu)} |\psi(x)|^2\, dx,$$

where $\Omega_\varepsilon(\mu) := \Omega \setminus \cup_i \overline{B}_\varepsilon(x_i)$. Given a dislocation μ, the elastic energy induced by μ, in the absence of external forces, is given by minimizing $E_\varepsilon(\psi)$ among all admissible strains.

This variational formulation has been recently considered in [4] to study the limit of the elastic energy induced by a fixed configuration of dislocations as the atomic scale ε tends to zero. The authors prove in particular that the energy[3] is of the order $|\log\varepsilon|$.

In this paper we study the asymptotic behavior of the elastic energy induced by the dislocations in terms of Γ-convergence, in this regime of energies, i.e., rescaling the energy functionals by $|\log\varepsilon|$, without assuming the dislocations to be fixed, uniformly bounded in mass, or *well separated*. Let us describe our continuum model in more detail.

Given a dislocation μ, the class of admissible strains $\mathcal{AS}_\varepsilon(\mu)$ associated with μ is given (we consider for simplicity $|\mathbf{b}| = 1$) by

$$\mathcal{AS}_\varepsilon(\mu) := \Bigg\{ \psi \in L^2(\Omega_\varepsilon(\mu);\mathbb{R}^2) : \operatorname{curl}\psi = 0 \text{ in } \Omega_\varepsilon(\mu) \text{ in the sense of distributions,} \\ \int_{\partial A} \psi(s)\cdot\tau(s)\, ds = \mu(A) \\ \text{for every open set } A \subset \Omega \text{ with } \partial A \text{ smooth and with } \partial A \subset \Omega_\varepsilon(\mu) \Bigg\}.$$

Here $\tau(s)$ is the oriented tangent vector to ∂A at the point s, and the integrand $\psi(s)\cdot\tau(s)$ is intended in the sense of traces (see Theorem 2, p. 204, of [6]). The (rescaled) elastic energy associated with μ is given by

$$\mathcal{F}_\varepsilon(\mu) := \frac{1}{|\log\varepsilon|}\Big(\min_{\psi\in\mathcal{AS}_\varepsilon(\mu)} E_\varepsilon(\psi) + |\mu|(\Omega)\Big). \tag{1.1}$$

The first term in the energy represents the elastic energy far from the dislocations, where the crystal is assumed to have a linear hyperelastic behavior (see Remark 2.6 for a partial justification of the use of linear elasticity in this region far from dislocations). The second term, $|\mu|(\Omega)$, is the *total variation* of μ on Ω and represents the elastic energy stored in the region surrounding the dislocations (the introduction of this energy in the continuum model will be fully justified by our discrete model; see Remarks 2.6 and 3.2 for more details).

[3]Though the Burgers vector should be rescaled by ε, in this and in the following results the Burgers vector is kept fixed. The relevant physical case can be recovered simply by introducing a supplementary rescaling term of the order $1/\varepsilon^2$ in the energy functionals.

In Theorem 2.4 we prove that the Γ-limit of the functionals $\mathcal{F}_\varepsilon$, with respect to the *flat convergence* of the dislocations (see (2.2)), is given by the functional $\mathcal{F}$ defined by

$$\mathcal{F}(\mu) := \frac{1}{2\pi}|\mu|(\Omega). \tag{1.2}$$

The asymptotic elastic energy per unit volume is essentially proportional to the number (and hence to the length) of the screw dislocations. Then we recover in the limit as $\varepsilon \to 0$ a *line tension model*.

A similar result was obtained in [8], [9], where the authors considered a phase field model for dislocations proposed by [11]. They study the asymptotic behavior, in different rescaling regimes, of the elastic energy given by the interaction of a nonlocal $H^{1/2}$ elastic energy, a nonlinear *Peierls potential*, and a *pinning condition* under the assumption that only one *slip system* is active. In particular, in the energy regime corresponding to a rescaling of the order $1/|\log\varepsilon|$, their Γ-limit is given by the sum of a bulk term, taking into account the pinning condition, and a surface term concentrated on the dislocation lines.

More in general, energy concentration phenomena as a result of the logarithmic rescaling are nowadays classical in the theory of *Ginzburg–Landau*-type functionals, to model vortices in superfluidity and superconductivity. We refer the reader to [3], [15], [10], [1], and to the references therein.

Even if we do not assume the dislocations to be fixed, our analysis shows that, as $\varepsilon_n \to 0$, the most convenient way to approximate a dislocation μ with multiplicity $z_i \equiv 1$, is the constant sequence $\mu_n \equiv \mu$. In this respect the main point is that there is no homogenization process able to approach an energy less than $1/2\pi \liminf_{\varepsilon_n} |\mu_n(\Omega)|$. The latter term can be interpreted as the quantity usually referred to as *geometrically necessary dislocations*. We conclude that in this energy regime there is no energetic advantage for the crystal to create micropatterns of dislocations.

These considerations become trivial if one assumes a priori a uniform bound for the number of dislocations. However, sequences $\{\mu_n\}$ with uniformly bounded energy (i.e., such that $\mathcal{F}_{\varepsilon_n}(\mu_n) \le C$) are not in general bounded in mass. The main reason is that one can easily construct a *short dipole* $\mu_n := \delta_{x_n} - \delta_{y_n}$, with $|\mu_n|(\Omega) = 2$, $|x_n - y_n| \to 0$, and whose energetic contribution is vanishing. On the other hand, it is clear that the flat norm of these dipoles is also vanishing. This is the reason why we study the Γ-convergence with respect to the flat convergence instead of the weak convergence of measures. We prove that the *equicoercivity* property holds with respect to the flat convergence: sequences μ_n with uniformly bounded energy, up to a subsequence, converge with respect to the flat norm. The proof of this result represents the main difficulty in our analysis.

Our strategy is to divide the dislocations in clusters such that in each cluster the distance between the dislocations is of order ε_n^δ for some $0 < \delta < 1$. The family of clusters with *zero effective multiplicity*, namely, such that the sum of the multiplicities in the cluster is equal to 0, will play the role of short dipoles. Using the estimate $|\mu_n|(\Omega) \le E|\log\varepsilon_n|$, which follows directly from $\mathcal{F}_{\varepsilon_n}(\mu_n) \le E$ and from the second term in (1.1), we deduce that these clusters give a vanishing contribution to the flat norm, of order $|\log\varepsilon_n|^2\varepsilon_n^\delta \to 0$. We identify the remaining clusters (with nonzero effective multiplicity) with Dirac masses, obtaining a sequence of measures $\tilde{\mu}_n := \sum_i z_i\delta_{x_i}$. Assume for a while that $\tilde{\mu}_n$ is uniformly bounded in mass so that (up to a subsequence) $\tilde{\mu}_n$ weakly converges to a measure μ. We prove that $\mu_n - \tilde{\mu}_n$ has vanishing flat norm, and we deduce the convergence of μ_n to μ with respect to the flat norm.

The main point in the previous argument is that $\tilde{\mu}_n$ is uniformly bounded in mass. This will be a consequence of the key Lemma 2.5, where we prove that each cluster with nonzero effective multiplicity gives a positive energetic contribution. It is in this step that we have to prevent the possibility of a homogenization process, able to approach a vanishing energy through a sequence μ_n with nonzero geometrically necessary dislocations. This analysis will be performed through an iterative process, which will require the introduction of several meso-scales. The choice of the number of meso-scales involved in this analysis as $\varepsilon_n \to 0$ will play a fundamental role in our proof.

The last part of this paper is devoted to a purely discrete model. We consider the illustrative case of a *square lattice* of size ε with *nearest-neighbor interactions*, following along the lines of the more general theory introduced in [2].

In this framework a displacement u is a function defined on the set Ω^0_ε of points of the lattice; the strains β are defined on the bonds of the lattice, i.e., on the class Ω^1_ε of the oriented segments of the square lattice. Finally a dislocation is represented by an integer function α defined on the class Ω^2_ε of the oriented squares of the lattice.

The class of admissible strains associated with a dislocation α is given by the strains $\xi : \Omega^1_\varepsilon \to \mathbb{R}$ satisfying

$$\mathbf{d}\,\xi = \alpha, \tag{1.3}$$

where the operator $\mathbf{d}$ is defined in (3.3). The condition expressed in (1.3) means that, for every $Q \in \Omega^2_\varepsilon$, the discrete circulation of ξ on ∂Q is equal to $\alpha(Q)$. The rescaled elastic energy induced by α is given by

$$\mathcal{F}^d_\varepsilon(\alpha) := \frac{1}{|\log \varepsilon|} \min_{\xi:\Omega^1_\varepsilon \to \mathbb{R}:\, \mathbf{d}\,\xi = \alpha} E^d_\varepsilon(\xi), \tag{1.4}$$

where the discrete elastic energy $E^d_\varepsilon(\xi)$ is defined in (3.2).

Every dislocation $\alpha : \Omega^2_\varepsilon \to \mathbb{Z}$ is induced by a function $\beta : \Omega^1_\varepsilon \to \mathbb{Z}$ defined on the bonds of the lattice such that $\mathbf{d}\,\beta = \alpha$. The class of admissible strains can then be written in the equivalent form

$$\{\beta + \mathbf{d}\,u,\ u : \Omega^0_\varepsilon \to \mathbb{R}\},$$

where $\mathbf{d}\,u : \Omega^1_\varepsilon \to \mathbb{R}$ is now the discrete gradient of u defined in (3.1). In this respect β can be interpreted as a discrete eigenstrain inducing the dislocation α. If $\alpha = \mathbf{d}\,\beta = 0$, then β is a compatible strain, i.e, $\beta = \mathbf{d}\,v$ for some displacement v, and the associated stored energy is equal to 0. Therefore α measures the degree of incompatibility of the eigenstrain β.

In Theorem 3.4 we restate our Γ-convergence result given in Theorem 2.4 in this discrete setting. The proof can be obtained as an immediate consequence of the results achieved in the continuum model, introducing an interpolation procedure with suitable commutative properties with respect to the chains $u \xrightarrow{\mathbf{d}} \xi \xrightarrow{\mathbf{d}} \alpha$ and $u \xrightarrow{\nabla} \psi \xrightarrow{\mathrm{curl}} \mu$ (see Proposition 3.3).

In the discrete model the behavior of the elastic stored energy is controlled by the lattice size ε, and it is not necessary (see Remark 3.2) to introduce a supplementary internal scale, as the core radius in the continuum case, to divide the stored elastic energy into two contributions, one concentrated in a region surrounding the dislocations and the other one far away. In this respect the discrete model seems very natural and provides a theoretical justification of the continuum model.

2. The continuum model. Here we introduce our continuum model for vertical screw dislocations in an infinite cylindrical crystal, in the setting of antiplanar elasticty. We will study, in terms of Γ-convergence, the asymptotic behavior of the elastic stored energy in the crystal induced by the screw dislocations as the atomic internal scale ε tends to 0. For the definition and the basic properties of Γ-convergence, we refer the reader to [5].

2.1. Description of the continuum model. In this section we introduce the space of *screw dislocations* X and the elastic energy functionals $\mathcal{E}_\varepsilon : X \to \mathbb{R}$. We are in the setting of antiplanar elasticty, so that the physical quantities involved in the model will be defined on an horizontal section $\Omega \subset \mathbb{R}^2$ of the infinite cylindrical crystal.

2.1.1. The space X of screw dislocations. Let Ω be an open bounded subset of $\mathbb{R}^2$. For any $x \in \Omega$ we denote by δ_x the Dirac mass centered at x. Let us denote by $\mathcal{M}(\Omega)$ the class of Radon measures on Ω. The space of screw dislocations X is given by

$$X := \left\{ \mu \in \mathcal{M}(\Omega) : \mu = \sum_{i=1}^{M} z_i \delta_{x_i},\ M \in \mathbb{N},\ x_i \in \Omega,\ z_i \in \mathbb{Z} \right\}. \tag{2.1}$$

The support set of μ defined by $\mathrm{supp}(\mu) := \{x_1, \dots, x_M\}$ represents the set where the dislocations are present, while the leading coefficients z_i in (2.1) are the *multiplicities* of the dislocations at the points x_i's. We endow X with the *flat norm* [4] $\|\mu\|_f$ defined by

$$\|\mu\|_f = \inf\{|S|,\ S \in \mathcal{S} : \partial S \llcorner \Omega = \mu\} \qquad \text{for every } \mu \in X. \tag{2.2}$$

Here $\mathcal{S}$ denotes the family of the finite formal sum of oriented segments L_i in $\overline{\Omega}$, with extreme points p_i and q_i and with integer multiplicity m_i; the mass of $S = \sum_{i=1}^{M} m_i L_i$ is given by

$$|S| := \sum_{i=1}^{M} |m_i||L_i| = \sum_{i=1}^{M} |m_i||q_i - p_i|,$$

and ∂S is defined by

$$\partial S := \sum m_i (\delta_{q_i} - \delta_{p_i}). \tag{2.3}$$

We will denote by $\mu_n \xrightarrow{f} \mu$ the convergence of μ_n to μ with respect to the flat norm.

Remark 2.1. Note that for every $\mu \in X$ we can find $S \in \mathcal{S}$ such that $\partial S \llcorner \Omega = \mu$. By linearity it is enough to consider the case $\mu = \delta_x$ for some $x \in \Omega$. Let $y \in \partial\Omega$ be the point of minimal distance from x, and let S be the segment joining y to x. Clearly we have $S \in \mathcal{S}$, $\partial S = \delta_x - \delta_y$ so that $\partial S \llcorner \Omega = \delta_x$. Moreover $\|\delta_x\|_f = |S| = \mathrm{dist}(x, \partial\Omega)$. In fact by definition $\|\mu\|_f \le |S| = \mathrm{dist}(x, \partial\Omega)$. To prove the opposite inequality it is enough to check that (by triangular inequality) any $S \in \mathcal{S}$ with $\partial S = \delta_x$ satisfies $|S| \ge \mathrm{dist}(x, \partial\Omega)$.

[4] Here we are adapting the classical definition of the flat norm to our context of Dirac masses confined in an open bounded set. For the canonical definition of the flat norm and its main properties, we refer the reader to [7], [12].

2.1.2. Admissible strains. Let us fix $\varepsilon > 0$. Given $\mu \in X$, we denote

$$\Omega_\varepsilon(\mu) := \Omega \setminus \bigcup_{x_i \in \operatorname{supp}(\mu)} \overline{B}_\varepsilon(x_i),$$

where $B_\varepsilon(x_i)$ denotes the open ball of center x_i and radius ε.

The class $\mathcal{AS}_\varepsilon(\mu)$ of *admissible strains* associated with μ is given by

$$\text{(2.4)} \quad \mathcal{AS}_\varepsilon(\mu) := \Big\{ \psi \in L^2(\Omega_\varepsilon(\mu); \mathbb{R}^2) : \operatorname{curl} \psi = 0 \text{ in } \Omega_\varepsilon(\mu) \text{ in the sense of distributions, } \int_{\partial A} \psi(s) \cdot \tau(s)\, ds = \mu(A) \text{ for every open set } A \subset \Omega \text{ with } \partial A \text{ smooth and with } \partial A \subset \Omega_\varepsilon(\mu) \Big\}.$$

Here $\tau(s)$ is the oriented tangent vector to ∂A at the point s, and the integrand $\psi(s) \cdot \tau(s)$ is intended in the sense of traces (see Theorem 2, p. 204, of [6]).

Remark 2.2. Let $\psi \in \mathcal{AS}_\varepsilon(\mu)$. By the definition (2.4), we have in particular that the circulation of ψ along ∂A is equal to 0 for every $A \subset \Omega_\varepsilon(\mu)$, which is consistent with $\operatorname{curl} \psi = 0$ in $\Omega_\varepsilon(\mu)$ in the sense of distributions. Note also that to define $\Omega_\varepsilon(\mu)$ we do not require that the balls $B_\varepsilon(x_i)$ are contained in Ω. However, only the balls compactly contained in Ω give a contribution to the circulation of the admissible fields in (2.4).

2.1.3. The elastic energy. The elastic energy associated with a strain $\psi \in \mathcal{AS}_\varepsilon(\mu)$ is given by

$$E_\varepsilon(\psi) := \|\psi(x)\|^2_{L^2(\Omega_\varepsilon(\mu);\mathbb{R}^2)}.$$

The elastic energy functional $\mathcal{E}_\varepsilon : X \to \mathbb{R}$ is defined by

$$\text{(2.5)} \qquad \mathcal{E}_\varepsilon(\mu) := \min_{\psi \in \mathcal{AS}_\varepsilon(\mu)} E_\varepsilon(\psi) + |\mu|(\Omega) \qquad \text{for every } \mu \in X.$$

The first contribution to the total energy represents the elastic energy stored in a region far from the dislocations. The second contribution to the total energy is the *total variation* of μ on Ω and represents the so-called *core energy*, namely, the energy stored in the balls $B_\varepsilon(x_i)$ (see Remarks 2.6 and 3.2 for some comment on the core energy in this model).

Remark 2.3. Note that the minimum problem in (2.5) is well posed. In fact, following the *direct method of calculus of variations*, let ψ_h be a minimizing sequence. We have that $\|\psi_h\|_{L^2(\Omega_\varepsilon(\mu);\mathbb{R}^2)} \le C$ for some positive constant C. Therefore (up to a subsequence) $\psi_h \rightharpoonup \psi$ for some $\psi \in L^2(\Omega_\varepsilon(\mu); \mathbb{R}^2)$. Moreover (see Theorem 2, p. 204, of [6]) we have $\psi \in \mathcal{AS}_\varepsilon(\mu)$. By the fact that the L^2 norm is lower semicontinuous with respect to the weak convergence, we deduce that ψ is a minimum point.

2.2. The Γ-convergence result. In this section we study the asymptotic behavior, as $\varepsilon \to 0$, of the elastic energy functionals $\mathcal{E}_\varepsilon$ defined in (2.5) in terms of Γ-convergence. To this aim let us rescale the functionals $\mathcal{E}_\varepsilon$ setting

$$\text{(2.6)} \qquad \mathcal{F}_\varepsilon := \frac{1}{|\log \varepsilon|} \mathcal{E}_\varepsilon,$$

and let us introduce the candidate Γ-limit $\mathcal{F} : X \to \mathbb{R}$ defined by

$$\mathcal{F}(\mu) := \frac{1}{2\pi}|\mu|(\Omega) \qquad \text{for every } \mu \in X. \tag{2.7}$$

THEOREM 2.4. *The following* Γ*-convergence result holds.*

(i) *Equicoercivity: Let* $\varepsilon_n \to 0$*, and let* $\{\mu_n\}$ *be a sequence in* X *such that* $\mathcal{F}_{\varepsilon_n}(\mu_n) \le E$ *for some positive constant* E *independent of* n*. Then (up to a subsequence)* $\mu_n \xrightarrow{f} \mu$ *for some* $\mu \in X$*.*

(ii) Γ*-convergence: The functionals* $\mathcal{F}_{\varepsilon_n}$ Γ*-converge to* $\mathcal{F}$ *as* $\varepsilon_n \to 0$ *with respect to the flat norm, i.e., the following inequalities hold.*

Γ*-liminf inequality:* $\mathcal{F}(\mu) \le \liminf \mathcal{F}_{\varepsilon_n}(\mu_n)$ *for every* $\mu \in X$*,* $\mu_n \xrightarrow{f} \mu$ *in* X*.*

Γ*-limsup inequality: Given* $\mu \in X$*, there exists* $\{\mu_n\} \subset X$ *with* $\mu_n \xrightarrow{f} \mu$ *such that* $\limsup \mathcal{F}_{\varepsilon_n}(\mu_n) \le \mathcal{F}(\mu)$*.*

2.2.1. Equicoercivity. The proof of the equicoercivity property is quite technical and requires some preliminary result. Before giving the rigorous proof, let us recall the main steps of our strategy.

The first step is to divide the dislocations in clusters of size $r = \varepsilon_n^\delta$ for some $0 < \delta < 1$. To this aim, let us set

$$A_r(\mu_n) := \bigcup_{x \in \operatorname{supp}(\mu_n)} B_r(x). \tag{2.8}$$

Each connected component of $A_r(\mu_n)$ represents a cluster of dislocations. By construction the distance between the dislocations belonging to the same cluster is of the order $r = \varepsilon_n^\delta$. The main point is that the family of clusters of dislocations with zero *effective multiplicity*, i.e., such that the sum of the multiplicities of the dislocations in each of these clusters is equal to 0, gives a vanishing contribution to the flat norm, while the number of the remaining clusters with nonzero effective multiplicity is uniformly bounded. This latter fact is more delicate and will be done in the following key lemma, which states that each cluster with nonzero effective multiplicity gives a positive energetic contribution.

LEMMA 2.5. *Let* $0 < \delta < 1$ *be fixed. Let* $\varepsilon_n \to 0$*, and let* $\{\mu_n\}$ *be a sequence such that* $\mathcal{F}_{\varepsilon_n}(\mu_n) \le E$ *for some positive* E *independent of* n*. Moreover assume that for every* n *there exists a connected component* C_n *of* $A_{\varepsilon_n^\delta}(\mu_n)$ *(defined according to* (2.8)*) with* $C_n \subset \Omega$ *and* $\mu_n(C_n) \ne 0$*. Then*

$$\liminf_n \frac{1}{|\log \varepsilon_n|} \int_{C_n} |\psi_n(x)|^2 \, dx \ge \frac{1}{2\pi}(1-\delta) \quad \textit{for every sequence } \{\psi_n\} \subset \mathcal{AS}_{\varepsilon_n}(\mu_n).$$

Before giving the formal proof of the lemma, let us explain its main ideas. Let C_n be a cluster of dislocations with effective multiplicity equal to $\lambda \ne 0$, and let γ be a closed curve surrounding C_n, which does not intersect any other cluster of dislocations. Then the circulation of every admissible strain $\psi_n \in \mathcal{AS}_{\varepsilon_n}(\mu_n)$ on γ is equal to λ. We get an estimate of the tangential component of ψ_n on γ and, hence, of the L^2 norm of ψ_n on γ. Extending this estimate on an annular neighborhood F_n of the cluster C_n, by means of polar coordinates, we want to obtain an estimate of the elastic energy stored around C_n independently of ε_n. However, the rigorous proof will require some additional effort. The main obstruction to the previous argument is that, in general, F_n may intersect other clusters of dislocations. Our strategy is then to iterate the previous construction in subclusters of C_n. We consider a certain number of exponents $0 < \delta = s_n^0 < s_n^1 < \cdots s_n^{M_n} \leq 1$, where $M_n \to \infty$ as $\varepsilon_n \to 0$. For *almost every scale* s_n^i we find a subcluster of C_n of size $\varepsilon^{s_n^i}$ with nonzero effective multiplicity, surrounded by some annulus F_n^i, such that the sets F_n^i are pairwise disjoint and the elastic energy stored in each F_n^i is of the order $s_n^i - s_n^{i-1}$. We deduce that the elastic energy stored in C_n is at least of the order $1 - \delta$.

The starting point of our analysis is the following estimate, which easily follows by the second term in (2.5) and by (2.6):

$$\sharp\, \mathrm{supp}(\mu_n) \leq E|\log \varepsilon_n| \qquad \text{for every } n \in \mathbb{N}, \tag{2.9}$$

where $\sharp\, \mathrm{supp}(\mu_n)$ denotes the number of elements in $\mathrm{supp}(\mu_n)$.

Proof of Lemma 2.5. Let us divide the proof of the lemma into four steps.

Step 1. In this Step we introduce the exponents $0 < \delta = s_n^0 < s_n^1 < \cdots s_n^{M_n} \leq 1$, and we select the meso-scales $\varepsilon_n^{s_n^i}$ which will be involved in our analysis.

For every n, let us set

$$s_n^i := \delta + \frac{i}{H \log|\log(\varepsilon_n)|} \qquad \text{for every } 0 \leq i \leq M_n,$$

where $H > 0$ is a fixed positive constant and M_n is the integer part of $H(1 - \delta)\log|\log(\varepsilon_n)|$. For every n and for every $0 \leq i \leq M_n$, let us set

$$A_n^i := A_{\varepsilon_n^{s_n^i}}(\mu_n) \cap C_n = \bigcup_{x \in \mathrm{supp}(\mu_n) \cap C_n} B_{\varepsilon_n^{s_n^i}}(x).$$

Let $\mathcal{C}_n^i$ be the family of the connected components $C_n^{i,j}$ of A_n^i with $\mu_n(C_n^{i,j}) \neq 0$. Let us now split the indices $i \in [0, M_n)$ into two families J_n and I_n by setting $i \in J_n$ if every element in $\mathcal{C}_n^i$ contains at least two elements of $\mathcal{C}_n^{i+1}$; $i \in I_n$ otherwise. Let us prove that

$$\liminf_n \frac{\sharp\, I_n}{M_n} = 1 - o(1/H), \tag{2.10}$$

where $\sharp E$ denotes the number of elements of a set E and $o(1/H) \to 0$ as $H \to \infty$. To this aim, note that if $i \in J_n$, then $\sharp\, \mathcal{C}_n^{i+1} \geq 2 \sharp\, \mathcal{C}_n^i$, and hence, using that $\sharp\, \mathcal{C}_n^i$ is nondecreasing with respect to i and recalling that by assumption $\sharp\, \mathcal{C}_n^0 = \sharp\, \{C_n\} = 1$, we have

$$\sharp\, \mathrm{supp}(\mu_n) \geq \sharp\, \mathcal{C}_n^{M_n} \geq 2^{\sharp\, J_n} \sharp\, \mathcal{C}_n^0 = 2^{\sharp\, J_n}.$$

By (2.9) we obtain

$$E|\log \varepsilon_n| \geq \sharp\, \mathrm{supp}(\mu_n) \geq 2^{\sharp\, J_n}.$$

Therefore $\sharp\, J_n \leq C \log|\log(\varepsilon_n)|$ for some positive constant C independent of H. We deduce that

$$\limsup_n \frac{\sharp\, J_n}{M_n} \leq \limsup_n \frac{C \log|\log(\varepsilon_n)|}{H(1-\delta)\log|\log(\varepsilon_n)| - 1} = \frac{C}{H(1-\delta)},$$

which together with $\sharp\, J_n + \sharp\, I_n \equiv M_n$ gives (2.10).

Step 2. In this step we define the annular sets F_n^i.

Let $i \in I_n$. By definition there exists C_n^{i,j_i} in $\mathcal{C}_n^i$ which contains exactly one element $C_n^{i+1,k_{i+1}}$ in $\mathcal{C}_n^{i+1}$. Let $p_n^i \in C_n^{i+1,k_{i+1}}$ be chosen arbitrarily. Let us define

$$R_1^i := E|\log\varepsilon_n|\varepsilon_n^{s_n^{i+1}}, \qquad R_2^i := \frac{1}{2}\varepsilon_n^{s_n^i}.$$

Moreover, for every connected component $C_n^{i+1,l}$ of A_n^{i+1} different from $C_n^{i+1,k_{i+1}}$, let us set

$$r_1^l := \min_{x\in C_n^{i+1,l}} |x-p_n^i| - \varepsilon_n^{s_n^{i+1}}, \qquad r_2^l := \max_{x\in C_n^{i+1,l}} |x-p_n^i| + \varepsilon_n^{s_n^{i+1}}. \tag{2.11}$$

Let us set

$$L_n^i := (R_1^i, R_2^i)\setminus \bigcup_l \overline{(r_1^l, r_2^l)}, \qquad F_n^i := \{x\in\mathbb{R}^2 : |x-p_n^i| \in L_n^i\}. \tag{2.12}$$

The set L_n^i is a finite union of open intervals, and hence it can be written in the form

$$L_n^i = \bigcup_{l=1}^{N_n^i} (\varepsilon_n^{\alpha_{i,n}^l}, \varepsilon_n^{\beta_{i,n}^l}). \tag{2.13}$$

Let us claim the following properties concerning L_n^i and F_n^i.

(a) For every i, n and for every $r \in L_n^i$, we have $\mu_n(B_r(p_n^i)) \neq 0$.

(b) For every n the sets F_n^i are pairwise disjoint.

(c) For every i we have $\sum_{l=1}^{N_n^i} |\alpha_{i,n}^l - \beta_{i,n}^l| = (1+o(1/n))/\big(H\log|\log(\varepsilon_n)|\big)$, where $o(1/n)$ is a function independent of i tending to 0 as $n\to\infty$.

Step 3. Let $\psi_n \in \mathcal{AS}_{\varepsilon_n}(\mu_n)$. Using the claim, we are in position to estimate the L^2 norm of ψ_n on each F_n^i using polar coordinates.

By property (a) and by the fact that ψ_n is an admissible strain (see (2.4)), we have

$$\int_{(0,2\pi)} |\psi_n(r,\theta)| r\,d\theta \geq |\mu_n(B_r(p_n^i))| \geq 1 \qquad \text{for every } r\in L_n^i.$$

Using Jensen's inequality and property (c) above, we deduce the following estimate:

$$\begin{aligned}
\int_{F_n^i} |\psi_n(x)|^2\,dx &= 2\pi \sum_{l=1}^{N_n^i} \int_{(\varepsilon_n^{\alpha_{i,n}^l},\varepsilon_n^{\beta_{i,n}^l})} r\left(\frac{1}{2\pi}\int_{(0,2\pi)} |\psi_n(r,\theta)|^2\,d\theta\right) dr\\
&\geq 2\pi \sum_{l=1}^{N_n^i} \int_{(\varepsilon_n^{\alpha_{i,n}^l},\varepsilon_n^{\beta_{i,n}^l})} r\left(\frac{1}{2\pi}\int_{(0,2\pi)} |\psi_n(r,\theta)|\,d\theta\right)^2 dr \geq \frac{1}{2\pi}\sum_{l=1}^{N_n^i} \int_{(\varepsilon_n^{\alpha_{i,n}^l},\varepsilon_n^{\beta_{i,n}^l})} \frac{1}{r}dr\\
&= \frac{1}{2\pi}\sum_{l=1}^{N_n^i} |\alpha_{i,n}^l - \beta_{i,n}^l||\log\varepsilon_n| = \frac{1}{2\pi}\frac{1+o(1/n)}{H\log|\log(\varepsilon_n)|}|\log\varepsilon_n|.
\end{aligned}$$

Summing the previous inequality over all $i\in I_n$ and dividing by $|\log\varepsilon_n|$, in view of (2.10) and property (b) above, we deduce

$$\begin{aligned}
\liminf_n \frac{1}{|\log\varepsilon_n|}\int_{C_n} |\psi_n(x)|^2\,dx &\geq \liminf_n \frac{1}{|\log\varepsilon_n|}\sum_{i\in I_n}\int_{F_n^i} |\psi_n(x)|^2\,dx\\
&\geq \liminf_n \frac{1}{2\pi}\frac{\sharp I_n(1+o(1/n))}{H\log|\log(\varepsilon_n)|} = \liminf_n \frac{1}{2\pi}\frac{M_n(1-o(1/H))}{H\log|\log(\varepsilon_n)|}.
\end{aligned}$$

Letting $H \to \infty$ and recalling that M_n is the integer part of $H(1-\delta)\log|\log(\varepsilon_n)|$, we obtain

(2.14)
$$\liminf_n \frac{1}{|\log\varepsilon_n|}\int_{C_n}|\psi_n(x)|^2\,dx \geq \liminf_n \frac{1}{2\pi}\frac{H(1-\delta)\log|\log(\varepsilon_n)|-1}{H\log|\log(\varepsilon_n)|} = \frac{1}{2\pi}(1-\delta).$$

Step 4. In order to conclude the proof, we have to prove the claim.

Property (a) follows directly by the construction of the intervals L^i_n. More precisely one can easily check that by construction every cluster of dislocations $C^{i+1,l}_n \in A^{i+1}_n$ intersecting $B_r(p^i_n)$ is actually contained in $B_r(p^i_n)$, and since $i \in I$, then all these clusters except one have zero effective multiplicity.

Let us pass to the proof of property (b). Let i_1, $i_2 \in I_n$, with $i_1 < i_2$. We will use the notation used in the previous constructions, with i replaced by i_1, i_2, respectively. In particular let $p^{i_1}_n \in C^{i_1+1,k_{i_1+1}}_n$, $p^{i_2}_n \in C^{i_2+1,k_{i_2+1}}_n$. We divide the proof into two cases.

In the first case we assume that $p^{i_2}_n \in C^{i_1+1,k_{i_1+1}}_n$. In this case, since $|p^{i_1}_n - p^{i_2}_n| \leq \varepsilon_n^{s^{i_1+1}_n}$, we have

$$R^{i_2}_2(p^{i_2}_n) + |p^{i_1}_n - p^{i_2}_n| = \frac{1}{2}\varepsilon_n^{s^{i_2}_n} + |p^{i_1}_n - p^{i_2}_n| \leq 2\varepsilon_n^{s^{i_1+1}_n} < E|\log\varepsilon_n|\varepsilon_n^{s^{i_1+1}_n} = R^{i_1}_1(p^{i_1}_n),$$

and hence $B_{R^{i_2}_2}(p^{i_2}_n) \subset B_{R^{i_1}_1}(p^{i_1}_n)$ so that $F^{i_1}_n$ and $F^{i_2}_n$ are disjoint.

Let us consider now the case $p^{i_2}_n \notin C^{i_1+1,k_{i_1+1}}_n$. In this case we have $p^{i_2}_n \in C^{i_1+1,l}_n$ for some connected component $C^{i_1+1,l}_n$ of $A^{i_1+1}_n$ different from $C^{i_1+1,k_{i_1+1}}_n$. Therefore by (2.11), (2.12) we deduce that

$$B_{\varepsilon_n^{s^{i_1+1}_n}}(p^{i_2}_n) \cap F^{i_1} = \emptyset.$$

On the other hand

$$R^{i_2}_2 = \frac{1}{2}\varepsilon_n^{s^{i_2}_n} \leq \frac{1}{2}\varepsilon_n^{s^{i_1+1}_n}.$$

We deduce that

$$B_{R^{i_2}_2}(p^{i_2}_n) \cap F^{i_1} = \emptyset.$$

This concludes the proof of (b).

Let us pass to the proof of (c). For every i we have

$$\begin{aligned}|L^i_n| &\geq (R^i_2 - R^i_1) - \tilde{C}|\log\varepsilon_n|\varepsilon_n^{s^{i+1}_n} \geq \frac{1}{2}\varepsilon_n^{s^i_n} - C|\log\varepsilon_n|\varepsilon_n^{s^{i+1}_n} \\ &= \varepsilon_n^{s^i_n}\left(\frac{1}{2} - C|\log\varepsilon_n|\varepsilon_n^{(s^{i+1}_n - s^i_n)}\right),\end{aligned} \tag{2.15}$$

where C is a constant depending only on E. On the other hand, fixing the quantity

$$\sum_{l=1}^{N^i_n}|\varepsilon_n^{\alpha^l_{i,n}} - \varepsilon_n^{\beta^l_{i,n}}|$$

in (2.13) and maximizing $|L^i_n|$ with respect to the position of the indices $\alpha^l_{i,n}$, $\beta^l_{i,n}$, we obtain

$$|L^i_n| \leq \varepsilon_n^{s^i_n}\left(\frac{1}{2} - \varepsilon_n^{\sum_{l=1}^{N^i_n}|\alpha^l_{i,n} - \beta^l_{i,n}|}\right). \tag{2.16}$$

From (2.15) and (2.16) we deduce that

$$\varepsilon_n^{s_n^i}\left(\frac{1}{2}-\varepsilon_n^{\sum_{l=1}^{N_n^i}|\alpha_{i,n}^l-\beta_{i,n}^l|}\right)\geq \varepsilon_n^{s_n^i}\left(\frac{1}{2}-C|\log\varepsilon_n|\varepsilon_n^{(s_n^{i+1}-s_n^i)}\right).$$

Therefore

$$\varepsilon_n^{\sum_{l=1}^{N_n^i}|\alpha_{i,n}^l-\beta_{i,n}^l|}\leq C|\log\varepsilon_n|\varepsilon_n^{(s_n^{i+1}-s_n^i)}=C|\log\varepsilon_n|\varepsilon_n^{\frac{1}{H\log|\log(\varepsilon_n)|}},$$

from which (c) easily follows. □

We are now in a position to prove the equicoercivity. The idea is to identify the clusters of dislocations with nonzero effective multiplicity with Dirac masses, obtaining a sequence of measures $\tilde{\mu}_n := \sum_i z_i\delta_{x_i}$. By Lemma 2.5 we will deduce that $\tilde{\mu}_n$ is uniformly bounded in mass so that (up to a subsequence) $\tilde{\mu}_n$ weakly converges to a measure μ. We prove that $\mu_n - \tilde{\mu}_n$ has a vanishing flat norm, and we deduce the convergence of μ_n to μ with respect to the flat norm.

Proof of the equicoercivity property. Let $0 < S < T < 1$. For every $S < \delta < T$, let us consider the set $A_{\varepsilon_n^\delta}(\mu_n)$ defined as in (2.8). Let us denote by $\mathcal{C}_{\delta,n}$ the family of connected components $C_{\delta,n}^1,\dots,C_{\delta,n}^{M_{\delta,n}}$ of $A_{\varepsilon_n^\delta}(\mu_n)$ which are contained in Ω and satisfy $\mu(C_{\delta,n}^l)\neq 0$. By Lemma 2.5 we deduce that $\sharp\mathcal{C}_{\delta,n}=M_{\delta,n}$ is bounded by a constant M independent of n and δ. For every n, let us consider the finite family of indices

$$I_n := \{t_n^1,\dots,t_n^{M_n}\},\quad S\leq t_n^1<t_n^2<\cdots<t_n^{M_n}\leq T,\quad M_n\leq M,$$

given by the discontinuity points of the function $\delta\to\sharp\mathcal{C}_{\delta,n}$. Up to a subsequence, we have that the set of accumulation points of I_n is of the type

$$I_\infty := \{\delta^1,\dots,\delta^H\},\quad S\leq \delta^1<\delta^2<\cdots<\delta^H\leq T,\quad H\leq M.$$

Let

$$[\delta_1,\delta_2]\subset (S,T)\setminus I_\infty.$$

For n big enough we have that the function $\delta\to\sharp\mathcal{C}_{\delta,n}$ is constant on $[\delta_1,\delta_2]$. Since each element of $\mathcal{C}_{\delta_1,n}$ contains at least one element of $\mathcal{C}_{\delta_2,n}$, we deduce that each element $C_{\delta_1,n}^l\in\mathcal{C}_{\delta_1,n}$ actually contains exactly one element $C_{\delta_2,n}^l\in\mathcal{C}_{\delta_2,n}$.

We want to prove that for every sequence $\{H_n\}\subset\mathcal{C}_{\delta_1,n}$ we have

$$\limsup_n|\mu_n(H_n)|\leq K \tag{2.17}$$

for some positive constant K independent of n.

Let G_n be the only element of $\mathcal{C}_{\delta_2,n}$ contained in H_n. The idea, as in the proof of Lemma 2.5, is to evaluate the elastic energy of every admissible strain $\psi_n\in\mathcal{AS}_{\varepsilon_n}(\mu_n)$ stored in the region between G_n and H_n using polar coordinates. To this aim, let $p_n\in G_n$, and let us define

$$R_1 := E|\log\varepsilon_n|\varepsilon_n^{\delta_2},\qquad R_2:=\varepsilon_n^{\delta_1}.$$

Moreover, for every connected component C_n^l of $A_{\varepsilon_n^{\delta_2}}(\mu_n)$ different from G_n, let us set

$$r_1^l := \min_{x\in C_n^l}|x-p_n|-\varepsilon_n^{\delta_2},\qquad r_2^l := \max_{x\in C_n^l}|x-p_n|+\varepsilon_n^{\delta_2}.$$

Let us set

$$L_n := (R_1, R_2) \setminus \bigcup_l \overline{(r_1^l, r_2^l)}, \qquad F_n := \{x \in \mathbb{R}^2 : |x - p_n| \in L_n\}.$$

The set L_n is a finite union of open intervals, and hence it can be written in the form

$$L_n = \bigcup_{l=1}^{N_n} (\varepsilon_n^{\alpha_n^l}, \varepsilon_n^{\beta_n^l}). \tag{2.18}$$

Arguing as in the proof of properties (a) and (c), in the proof of Lemma 2.5, we deduce that the following properties hold.

(i) For every n and for every $r \in L_n$, we have $\mu_n(B_r(p_n)) = \mu_n(G_n)$.

(ii) $\sum_{l=1}^{N_n} |\alpha_n^l - \beta_n^l| = (1 + o(1/n))(\delta_2 - \delta_1)$, where $o(1/n) \to 0$ as $n \to \infty$.

Let $\psi_n \in \mathcal{AS}_{\varepsilon_n}(\mu_n)$. By the fact that ψ_n is an admissible strain and by property (i), we deduce that

$$\int_{(0,2\pi)} |\psi_n(r,\theta)| r d\theta \geq |\mu(G_n)| \qquad \text{for every } r \in L_n.$$

Using Jensen's inequality and property (ii) above we obtain

$$\begin{aligned}
&\int_{F_n} |\psi_n(x)|^2\, dx = 2\pi \sum_{l=1}^{N_n} \int_{(\varepsilon_n^{\alpha_n^l}, \varepsilon_n^{\beta_n^l})} r \left(\frac{1}{2\pi} \int_{(0,2\pi)} |\psi_n(r,\theta)|^2\, d\theta \right) dr \\
&\geq 2\pi \sum_{l=1}^{N_n} \int_{(\varepsilon_n^{\alpha_n^l}, \varepsilon_n^{\beta_n^l})} r \left(\frac{1}{2\pi} \int_{(0,2\pi)} |\psi_n(r,\theta)|\, d\theta \right)^2 dr \geq \frac{1}{2\pi} |\mu(G_n)|^2 \sum_{l=1}^{N_n} \int_{(\varepsilon_n^{\alpha_n^l}, \varepsilon_n^{\beta_n^l})} \frac{1}{r} dr \\
&= \frac{1}{2\pi} |\mu(G_n)|^2 \sum_{l=1}^{N_n} |\alpha_n^l - \beta_n^l| |\log \varepsilon_n| = \frac{1}{2\pi} |\mu(G_n)|^2 (1 + o(1/n))(\delta_2 - \delta_1) |\log \varepsilon_n|.
\end{aligned}$$

Dividing by $|\log \varepsilon_n|$ in the previous inequality and noticing that $\mu_n(G_n) = \mu_n(H_n)$, we deduce

$$E \geq \limsup_n \mathcal{F}_n(\mu_n) \geq \frac{1}{2\pi}(\delta_2 - \delta_1) \limsup_n (\mu_n(G_n))^2 = \frac{1}{2\pi}(\delta_2 - \delta_1) \limsup_n (\mu_n(H_n))^2, \tag{2.19}$$

and this concludes the proof of (2.17).

Now we construct the sequence $\{S_n\}$ of oriented segments in $\mathcal{S}$ (see (2.2)) of the form $S_n = F_n + N_n$ such that

$$\partial S_n \llcorner \Omega = \mu_n, \quad |\partial F_n| \leq 2MK, \quad |N_n| \to 0, \tag{2.20}$$

which is clearly enough to guarantee the compactness of the sequence μ_n.

To this aim, in every element $H_n^l \in \mathcal{C}_{\delta_1,n}$ fix a point p_n^l, and consider the measure

$$\tilde{\mu}_n := \sum_{l=1}^{\sharp \mathcal{C}_{\delta_1,n}} \mu(H_n^l)\delta_{p_n^l}.$$

We have that $|\tilde{\mu}_n| \le MK$, and hence we can find $F_n \in \mathcal{S}$ satisfying

$$|\partial F_n| \le 2MK, \quad \partial F_n \llcorner \Omega = \tilde{\mu}_n. \tag{2.21}$$

Now let us denote by $\mathcal{I}_n$ the union of the connected components of $A_{\varepsilon_n^{\delta_1}}(\mu_n)$ strictly contained in Ω and by $\mathcal{K}_n$ the union of the connected components of $A_{\varepsilon_n^{\delta_1}}(\mu_n)$ intersecting $\partial\Omega$. We clearly have

$$\operatorname{supp}(\mu_n) \subset \mathcal{I}_n \cup \mathcal{K}_n. \tag{2.22}$$

Moreover by construction $(\mu_n - \tilde{\mu}_n)(I_n^l) = 0$ for every $I_n^l \in \mathcal{I}_n$. Therefore, using that $\sharp \operatorname{supp}(\mu_n) \le E|\log\varepsilon_n|$ and that for every $I_n^l \in \mathcal{I}_n$ we have $\operatorname{diam}(I_n^l) \le E|\log\varepsilon_n|\varepsilon_n^{\delta_i}$, we can easily find $V_n \in \mathcal{S}$ such that

$$\partial V_n = (\mu_n - \tilde{\mu}_n)\llcorner \mathcal{I}_n, \qquad |V_n| \le \operatorname{diam}(I_n^l)\, \sharp \operatorname{supp}(\mu_n) \le \varepsilon_n^{\delta_1} E^2 |\log\varepsilon_n|^2. \tag{2.23}$$

On the other hand, since for every $x \in \operatorname{supp}(\mu_n) \cap \mathcal{K}_n$ we have $d(x,\partial\Omega) \le E|\log\varepsilon_n|\varepsilon_n^{\delta_1}$, we can also find $W_n \in \mathcal{S}$ (joining each $x \in \operatorname{supp}(\mu_n) \cap \mathcal{K}_n$ with a point of $\partial\Omega$) such that

$$\partial W_n \llcorner \Omega = \mu_n \llcorner \mathcal{K}_n, \qquad |W_n| \le \varepsilon_n^{\delta_1} E^2 |\log\varepsilon_n|^2. \tag{2.24}$$

Setting $N_n := V_n + W_n$, by (2.21), (2.22), (2.23), and (2.24) we deduce that (2.20) holds true. ∎

2.2.2. Γ-convergence. Here we prove the Γ-convergence result.

Proof of the Γ*-limsup inequality.* It is enough to prove the Γ-limsup inequality assuming that $|\mu(x)| = 1$ for every $x \in \operatorname{supp}(\mu)$. In fact the class of measures satisfying this assumption is dense in energy, and with respect to the flat convergence in X (more precisely, given $\delta > 0$ and $\tilde{\mu} \in X$), there exists $\mu \in X$ with $\|\mu - \tilde{\mu}\|_f \le \delta$ and $\mathcal{F}(\mu) = \mathcal{F}(\tilde{\mu})$, satisfying $|\mu(x)| = 1$ for every $x \in \operatorname{supp}(\mu)$.

The recovering sequence is given by the constant sequence $\mu_n \equiv \mu$. We have to construct a sequence of admissible strains $\psi_n \in \mathcal{AS}_\varepsilon(\mu)$ satisfying

$$\mathcal{F}(\mu) \ge \limsup \frac{1}{|\log\varepsilon_n|} \int_{\Omega_{\varepsilon_n}(\mu)} |\psi_n|^2. \tag{2.25}$$

To this aim, for every $x_i \in \operatorname{supp}(\mu) \cap \Omega$, we consider the field ψ_{x_i}, which in polar coordinates is defined by

$$\psi_{x_i}(r,\theta) := \frac{1}{2\pi r}\tau_i(r,\theta),$$

where $\tau_i(r,\theta)$ is the unit tangent vector to $\partial B_r(x_i)$ at the point with coordinates (r,θ).

The recovering sequence ψ_n is defined by

$$\psi_n := \sum_{x_i \in \operatorname{supp}(\mu)} \psi_{x_i} \llcorner \Omega_{\varepsilon_n}. \tag{2.26}$$

It can be easily proved that $\psi_n \in \mathcal{AS}_{\varepsilon_n}(\mu_n)$, and

$$\limsup \frac{1}{|\log \varepsilon_n|} \int_{\Omega_{\varepsilon_n}(\mu)} |\psi_n|^2 = \limsup \sum_{x_i \in \operatorname{supp}(\mu)} \frac{1}{|\log \varepsilon_n|} \int_{\Omega_{\varepsilon_n}(\mu)} |\psi_{x_i}|^2 = \mathcal{F}(\mu),$$

and this concludes the proof of (2.25). □

Proof of the Γ*-liminf inequality.* Let $\mu \in X$, and let $\mu_n \xrightarrow{f} \mu$ in X. We can assume that $\liminf \mathcal{F}_n(\mu_n) \le E < \infty$. Let us fix $0 < S < T < 1$, and let us consider the set $A_{\varepsilon_n^T}(\mu_n)$ defined as in (2.8). By Lemma 2.5 we deduce that there exists a finite number of connected component $C_n^1, \dots, C_n^{L_n}$ of $A_{\varepsilon_n^T}(\mu_n)$, with L_n uniformly bounded by a constant M independent of n such that $C_n^l \subset \Omega$ and $\mu_n(C_n^l) \neq 0$. Let us denote by $\sharp_n : (S,T) \to \{1, \dots M\}$ the function that counts the number of connected components of $A_{\varepsilon_n^s}(\mu_n)$ containing at least one C_n^l. For every n, let us consider the finite family of indices

$$I_n := \{t_n^1, \dots, t_n^{M_n}\}, \quad S \le t_n^1 < t_n^2 < \dots < t_n^{M_n} \le T, \quad M_n \le M,$$

given by the discontinuity points of $\sharp_n$. Up to a subsequence, we have that the set of accumulation points of I_n contained in (S,T) is of the type

$$I_\infty := \{\delta^1, \dots, \delta^H\}, \quad S < \delta^1 < \delta^2 < \dots < \delta^H < T, \quad H \le M.$$

Let us set $\delta^0 = S$, $\delta^{H+1} = T$, and for every $0 \le i \le H$ consider intervals (a^i, b^i) with $\delta^i < a^i < b^i < \delta^{i+1}$. For any fixed $\tau > 0$, we can always assume that

$$\sum_i (b^i - a^i) \ge T - S - \tau.$$

For every $s \in (a^i, b^i)$, we have exactly $\sharp_n(b^i)$ connected components $K_n^{s,1}, \dots, K_n^{s,\sharp_n(b^i)}$ of $A_{\varepsilon_n^s}(\mu_n)$ containing at least one C_n^l. For every $1 \le j \le \sharp_n(b^i)$ we arbitrarily fix a point $p_n^{i,j} \in K_n^{b^i,j}$. Let us define

$$R_1^{i,n} := E|\log \varepsilon_n| \varepsilon_n^{b^i}, \qquad R_2^{i,n} := \varepsilon_n^{a^i}.$$

Moreover, for every connected component $C_n^{i,l}$ of $A_{\varepsilon_n^{b^i}}(\mu_n)$ different from $K_n^{b^i,j}$, let us set

$$r_{1,n}^{i,j,l} := \min_{x \in C_n^{i,l}} |x - p_n^{i,j}| - \varepsilon_n^{b_i}, \qquad r_{2,n}^{i,j,l} := \max_{x \in C_n^{i,l}} |x - p_n^{i,j}| + \varepsilon_n^{b_i}.$$

Let us set

$$L_n^{i,j} := (R_1^{i,n}, R_2^{i,n}) \setminus \bigcup_l \overline{(r_{1,n}^{i,j,l}, r_{2,n}^{i,j,l})}, \qquad F_n^{i,j} := \{x \in \mathbb{R}^2 : |x - p_n^{i,j}| \in L_n^{i,j}\}.$$

The sets $L_n^{i,j}$ are a finite union of open intervals, and hence they can be written in the form

$$L_n^{i,j} = \bigcup_{l=1}^{N_n^{i,j}} (\varepsilon_n^{\alpha_n^{i,j,l}}, \varepsilon_n^{\beta_n^{i,j,l}}).$$

Let us denote by $\mathcal{H}_n^i$ the family of sets $F_n^{i,j}$ which are strictly contained in Ω. The following properties concerning $L_n^{i,j}$, $F_n^{i,j}$, and $\mathcal{H}_n^i$ can be readily verified by the reader.

(a) For every i, j, for n big enough, and for every $r \in L_n^{i,j}$, we have $\mu_n(B_r(p_n^{i,j})) \equiv \mu(F_n^{i,j})$.
(b) For n big enough, the sets $F_n^{i,j}$ are pairwise disjoint.
(c) For every i, $\mu_n \llcorner \big(\cup_{F_n^{i,j} \in \mathcal{H}_n^i} F_n^{i,j}\big) \xrightarrow{f} \mu$.
(d) For every i, j, $\sum_{l=1}^{N_n^{i,j}} |\alpha_n^{i,j,l} - \beta_n^{i,j,l}| = (1 + o(1/n))(b^i - a^i)$, where $o(1/n) \to 0$ as $n \to \infty$.

Arguing as in the proof of (2.19) we obtain that, for every $\psi_n \in \mathcal{AS}_{\varepsilon_n}(\mu_n)$ and for every $F_n^{i,j} \in \mathcal{H}_n^i$,

$$\frac{1}{|\log \varepsilon_n|} \int_{F_n^{i,j}} |\psi_n|^2 \, dx \geq \frac{1}{2\pi} \big(\mu_n(F_n^{i,j})\big)^2 (1 + o(1/n))(b^i - a^i).$$

Summing the previous inequality over all $F_n^{i,j} \in \mathcal{H}_n^i$, we obtain

$$\frac{1}{|\log \varepsilon_n|} \sum_{F_n^{i,j} \in \mathcal{H}_n^i} \int_{F_n^{i,j}} |\psi_n|^2 \, dx \geq \frac{1}{2\pi} (1 + o(1/n)) \sum_{F_n^{i,j} \in \mathcal{H}_n^i} |\mu_n(F_n^{i,j})|(b^i - a^i). \tag{2.27}$$

Recalling that the diameter of $F_n^{i,j}$ tends to 0 as $n \to \infty$, by property (c) we easily deduce that, for every fixed i,

$$\liminf_n \sum_{F_n^{i,j} \in \mathcal{H}_n^i} |\mu_n(F_n^{i,j})| \geq |\mu|(\Omega). \tag{2.28}$$

Letting $n \to \infty$ in (2.27) and using (2.28), we obtain

$$\begin{aligned}
\liminf_n \mathcal{F}(\mu_n) &\geq \liminf_n \frac{1}{|\log \varepsilon_n|} \int_{\Omega_{\varepsilon_n}(\mu)} |\psi_n|^2 \, dx \\
&\geq \liminf_n \frac{1}{|\log \varepsilon_n|} \sum_i \sum_{F_n^{i,j} \in \mathcal{H}_n^i} \int_{F_n^{i,j}} |\psi_n|^2 \, dx \\
&\geq \liminf_n \frac{1}{2\pi} \sum_i \sum_{F_n^{i,j} \in \mathcal{H}_n^i} |\mu_n(F_n^{i,j})|(b^i - a^i) \geq (T - S - \tau) \frac{1}{2\pi} |\mu|(\Omega).
\end{aligned}$$

Letting $S \to 0$, $T \to 1$ and $\tau \to 0$, we deduce the Γ-liminf inequality. ◻

Remark 2.6. Let $C, C' > 0$ be fixed positive constants. Here we observe that nothing changes in our Γ-convergence result if in the definition of $\Omega_\varepsilon(\mu)$ we remove balls of radius $C\varepsilon$ instead of ε and if we multiply the second term $|\mu|(\Omega)$ in (2.5) by C'. More precisely given $\mu \in X$, define

$$\Omega_\varepsilon^C(\mu) := \Omega \setminus \bigcup_{x_i \in \operatorname{supp}(\mu)} B_{C\varepsilon}(x_i). \tag{2.29}$$

Define consequently the space of admissible strains $\mathcal{AS}_\varepsilon^C(\mu)$ associated with μ as follows:

(2.30)

$$\begin{aligned}
\mathcal{AS}_\varepsilon^C(\mu) := \bigg\{ \psi \in L^2(\Omega_\varepsilon^C(\mu); \mathbb{R}^2) :\ &\operatorname{curl} \psi = 0 \text{ in } \Omega_\varepsilon^C(\mu) \text{ in the sense of distributions}, \\
&\int_{\partial A} \psi(s) \cdot \tau(s) \, ds = \mu(A) \\
&\text{for every open set } A \subset \Omega \text{ with } \partial A \text{ smooth and with } \partial A \subset \Omega_\varepsilon^C(\mu) \bigg\}.
\end{aligned}$$

Finally let $\mathcal{E}_\varepsilon^{C,C'}(\mu)$ be defined by

$$\mathcal{E}_\varepsilon^{C,C'}(\mu) := \min_{\psi \in \mathcal{AS}_\varepsilon^C(\mu)} \int_{\Omega_\varepsilon^C(\mu)} |\psi(x)|^2 \, dx + C'|\mu|(\Omega), \tag{2.31}$$

and let $\mathcal{F}_\varepsilon^{C,C'} := 1/|\log\varepsilon|\mathcal{E}_\varepsilon^{C,C'}$ be the corresponding rescaled functionals. Then Theorem 2.4 still holds true with $\mathcal{F}_\varepsilon$ replaced by $\mathcal{F}_\varepsilon^{C,C'}$.

In this respect the choice of the core radius and of the core energy does not play an essential role in the asymptotic behavior of the functionals $\mathcal{F}_\varepsilon$ as $\varepsilon \to 0$. This fact also gives a partial justification of the use of linearized elasticity in $\Omega_\varepsilon^C(\mu)$. In fact, the recovering sequence ψ_n given in (2.26) satisfies

$$\|\psi_n\|_{L^\infty(\Omega_{\varepsilon_n}^C(\mu_n);\mathbb{R}^2)} \leq \frac{1}{2\pi C\varepsilon_n} + O(n),$$

where $O(n)$ is uniformly bounded with respect to n. Recalling that the admissible strains should be rescaled by ε_n (because the Burgers vector has to be rescaled by ε_n), we deduce that the modulus of the gradient of the rescaled recovering sequence can be chosen arbitrarily small, choosing C big enough. This is our partial justification of the use of linear elasticity.

3. The discrete model. Here we give a Γ-convergence result in a discrete model for the stored energy associated with a configuration of screw dislocations, as the atomic distance ε tends to 0. The model follows the general theory of eigenstrains (we refer the reader to [13]): a dislocation in the crystal is associated with a preexisting plastic strain in the reference lattice. In the next section we will describe our discrete model, which follows the lines of the more general theory introduced in [2].

3.1. Description of the discrete model. We will consider the illustrative case of a square lattice, with nearest-neighbor interactions. Let $\Omega \subset \mathbb{R}^2$ be a horizontal section of the region occupied by the cylindrical crystal. We will assume for simplicity Ω to be polygonal. In the reference configuration, the lattice of atoms is given by the set

$$\Omega_\varepsilon^0 := \{x \in \varepsilon\mathbb{Z}^2 \cap \overline{\Omega}\}.$$

We denote by Ω_ε^1 the class of *bonds* in Ω, i.e., the class of oriented ε-segments $[x, x+\varepsilon e_i]$, where e_1, e_2 is the canonical basis of $\mathbb{R}^2$ and x, $x+\varepsilon e_i \in \Omega_\varepsilon^0$.

Given a function $u : \Omega_\varepsilon^0 \to \mathbb{R}$, let us introduce the (rescaled) *discrete gradient* of u, $\mathbf{d}\, u : \Omega_\varepsilon^1 \to \mathbb{R}$, defined by

$$\mathbf{d}\, u([x, x+\varepsilon e_i]) := u(x+\varepsilon e_i) - u(x) \qquad \text{for every } [x, x+\varepsilon e_i] \in \Omega_\varepsilon^1. \tag{3.1}$$

Given a strain $\xi : \Omega_\varepsilon^1 \to \mathbb{R}$, the elastic energy associated with ξ is given by

$$E^d(\xi) := \sum_{v \in \Omega_\varepsilon^1} a(v)(\xi(v))^2, \tag{3.2}$$

where the function $a(v) \in \{1/2, 1\}$, introduced only to simplify some interpolation procedure (see property (b) of Proposition 3.3), is defined by

$$a([x, x+\varepsilon e_i]) := \begin{cases} \frac{1}{2} & \text{if } x,\ x+\varepsilon e_i \in \partial\Omega, \\ 1 & \text{otherwise.} \end{cases}$$

The elastic energy associated with a displacement $u:\Omega^0_\varepsilon\to\mathbb{R}$, in absence of dislocations, is given by $E^d(\mathbf{d}\,u)$.

To model the presence of dislocations, following [2] we introduce the class Ω^2_ε of oriented ε-squares $[x,x+\varepsilon e_2,x+\varepsilon e_1+\varepsilon e_2]$ with $x,\ x+\varepsilon e_2,\ x+\varepsilon e_1+\varepsilon e_2, x+\varepsilon e_1\in\Omega^0_\varepsilon$. Given $Q:=[x,x+\varepsilon e_2,x+\varepsilon e_1+\varepsilon e_2]\in\Omega^2_\varepsilon$, let us denote by $\tilde Q\subset\mathbb{R}^2$ the convex envelope of $\{x,x+\varepsilon e_1,x+\varepsilon e_2,x+\varepsilon e_1+\varepsilon e_2\}$. For simplicity we will always assume that $\overline{\Omega}=\cup_{Q_i\in\Omega^2_\varepsilon}\tilde Q_i$.

In this discrete setting, a dislocation is represented by a function $\alpha:\Omega^2_\varepsilon\to\mathbb{Z}$. The squares in the support of α represent the zone where a dislocation is present, while the value of α on these squares represents the multiplicity of the dislocation.

Given $\xi:\Omega^1_\varepsilon\to\mathbb{R}$, the function $\mathbf{d}\,\xi:\Omega^2_\varepsilon\to\mathbb{R}$ is defined by

$$\begin{aligned}(3.3)\quad \mathbf{d}\,\xi([x,x+\varepsilon e_2,x+\varepsilon e_1+\varepsilon e_2]) &:= \xi([x,x+\varepsilon e_2])+\xi([x+\varepsilon e_2,x+\varepsilon e_1+\varepsilon e_2])\\ -\xi([x+\varepsilon e_1,x+\varepsilon e_1+\varepsilon e_2])-\xi([x,x+\varepsilon e_1]) &\qquad \text{for every } [x,x+\varepsilon e_2,x+\varepsilon e_1+\varepsilon e_2]\in\Omega^2_\varepsilon.\end{aligned}$$

The elastic energy associated with a dislocation $\alpha:\Omega^2_\varepsilon\to\mathbb{Z}$ is given by

$$\mathcal{E}^d_\varepsilon(\alpha):=\min_{\xi:\Omega^1_\varepsilon\to\mathbb{R}:\,\mathbf{d}\,\xi=\alpha}E^d_\varepsilon(\xi). \tag{3.4}$$

Remark 3.1. Note that if α is a dipole of the type

$$\alpha(Q):=\begin{cases}-1 & \text{if } Q=[x,x+\varepsilon e_2,x+\varepsilon(e_1+e_2)],\\ +1 & \text{if } Q=[x+\varepsilon z e_1,x+\varepsilon(e_2+ze_1),x+\varepsilon(e_1+e_2+ze_1)],\\ 0 & \text{otherwise}\end{cases}$$

for some $x\in\Omega^0_\varepsilon$, $z\in\mathbb{Z}$, then $\alpha=\mathbf{d}\,\beta$, with β defined by

$$\beta(v):=\begin{cases}1 & \text{if } v=[x+\varepsilon s e_1,x+\varepsilon(se_1+e_2)] \text{ with } s\in\{1,\dots,z\},\\ 0 & \text{otherwise.}\end{cases}$$

Actually for every $\alpha:\Omega^2_\varepsilon\to\mathbb{Z}$ we can find β with $\mathbf{d}\,\beta=\alpha$. By linearity, it is sufficient to check it in the case

$$\alpha(Q):=\begin{cases}1 & \text{if } Q=[x,x+\varepsilon e_1,x+\varepsilon(e_1+e_2)],\\ 0 & \text{otherwise.}\end{cases}$$

We have $\alpha=\mathbf{d}\,\beta$, where

$$\beta(v):=\begin{cases}1 & \text{if } v=[x-\varepsilon s\,e_1,x-\varepsilon s\,e_1+\varepsilon\,e_2] \text{ with } s\in\{0\cup\mathbb{N}\},\\ 0 & \text{otherwise.}\end{cases}$$

Note that there are many β inducing the same α (such that $\mathbf{d}\,\beta=\alpha$). More precisely if $\mathbf{d}\,\beta=\alpha$, then α is induced exactly by

$$\{\beta+\mathbf{d}\,u,\ u:\Omega^0_\varepsilon\to\mathbb{R}\}.$$

This follows by the fact that if $\xi:\Omega^1_\varepsilon\to\mathbb{R}$ is such that $\mathbf{d}\,\xi=0$, then there exists $u:\Omega^0_\varepsilon\to\mathbb{R}$ such that $\xi=\mathbf{d}\,u$ and $\mathbf{d}\,\mathbf{d}\,u(Q)=0$ for every $u:\Omega^0_\varepsilon\to\mathbb{R}$ and for every $Q\in\Omega^2_\varepsilon$.

We deduce that if $d\beta = \alpha$, then

$$\mathcal{E}^d_\varepsilon(\alpha) := \min_{u:\Omega^0_\varepsilon \to \mathbb{R}} E^d_\varepsilon(\mathbf{d}\, u - \beta).$$

Therefore β can be interpreted as an eigenstrain associated with the dislocation α. However, we stress that the energy depends on α and not on the particular choice of the eigenstrain inducing α.

3.2. The Γ-convergence result. To study the asymptotic behavior of the elastic energy functionals $\mathcal{E}^d_\varepsilon$ as $\varepsilon \to 0$ in terms of Γ-convergence, it is convenient to define a common space of configurations of dislocations independent of ε. To this aim, to every dislocation $\alpha : \Omega^2_\varepsilon \to \mathbb{Z}$ we associate the measure

$$\hat{\mu}(\alpha) := \sum_{Q \in \Omega^2_\varepsilon} \alpha(Q)\delta_{x(Q)},$$

where $x(Q)$ denotes the center of Q. Therefore, as in the continuum case, the space of dislocations is the space X defined in (2.1). Moreover we denote by X_ε the subspace of X given by the measures μ such that $\mu = \hat{\mu}(\alpha)$ for some $\alpha \in \Omega^2_\varepsilon$. Finally, given $\mu \in X_\varepsilon$, we will denote by $\tilde{\alpha}(\mu) : \Omega^2_\varepsilon \to \mathbb{Z}$ the (unique) dislocation satisfying $\hat{\mu}(\tilde{\alpha}(\mu)) = \mu$.

The class of discrete admissible strains associated with ε and $\mu \in X_\varepsilon$ is defined by

$$\mathcal{AS}^d_\varepsilon(\mu) := \{\xi : \Omega^1_\varepsilon \to \mathbb{R} : \mathbf{d}\,\xi = \tilde{\alpha}(\mu)\}.$$

The rescaled energy functionals take the form

$$\mathcal{F}^d_\varepsilon(\mu) := \begin{cases} \frac{1}{|\log \varepsilon|}\, \mathcal{E}_\varepsilon(\tilde{\alpha}(\mu)) & \text{if } \mu \in X_\varepsilon, \\ +\infty & \text{in } X \setminus X_\varepsilon. \end{cases} \tag{3.5}$$

Remark 3.2. Here we notice that in the discrete model we do not need to introduce the core energy $|\mu|(\Omega)$ as in the continuum case to obtain an estimate similar to (2.9). The term $|\mu|(\Omega)$, in the continuum model, represents the energy stored in a region surrounding the dislocations, whose diameter is comparable to the atomic distance. This interpretation is fully justified by the following easy computation: Let $\mu \in X_\varepsilon$, let $x \in \text{supp}(\mu)$, and let $Q_\varepsilon(x)$ be the ε-square centered at x. For every admissible strain $\xi \in \mathcal{AS}^d_\varepsilon(\mu)$, we have by definition

$$\sum_{v \in \partial Q_\varepsilon(x)} \xi(v) = \mu(x).$$

We deduce that

$$\sum_{v \in \partial Q_\varepsilon(x)} |\xi(v)|^2 \geq C, \tag{3.6}$$

where C is a constant independent of ε. Therefore, in the discrete model, the energy stored in the bonds near the dislocations turns out to be controlled from below by $|\mu|(\Omega)$. As observed in Remark 2.6, a sharp computation of this energy becomes unnecessary in the continuum model in the study of the asymptotic behavior of the elastic energy as $\varepsilon \to 0$.

By (3.6) we deduce (as in (2.9)) that if $\varepsilon_n \to 0$ and $\{\mu_n\}$ is a sequence in X such that, for every $n \in \mathbb{N}$, $\mathcal{F}^d{}_{\varepsilon_n}(\mu_n) \le E$ for some positive constant E, then

$$\sharp\,\mathrm{supp}(\mu_n) \le CE|\log(\varepsilon_n)| \qquad \text{for every } n \in \mathbb{N}, \tag{3.7}$$

where C is a fixed positive constant independent of ε.

The candidate Γ-limit of the functionals $\mathcal{F}^d_\varepsilon$, as in the continuum case (see (2.7)), is the functional $\mathcal{F} : X \to \mathbb{R}$ defined by

$$\mathcal{F}(\mu) := \frac{1}{2\pi}|\mu|(\Omega) \qquad \text{for every } \mu \in X. \tag{3.8}$$

Now we provide some interpolation procedures which will be used in the proof of the Γ-convergence result. Let $u : \Omega^0_\varepsilon \to \mathbb{R}$. Let us introduce its extension $\tilde{u} : \Omega \to \mathbb{R}$, defined in the following way: We divide every $Q \in \Omega^2_\varepsilon$ (more precisely, every $\tilde{Q}$ with $Q \in \Omega^2_\varepsilon$) into two triangles. In each triangle T, $\tilde{u}$ is the only affine function coinciding with u on the vertices of T. In a similar way, given a function $\xi : \Omega^1_\varepsilon \to \mathbb{R}$ we define $\tilde{\xi} : \Omega \to \mathbb{R}^2$ imposing on each triangle

$$\tilde{\xi} \equiv (\xi(v_1(T)), \xi(v_2(T))), \tag{3.9}$$

where $v_1(T)$ and $v_2(T)$ are the horizontal (parallel to e_1) and the vertical (parallel to e_2) edges of T, respectively. We collect in the following proposition some properties satisfied by the interpolated functions introduced above.

Proposition 3.3. *The following facts hold.*

(a) *For every* $u : \Omega^0_\varepsilon \to \mathbb{R}$, *we have* $\nabla\tilde{u} = \frac{1}{\varepsilon}\widetilde{\mathbf{d}u}$.

(b) *For every* $\xi : \Omega^1_\varepsilon \to \mathbb{R}$, *we have* $E^d(\xi) = \|\tilde{\xi}\|^2_{L^2(\Omega;\mathbb{R}^2)}$.

(c) *The function* $\tilde{\xi}$ *belongs to the class* $\mathcal{AS}^C_\varepsilon(\hat{\mu}(\mathbf{d}\,\xi))$ *defined in* (2.30) *for every* $C \ge 2^{1/2}$.

Now we are in a position to give our Γ-convergence result in this discrete model, for the elastic energy functionals $\mathcal{F}^d_\varepsilon$ as $\varepsilon \to 0$.

Theorem 3.4. *The following* Γ*-convergence result holds.*

(i) *Equicoercivity: Let* $\varepsilon_n \to 0$, *and let* $\{\mu_n\}$ *be a sequence in* X *such that* $\mathcal{F}^d_{\varepsilon_n}(\mu_n) \le E$ *for some positive constant* E *independent of* n. *Then (up to a subsequence)* $\mu_n \xrightarrow{f} \mu$ *for some* $\mu \in X$.

(ii) Γ*-convergence: The functionals* $\mathcal{F}^d_{\varepsilon_n}$ Γ*-converge to* $\mathcal{F}$ *as* $\varepsilon_n \to 0$ *with respect to the flat norm; i.e., the following inequalities hold.*

Γ*-liminf inequality:* $\mathcal{F}(\mu) \le \liminf \mathcal{F}^d_{\varepsilon_n}(\mu_n)$ *for every* $\mu \in X$, $\mu_n \xrightarrow{f} \mu$ *in* X.

Γ*-limsup inequality: Given* $\mu \in X$, *there exists* $\{\mu_n\} \subset X$ *with* $\mu_n \xrightarrow{f} \mu$ *such that* $\limsup \mathcal{F}^d_{\varepsilon_n}(\mu_n) \le \mathcal{F}(\mu)$.

Proof. We begin by proving the equicoercivity property.

Equicoercivity: Let $\xi_n \in \mathcal{AS}^d_{\varepsilon_n}(\mu_n)$ be such that

$$\frac{1}{|\log \varepsilon_n|} E^d_{\varepsilon_n}(\xi_n) \le E + 1.$$

Let $C \ge 2^{1/2}$. By Proposition 3.3 we have that the functions $\tilde{\xi}_n$ introduced in (3.9) are in the class $\mathcal{AS}^C_\varepsilon(\mu_n)$ defined in (2.30). By Proposition 3.3 and by (3.7) we deduce that $\mathcal{F}^{C,1}_{\varepsilon_n}(\mu_n) \le K$ for some positive constant $K > 0$. Therefore by Theorem 2.4 and Remark 2.6 we deduce that the equicoercivity property holds.

Γ-*liminf inequality:* Let $\mu_n \xrightarrow{f} \mu$ in X, and let $\xi_n \in \mathcal{AS}^d_{\varepsilon_n}(\mu_n)$ be such that

$$\liminf \frac{1}{|\log \varepsilon_n|} E^d_{\varepsilon_n}(\xi_n) = \liminf \mathcal{F}^d_{\varepsilon_n}(\mu_n).$$

By Proposition 3.3 we have that $\tilde{\xi}_n \in \mathcal{AS}^C_\varepsilon(\mu_n)$, with $C \geq 1/2$. By the Γ-liminf inequality of Theorem 2.4 and by Remark 2.6, we deduce that for every positive constant $C' > 0$ we have

$$\mathcal{F}(\mu) \leq \liminf \frac{1}{|\log \varepsilon_n|} \left(\int_{\Omega^C_{\varepsilon_n}(\mu_n)} |\tilde{\xi}_n|^2 + C'|\mu_n|(\Omega) \right).$$

By the arbitrariness of C', by Remark 3.2, and by Proposition 3.3 we deduce

$$\mathcal{F}(\mu) \leq \liminf_n \frac{1}{|\log \varepsilon_n|} \int_{\Omega^C_{\varepsilon_n}(\mu_n)} |\tilde{\xi}_n|^2 \leq \liminf_n \frac{1}{|\log \varepsilon_n|} E^d_{\varepsilon_n}(\xi_n) = \liminf_n \mathcal{F}^d_{\varepsilon_n}(\mu_n);$$

i.e., the Γ-liminf inequality holds true.

Γ-*limsup inequality:* It is enough to prove the Γ-limsup inequality assuming that $|\mu(x)| = 1$ for every $x \in \operatorname{supp}(\mu)$. In fact the class of measures satisfying this assumption is dense in energy and with respect to the flat convergence in X.

The recovering sequence is given by the constant sequence $\mu_n \equiv \mu$. We have to construct a sequence of admissible strains $\xi_n \in \mathcal{AS}^d_{\varepsilon_n}(\mu_n)$ satisfying

$$\mathcal{F}(\mu) \geq \limsup \frac{1}{|\log \varepsilon_n|} E^d_{\varepsilon_n}(\xi_n). \tag{3.10}$$

Let us fix $x_i \in \operatorname{supp}(\mu) \cap \Omega$. For every $v := [v_1, v_2] \in \Omega^1_\varepsilon$, let us denote by $T(v)$ the triangle whose vertices are x_i, v_1, and v_2 and, by $\theta_{x_i}(v) \in [0, 2\pi)$, its angle at the point x_i. We consider the field $\xi^n_{x_i} : \Omega^1_\varepsilon \to \mathbb{R}$ defined by

$$\xi^n_{x_i}(v) := \theta_{x_i}(v) o(T) \qquad \text{for every } v \in \Omega^1_{\varepsilon_n},$$

where $o(T) \in \{-1, 1\}$ is equal to 1 if the oriented segments $[x, v_1]$, $[v_1, v_2]$, $[v_2, x]$ induce a clockwise orientation to ∂T; $o(T) = -1$ otherwise.

Let us fix $0 < \delta < 1$. We set

$$A^n_i := \{x \in \Omega : \, |x - x_i| < \varepsilon_n^{1-\delta}\},$$
$$B^n_i := \Omega \setminus \overline{A}^n_i.$$

Let us consider the function $\tilde{\xi}^n_{x_i} : \Omega \to \mathbb{R}^2$ of $\mathbb{R}^n$ defined in (3.9). By construction, for n big enough, $\tilde{\xi}^n_{x_i}$ satisfies the following.

(i) For every $x \in A^n_i$,

$$|\tilde{\xi}^n_{x_i}(x)| \leq \frac{C}{\max\{|x - x_i|, \varepsilon_n\}},$$

where C is a positive constant independent of ε_n.

(ii) For every $x \in B^n_i$,

$$|\tilde{\xi}^n_{x_i}(x)| = \frac{1 + o(\varepsilon_n)}{|x - x_i|},$$

where $o(\varepsilon_n) \to 0$ as $\varepsilon_n \to 0$.

The recovering sequence ξ_n is defined by

$$\xi_n := \sum_i \xi^n_{x_i}.$$

It can be easily proved that $\xi_n \in \mathcal{AS}^d_{\varepsilon_n}(\mu_n)$. By Proposition 3.3 and by properties (i) and (ii) above it follows that

$$\begin{aligned}\lim_{\varepsilon_n \to 0} \frac{1}{|\log \varepsilon_n|} E^d_{\varepsilon_n}(\xi_n) &= \lim_{\varepsilon_n \to 0} \frac{1}{|\log \varepsilon_n|} \|\tilde{\xi}_n\|^2_{L^2(\Omega;\mathbb{R}^2)} \\ &= \frac{1}{2\pi} \sharp \operatorname{supp}(\mu)(1 + o(\delta)) = \mathcal{F}(\mu)(1 + o(\delta)),\end{aligned}$$

where $o(\delta) \to 0$ as $\delta \to 0$, and this concludes the proof of (3.10) and of the Γ-convergence result. $\square$

Acknowledgments. I wish to thank Adriana Garroni, Stefan Müller, and Michael Ortiz for interesting and fruitful discussions.

REFERENCES

[1] G. Alberti, S. Baldo, and G. Orlandi, *Variational convergence for functionals of Ginzburg-Landau type*, Indiana Univ. Math. J., 54 (2005), pp. 1411–1472.
[2] M. P. Ariza and M. Ortiz, *Discrete crystal elasticity and discrete dislocations in crystals*, Arch. Ration. Mech. Anal., 178 (2006), pp. 149–226.
[3] F. Bethuel, H. Brezis, and F. Hélein, *Ginzburg-Landau Vortices*, Progr. Nonlinear Differential Equations Appl. 13, Birkhäuser Boston, Boston, 1994.
[4] P. Cermelli and G. Leoni, *Energy and forces on dislocations*, SIAM J. Math. Anal., 37 (2005), pp. 1131–1160.
[5] G. Dal Maso, *An Introduction to Γ-Convergence*, Birkhäuser Boston, Boston, 1993.
[6] R. Dautray and J. L. Lions, *Mathematical Analysis and Numerical Nethods for Science and Technology*, Vol. 3, Springer, Berlin, 1988.
[7] H. Federer, *Geometric Measure Theory*, Grundlehren Math. Wiss. 153, Springer, New York, 1969.
[8] A. Garroni and S. Müller, *Γ-limit of a phase-field model of dislocations*, SIAM J. Math. Anal., 36 (2005), pp. 1943–1964.
[9] A. Garroni and S. Müller, *A variational model for dislocations in the line tension limit*, Arch. Ration. Mech. Anal., 181 (2006), pp. 535–578.
[10] R. L. Jerrard and H. M. Soner, *The Jacobian and the Ginzburg-Landau energy*, Calc. Var. Partial Differential Equations, 14 (2002), pp. 151–191.
[11] M. Koslowski, A. M. Cuitino, and M. Ortiz, *A phase-field theory of dislocation dynamics, strain hardening and hysteresis in ductile single crystal*, J. Mech. Phys. Solids, 50 (2002), pp. 2597–2635.
[12] F. Morgan, *Geometric Measure Theory. A Beginner's Guide*, Academic Press, Boston, 1988.
[13] T. Mura, *Micromechanics of Defects in Solids*, Kluwer Academic, Boston, 1987.
[14] R. Phillips, *Crystals, Defects and Microstructures: Modelling Across Scales*, Cambridge University Press, New York, 2001.
[15] E. Sandier, *Lower bounds for the energy of unit vector fields and applications*, J. Funct. Anal., 152 (1998), pp. 379–403.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 470–488

ANALYSIS OF UNIVARIATE NONSTATIONARY SUBDIVISION SCHEMES WITH APPLICATION TO GAUSSIAN-BASED INTERPOLATORY SCHEMES*

NIRA DYN†, DAVID LEVIN†, AND JUNGHO YOON‡

Abstract. This paper is concerned with nonstationary subdivision schemes. First, we derive new sufficient conditions for C^ν smoothness of such schemes. Next, a new class of interpolatory $2m$-point nonstationary subdivision schemes based on Gaussian interpolation is presented. These schemes are shown to be $C^{L+\mu}$ with $L \in \mathbb{Z}_+$ and $\mu \in (0,1)$, where L is the integer smoothness order of the known $2m$-point Deslauriers–Dubuc interpolatory schemes.

Key words. nonstationary subdivision, radial basis function, Gaussian, asymptotical equivalence, interpolation

AMS subject classifications. 41A05, 41A25, 41A30, 65D10, 65D17

DOI. 10.1137/050638217

1. Introduction.

Subdivision is a powerful tool for the fast construction of smooth curves and surfaces from a set of control points by means of iterative refinements. In this paper, we consider subdivision schemes for curves. A univariate binary uniform stationary subdivision scheme defines recursively new sets of points $P^k = \{p_j^k : j \in \mathbb{Z}\}$ at level $k > 0$ from a given set of control points at level zero $P^0 = \{p_j^0 : j \in \mathbb{Z}\}$, formally, by

$$P^{k+1} = SP^k, \quad k = 0, 1, \ldots.$$

A point of P^k is defined by a finite linear combination of points of P^{k-1} with two different rules,

$$p_j^{k+1} = \sum_{n\in\mathbb{Z}} a_{j-2n} p_n^k, \quad k \in \mathbb{Z}_+,\ j \in \mathbb{Z}.$$

Nonstationary subdivision schemes consist of recursive refinements of an initial sparse sequence with the use of rules that may vary from level to level but are the same everywhere on the same level. Therefore, in the binary case, starting with the control points $P^0 = \{p_n^0 : n \in \mathbb{Z}\}$, we define new sets of points $P^k = \{p_n^k : n \in \mathbb{Z}\}$ generated by the relation

$$p_j^{k+1} = \sum_{n\in\mathbb{Z}} a_{j-2n}^{[k]} p_n^k, \quad k \in \mathbb{Z}_+,\ j \in \mathbb{Z}, \tag{1.1}$$

where the set of coefficients $a^{[k]} := \{a_n^{[k]}\}$ is termed the mask of the rule at level k. We denote this rule by $S_{a^{[k]}}$ and the corresponding nonstationary scheme by $\{S_{a^{[k]}}\}$. It is

*Received by the editors August 16, 2005; accepted for publication (in revised form) August 31, 2006; published electronically July 4, 2007.
http://www.siam.org/journals/sima/39-2/63821.html

†School of Mathematical Sciences, Sackler Faculty of Exact Sciences, Tel-Aviv University, Ramat-Aviv, Tel-Aviv 69978, Israel (niradyn@post.tau.ac.il, levin@post.tau.ac.il).

‡Department of Mathematics, Ewha W. University, Seoul 120-750, South Korea (yoon@math.ewha.ac.kr). The work of the third author has been supported by grant R01-2006-000-10424-0 from the Korea Science and Engineering Foundation in Ministry of Science and Technology.

common to assume that for each level k, only a finite number of coefficients $a_n^{[k]} \in \mathbb{R}$ are nonzero so that changes in a control point affect only its local neighborhood. This property clearly facilitates the practical implementation of (1.1). A subdivision scheme is said to be stationary when the masks $a_n^{[k]}$ are independent of the levels; then we use the notation $a_n := a_n^{[k]}$. We denote this rule by S_a. Nonstationary subdivision schemes are useful because they can provide design flexibility, and the masks can be adapted to the geometrical configuration of the given data. Nonstationary subdivision schemes are studied in [2, 8, 10, 18], while a general treatment of stationary schemes can be found in [1, 5, 6, 7, 9].

The analysis of a subdivision scheme can be reduced to the case of initial control points in $\mathbb{R}$ since each component of the curve is a scalar function generated by the same subdivision scheme. Therefore, starting with values $f^0 = \{f_n^0 \in \mathbb{R} : n \in \mathbb{Z}\}$, we consider $f^k = \{f_n^k \in \mathbb{R} : n \in \mathbb{Z}\}$ generated by the relation

$$f_j^{k+1} = \sum_{n\in\mathbb{Z}} a_{j-2n}^{[k]} f_n^k, \quad k \in \mathbb{Z}_+. \tag{1.2}$$

DEFINITION 1.1. *A binary subdivision scheme is said to be C^ν if for the initial data $\delta = \{f_n^0 = \delta_{n,0} : n \in \mathbb{Z}\}$ there exists a limit function $\phi_0 \in C^\nu(\mathbb{R})$, $\phi_0 \not\equiv 0$, satisfying*

$$\lim_{k\to\infty} \sup_{n\in\mathbb{Z}} |f_n^k - \phi_0(2^{-k}n)| = 0. \tag{1.3}$$

Natural questions in the analysis of subdivision schemes concern the conditions for convergence and the conditions for the limit functions to be C^ν. In particular, in this study, we are interested in the class of interpolatory subdivision schemes which refine data by inserting values corresponding to intermediate points, using linear combinations of neighboring points. The general form of their refinement rules is as follows:

$$\begin{aligned} f_{2j}^{k+1} &= f_j^k, \\ f_{2j+1}^{k+1} &= \sum_{n\in\mathbb{Z}} a_{2n+1}^{[k]} f_{j-n}^k, \quad j \in \mathbb{Z},\ k \in \mathbb{Z}_+. \end{aligned}$$

Examples of such stationary schemes are the four-point scheme by Dyn, Gregory, and Levin [7] and the Deslauriers–Dubuc schemes [6], where finer level points are determined by local polynomial interpolation of the coarse level points. When the finer level points are determined by $2m$-point interpolation from a space of exponential polynomials, the resulting scheme is nonstationary and has smoothness properties, as in the $2m$-point Deslauriers–Dubuc scheme [10].

An analysis of the smoothness of nonstationary subdivision schemes is discussed in [8]; however, the conditions given in [8] are too strong. Thus, the first objective of this paper is to provide a new tool for the regularity analysis, improving the conditions in [8]. It can be applied to a wide class of nonstationary subdivision schemes, both interpolatory and noninterpolatory. Further, the results can be used directly for the smoothness analysis of nonstationary wavelet systems, which is one of the important issues in wavelet construction. Second, in this paper, we study a new class of nonstationary interpolatory subdivision schemes, where the value at the inserted point is obtained by radial basis function (RBF) interpolation to data at $2m$ points symmetric to the inserted point. Among the many possible RBFs, we employ the Gaussian function $G(x) = e^{-|x|^2/\lambda^2}$ with λ as a shape parameter. We show that the resulting $2m$-point subdivision scheme converges and has the smoothness $C^{L+\mu}$ with $L \in \mathbb{Z}_+$ and $\mu \in (0, 1)$, where L is the integer smoothness order of the $2m$-point Deslauriers–Dubuc scheme. The proof of these results is based on the new sufficient condition for smoothness of nonstationary schemes. Moreover, we will see that the scheme itself has its own advantages in view of approximation.

The paper is organized as follows. In section 2, we present known conditions for the convergence and smoothness of nonstationary schemes and derive new sufficient conditions for the smoothness of such schemes. In section 3, along with the basic setting of RBF interpolation, we present a new family of interpolatory subdivision schemes based on Gaussian interpolation. Next, we show using the results of section 2 that the new $2m$-point schemes have the same integer smoothness as the $2m$-point Deslauriers–Dubuc interpolatory scheme. Finally, in section 4, we illustrate the performance of the new interpolatory schemes by some numerical examples.

2. Sufficient conditions for smoothness of nonstationary schemes. Nonstationary subdivision schemes define recursively values $f^k := \{f_n^k : n \in \mathbb{Z}\}$ by rules depending on the level k, as in (1.2). To simplify the presentation of a subdivision scheme and its analysis, it is convenient to assign to each rule, defined by a mask $a^{[k]} = \{a_n^{[k]}\}$, the Laurent polynomial

$$a^{[k]}(z) := \sum_{n \in \mathbb{Z}} a_n^{[k]} z^n.$$

Assume here that for each level k, supp $(a^{[k]}) \subset [-N, N]$ for some integer $N > 0$. This implies that the Laurent polynomials $a^{[k]}(z)$ have a finite degree.

A stationary subdivision scheme $\{S_a\}$ has the formal relation $f^k = S_a^k f^0$. The limit function of a C^0 stationary scheme is denoted by $S_a^\infty f^0$. In particular, for the given data $\delta = \{\delta_{0,n} : n \in \mathbb{Z}\}$ at level 0, with the Kronecker delta $\delta_{n,0}$, the basic limit function of $\{S_a\}$ is defined by

$$\phi = S_a^\infty \delta.$$

For a nonstationary subdivision scheme $\{S_{a^{[k]}}\}$, we have the formal relation

$$f^k = S_{a^{[k-1]}} \cdots S_{a^{[0]}} f^0.$$

Further, for a convergent scheme $\{S_{a^{[k]}}\}$, its *basic limit function* is the function

$$\phi_0 := \lim_{k \to \infty} S_{a^{[k]}} \cdots S_{a^{[0]}} \delta.$$

It clearly follows from the linearity of (1.2) that for any initial data $f^0 = \{f_n^0 : n \in \mathbb{Z}\} \in \ell^\infty(\mathbb{Z})$, the limit function of $\{S_{a^{[k]}}\}$ can be written as

$$f^\infty = \sum_{n \in \mathbb{Z}} f_n^0 \phi_0(\cdot - n).$$

First we cite a basic result about the smoothness of stationary subdivision schemes.

THEOREM 2.1 (smoothing factors in stationary schemes [5]). *Consider a stationary subdivision scheme $\{S_a\}$ with the Laurent polynomial*

$$a(z) = \frac{1}{2}(1 + z)b(z),$$

where the subdivision scheme $\{S_b\}$ *corresponding to* $b(z)$ *is* C^γ*. Then the scheme* $\{S_a\}$ *is convergent, and its basic limit function* ϕ *is in* $C^{\gamma+1}$*.*

Now, for the analysis of the smoothness of nonstationary schemes, we adopt the notion of asymptotically equivalent schemes [8]: A nonstationary subdivision scheme $\{S_{a^{[k]}}\}$ is *asymptotically equivalent* to a stationary scheme $\{S_a\}$, $\{S_{a^{[k]}}\} \approx \{S_a\}$, if

$$\sum_{k\in\mathbb{Z}_+} \|S_{a^{[k]}} - S_a\|_\infty < \infty, \tag{2.1}$$

where

$$\|S_{a^{[k]}}\|_\infty = \max\left\{\sum_{n\in\mathbb{Z}} |a^{[k]}_{2n}|,\ \sum_{n\in\mathbb{Z}} |a^{[k]}_{1+2n}|\right\}.$$

THEOREM 2.2 (see [8]). *Let* $\{S_a\}$ *be a* C^0 *stationary subdivision scheme, and let* $\{S_{a^{[k]}}\} \approx \{S_a\}$ *with* $\operatorname{supp}(a) = \operatorname{supp}(a^{[k]})$ *for* $k \in \mathbb{Z}_+$*. Then* $\{S_{a^{[k]}}\}$ *is* C^0*, and if*

$$\|S_{a^{[k]}} - S_a\|_\infty \le c2^{-k}, \quad k \in \mathbb{Z}_+,$$

then the basic limit function ϕ_0 *of* $\{S_{a^{[k]}}\}$ *is Hölder continuous with some exponent* $\nu > 0$*.*

An analysis of the smoothness of nonstationary subdivision schemes is also discussed in [8]; however, the conditions given in [8] are too strong. Thus, the purpose of this section is to provide less restrictive sufficient conditions for the smoothness of nonstationary schemes. Specifically, we will show that a factor $(1 + r_k z)$ in the Laurent polynomials $a^{[k]}(z)$ is a smoothing factor if $|1 - r_k| \le c2^{-k}$, while in [8], r_k is required to satisfy the condition

$$r_k = e^{c2^{-k-1}}(1 + \epsilon_k) \quad \text{with} \quad \sum_{k=K}^{\infty} |\epsilon_k| 2^k < \infty.$$

Furthermore, we infer results on the smoothness of an interpolatory nonstationary scheme from the smoothness of a stationary scheme, which is asymptotically equivalent to it.

Let $a(z)$ be the Laurent polynomial associated with a stationary scheme $\{S_a\}$ with the property $a^{(\ell)}(-1) = 0$ for $\ell = 0, \dots, M-1$ and $a^{(M)}(-1) \ne 0$. Accordingly, it can be written as

$$a(z) = \sum_{n\in\mathbb{Z}} a_n z^n = 2^{-M}(1+z)^M b(z) \tag{2.2}$$

with $b(-1) \ne 0$. When a scheme $\{S_{a^{[k]}}\}$ is asymptotically equivalent to $\{S_a\}$, by definition, $|a^{[k]}_n - a_n| = o(1)$ as k tends to ∞. Hence, the Laurent polynomial $a^{[k]}(z)$ associated with $S_{a^{[k]}}$ has M-roots in the neighborhood $z = -1$ in the complex plane, and it can be written in the form of

$$a^{[k]}(z) = b^{[k]}(z) \prod_{n=1}^{M} \frac{1}{2}(1 + r_{k,n} z), \quad b^{[k]}(-1) = c + o(1), \quad c \ne 0, \tag{2.3}$$

with $r_{k,n}$ complex numbers such that $r_{k,n} \to 1$ as k tends to ∞. Moreover, since $a^{(\ell)}(-1) = 0$ for $\ell = 0, \dots, M-1$ (see (2.2)), it is easy to see that $D^\ell a^{[k]}(-1) = o(1)$

as $k \to \infty$, where D^ℓ indicates the differential operator of order ℓ. In this study, we require $D^\ell a^{[k]}(-1)$ to satisfy the following stronger condition.

CONDITION A. *A nonstationary subdivision scheme $\{S_{a^{[k]}}\}$ satisfies this condition if the corresponding Laurent polynomials $\{a^{[k]}(z)\}$ are of the form* (2.3) *and if*

$$|D^\ell a^{[k]}(-1)| \le c2^{-(M-\ell)k}, \quad \ell = 0, \dots, M-1, \quad k \in \mathbb{Z}_+.$$

In what follows, we show that if

$$a^{[k]}(z) = c^{[k]}(z) \prod_{n=1}^{L} \frac{1}{2}(1 + r_{k,n} z)$$

and if $\{S_{c^{[k]}}\}$ is $C^{N+\mu}$ with $N \in \mathbb{Z}_+$ and $\mu \in (0,1)$, then $\{S_{a^{[k]}}\}$ with Laurent polynomials of the form (2.3) satisfying Condition A has the smoothness $C^{N+L+\nu}$ with $\nu \in (0,1)$. First we show that a factor $(1 + r_k z)$ in the Laurent polynomials of a nonstationary scheme with $|1 - r_k| \le c2^{-k}$ is a smoothing factor. For this, we cite the following lemma.

LEMMA 2.3 (see [8]). *Consider a nonstationary subdivision scheme $\{S_{a^{[k]}}\}$ with Laurent polynomials of the form*

$$a^{[k]}(z) = \frac{1}{2}(1 + r_k z) b^{[k]}(z).$$

Let ϕ_a, ϕ_b, and h be the basic limit functions of $\{S_{a^{[k]}}\}$, $\{S_{b^{[k]}}\}$, and $\{S_{1+r_k z}\}$, respectively. Then

$$\phi_a = \int_{\mathbb{R}} \phi_b(\cdot - t) h(t)\, dt.$$

For the following analysis, it is necessary to remark that the basic limit function h of $\{S_{1+r_k z}\}$ is bounded and satisfies the following properties:

$$\text{(2.4)} \qquad \begin{aligned} &\text{(a)}\ \operatorname{supp} h = [0,1), \\ &\text{(b)}\ h\big((j + 2^{-1})2^{-k}\big) = r_k h(j2^{-k}), \quad k \in \mathbb{Z}_+,\ j = 0, \dots, 2^k - 1; \end{aligned}$$

see Example 2 in [8] for the details.

LEMMA 2.4. *Let ϕ_b and h be the basic limit functions of $\{S_{b^{[k]}}\}$ and $\{S_{1+r_k z}\}$, respectively. Suppose that*

$$|1 - r_k| \le c2^{-k}, \quad k \ge K \in \mathbb{Z}_+. \tag{2.5}$$

For each $k \in \mathbb{Z}_+$, define the sequence of functions

$$I_k(x) = \int_{x-1}^{x} \phi_b(t) h_k(x - t)\, dt, \tag{2.6}$$

where

$$h_k(t) = h(j2^{-k}), \quad j2^{-k} \le t < (j+1)2^{-k}, \quad j = 0, \dots, 2^k - 1. \tag{2.7}$$

If ϕ_b is Hölder continuous with some exponent $\nu > 0$, then I_k satisfies the following properties:

(a) *For any $k, \ell \geq K \in \mathbb{Z}_+$ with $\ell > k$,*

$$|I'_\ell(x) - I'_k(x)| \leq c2^{-\nu k}. \tag{2.8}$$

(b) *There exists $\delta_0 > 0$ such that for any $\delta < \delta_0$,*

$$|I'_k(x+\delta) - I'_k(x)| \leq c\delta^\mu, \quad k \geq K \in \mathbb{Z}_+,$$

for some $\mu \in (0,1)$.

Proof. (a) By (2.7), we have

$$I'_k(x) = \sum_{j=0}^{2^k-1} h(j2^{-k})[\phi_b(x - j2^{-k}) - \phi_b(x-(j+1)2^{-k})] \in C(\mathbb{R}), \tag{2.9}$$

and after some calculations, we get

$$\begin{aligned} I'_{k+1}(x) - I'_k(x) = &\sum_{j=0}^{2^k-1} [h((j+2^{-1})2^{-k}) - h(j2^{-k})] \\ &\cdot [\phi_b(x-(j+2^{-1})2^{-k}) - \phi_b(x-(j+1)2^{-k})]. \end{aligned}$$

Here, since ϕ_b is Hölder continuous with exponent $\nu > 0$,

$$|\phi_b(x-(j+2^{-1})2^{-k}) - \phi_b(x-(j+1)2^{-k})| \leq c2^{-\nu k} \tag{2.10}$$

with a constant $c > 0$ independent of j and x. Thus, in view of (2.4), (2.5), and the boundedness of h, we obtain the expression

$$|I'_{k+1}(x) - I'_k(x)| \leq c2^{-\nu k} \sum_{j=0}^{2^k-1} |(r_k - 1)h(j2^{-k})| \leq c'2^{-\nu k}. \tag{2.11}$$

It clearly induces the required result of (a).

(b) For the given Hölder exponent $\nu > 0$ and $\delta > 0$, choose $p = \frac{1}{\nu}$ and an integer $\tau > 0$ such that

$$(2^{-4p})^\tau \leq \delta \leq (2^{-2p})^\tau.$$

Note that from this inequality we can also obtain

$$\delta^{\frac{1}{2p}} \leq 2^{-\tau} \leq \delta^{\frac{1}{4p}}. \tag{2.12}$$

Then, by the triangle inequality,

$$\begin{aligned} I'_k(x+\delta) - I'_k(x)| \leq &|I'_k(x+\delta) - I'_\tau(x+\delta)| \\ &+ |I'_\tau(x+\delta) - I'_\tau(x)| + |I'_\tau(x) - I'_k(x)|. \end{aligned} \tag{2.13}$$

Assuming $k > \tau$, we apply (2.8) and (2.12) to obtain

$$|I'_\tau(x) - I'_k(x)| \leq c2^{-\tau\nu} \leq c\delta^{\frac{\nu}{4p}} \leq c\delta^{\frac{\nu^2}{4}}, \tag{2.14}$$

which estimates the first and last terms on the right-hand side of (2.13). Next, recalling that $\operatorname{supp} h = [0,1)$, the summation in (2.9) can be rewritten as follows (with k replaced by τ):

$$I'_\tau(x) = \sum_{j=0}^{2^\tau} [h(j2^{-\tau}) - h((j-1)2^{-\tau})]\phi_b(x - j2^{-\tau}).$$

Therefore,

$$|I'_\tau(x+\delta) - I'_\tau(x)| \le \sum_{j=0}^{2^\tau} |h(j2^{-\tau}) - h((j-1)2^{-\tau})||\phi_b(x+\delta-j2^{-\tau}) - \phi_b(x-j2^{-\tau})|.$$

Here, h is bounded and ϕ_b is Hölder continuous with exponent $\nu > 0$. Thus, due to (2.12) and the fact that $p = 1/\nu$, we get

$$|I'_\tau(x+\delta) - I'_\tau(x)| \le c2^\tau\delta^\nu \le c2^{-\tau} \le c\delta^{\frac{\nu}{4}}.$$

Finally, combining this bound with (2.13) and (2.14), we conclude that

$$|I'_k(x+\delta) - I'_k(x)| \le c\delta^{\frac{\nu^2}{4}}.$$

Taking $\mu := \frac{\nu^2}{4}$, we finish the proof. □

LEMMA 2.5. *Consider a nonstationary subdivision scheme* $\{S_{a^{[k]}}\}$ *with Laurent polynomials of the form*

$$a^{[k]}(z) = \frac{1}{2}(1 + r_k z)b^{[k]}(z).$$

Suppose that

$$|1 - r_k| \le c2^{-k}, \quad k \ge K \in \mathbb{Z}_+, \tag{2.15}$$

and that the scheme corresponding to $\{S_{b^{[k]}}\}$ *is* $C^{L+\nu}$ *with* $L \in \mathbb{Z}_+$ *and* $\nu \in (0,1)$. *Then* $\{S_{a^{[k]}}\}$ *is* $C^{L+1+\mu}$ *for some* $\mu \in (0,1)$.

Proof. Due to Lemma 2.3, we find that

$$\phi_a = \int_{\mathbb{R}} \phi_b(\cdot - t)h(t)\,dt,$$

where ϕ_a, ϕ_b, and h are the basic limit functions of $\{S_{a^{[k]}}\}$, $\{S_{b^{[k]}}\}$, and $\{S_{1+r_k z}\}$, respectively. Note that h is bounded and $\operatorname{supp}\{h\} = [0,1)$ [8]. It is sufficient to prove the lemma for $\ell = 0$, since

$$D^\ell\phi_a = \int_{\mathbb{R}} D^\ell\phi_b(\cdot - t)h(t)\,dt.$$

To this end, invoking the definition of the function I_k in (2.6), we find that $I_k(x) \to \phi_a(x)$ uniformly as $k \to \infty$. Further, according to Lemma 2.4(a), $\{I'_k\}$ is uniformly convergent, which means that the limit of $\{I'_k\}$ is continuous and that is ϕ'_a. Using this fact, we can conclude from Lemma 2.4(b) that ϕ'_a is Hölder continuous with exponent $\mu > 0$. It completes the proof. □

We now show that Condition A on $\{S_{a^{[k]}}\}$ implies the condition (2.15) for all the factors in the representation (2.3). To prove this we need the following two lemmas. Without loss of generality, we rearrange the set $r_{k,n}$ in (2.3) such that

$$|1-r_{k,n}| = \max\{|1-r_{k,\ell}| : \ell = n,\dots,M\}, \quad n = 1,\dots,M, \tag{2.16}$$

that is, $|1-r_{k,n}| \geq |1-r_{k,n+1}|$. The following lemma shows that if Condition A is satisfied, $|1-r_{k,n}| \leq c2^{-k}$. For this proof, we use the notation

$$\{x_k\} \asymp \{y_k\}$$

for two sequences of nonzero reals if there exist some constants $c_1, c_2 > 0$ such that $c_1 \leq |x_k\, y_k^{-1}| \leq c_2$ for all k.

LEMMA 2.6. *Suppose that Condition A holds for the scheme $\{S_{a^{[k]}}\}$. Then*

$$|1-r_{k,n}| \leq c2^{-k}, \quad k \geq K \in \mathbb{Z}_+. \tag{2.17}$$

Proof. Denote $|1-r_{k,1}| =: \omega_k$. Since $|1-r_{k,n}| \leq \omega_k$ for any $n \leq M$, it is sufficient to show that $\sup_k |2^k\omega_k| \leq c$ for a constant $c > 0$. Now, suppose that $\sup_k |2^k\omega_k| = \infty$, which means that there exists a sequence $\{k_\ell\}$ such that

$$|2^{k_\ell}\omega_{k_\ell}| \leq |2^{k_{\ell+1}}\omega_{k_{\ell+1}}| \to \infty \quad \text{as} \quad k_\ell \to \infty. \tag{2.18}$$

Then, recalling $|1-r_{k,n+1}| \leq |1-r_{k,n}|$, we will derive a contradiction by considering the following two cases.

Case 1. $\{\omega_{k_\ell}\} \asymp \{|1-r_{k_\ell,n}|\}$ for $n = 1,\dots,M$.

In this case, it is clear from (2.3) that $\{a_{k_\ell}(-1)\} \asymp \{\omega_{k_\ell}^M\}$. By Condition A, $|a_{k_\ell}(-1)| \leq c2^{-k_\ell M}$, we get the bound $|2^{k_\ell}\omega_{k_\ell}| \leq c$ for any k_ℓ, in contradiction to (2.18).

Case 2. $\{\omega_{k_\ell}\} \asymp \{|1-r_{k_\ell,n}|\}$ for $n = 1,\dots,s < M$.

That is, there exists a subsequence $\{k_j\} \subset \{k_\ell\}$ such that for any $n > s$, $|1-r_{k_j,n}|\omega_{k_j}^{-1} \to 0$ as $k_j \to \infty$, i.e.,

$$|1-r_{k_j,n}| = o(\omega_{k_j}), \quad n > s. \tag{2.19}$$

Then we use the following lemma.

LEMMA 2.7. *Let*

$$F_{k_j}(z) := \prod_{n=1}^{M} \frac{1}{2}(1+r_{k_j,n}z).$$

Under the condition of Case 2, *we have*

$$\{F_{k_j}^{(M-s)}(-1)\} \asymp \{\omega_{k_j}^s\} \quad \text{and} \quad |F_{k_j}^{(M-s-\ell)}(-1)| = o(\omega_{k_j}^{s+\ell}) \quad \forall \ell > 0.$$

Proof. For the given $s < M$, denote $I_s := \{1,2,\dots,s\}$ and let Λ_s be the collection of all subsets of $\{1,2,\dots,M\} = I_M$ with cardinality s, i.e.,

$$\Lambda_s := \{I \subset I_M \ : \ \#I = s\}.$$

Then,

$$F_{k_j}^{(M-s)}(-1) = \left(\prod_{n\in I_s}(1-r_{k_j,n}) + \sum_{I\in\Lambda_s\setminus I_s}\prod_{n\in I}(1-r_{k_j,n})\right)\left(\frac{1}{2^M}+o(1)\right). \tag{2.20}$$

Since $|1-r_{k_j,n}| \geq |1-r_{k_j,n+1}|$,

$$\left\{\prod_{n\in I_s}|1-r_{k_j,n}|\right\} \asymp \{\omega_{k_j}^s\} \quad \text{and} \quad \prod_{n\in I}|1-r_{k_j,n}| = o(\omega_{k_j}^s).$$

Thus,

$$\{F_{k_j}^{(M-s)}(-1)\} \asymp \{\omega_{k_j}^s\}.$$

In a similar way, we can prove the relation $|F_{k_j}^{(M-s-\ell)}(-1)| = o(\omega_{k_j}^{s+\ell})$ for all $\ell > 0$. ◻

Now, we turn to the proof of Lemma 2.6 in Case 2. It follows from (2.3) that for some suitable constants c_ℓ with $\ell = 0, \ldots, M-s$, we have

$$\begin{aligned}(2.21)\quad a_{k_j}^{(M-s)}(-1) &= \sum_{\ell=0}^{M-s}\binom{2m-s}{\ell} b_{k_j}^{(\ell)}(-1)F_{k_j}^{(M-s-\ell)}(-1)\\ &= b_{k_j}(-1)F_{k_j}^{(M-s)}(-1) + \sum_{\ell=1}^{M-s}\binom{2m-s}{\ell} b_{k_j}^{(\ell)}(-1)F_{k_j}^{(M-s-\ell)}(-1).\end{aligned}$$

Since $b_{k_j}(-1) = c + o(1)$ with a constant $c \neq 0$, identity (2.21) leads to $\{a_{k_j}^{(M-s)}(-1)\} \asymp \{\omega_{k_j}^s\}$ by Lemma 2.7. Also, from Condition A, $|a_{k_j}^{(M-s)}(-1)| \leq c2^{-k_j s}$, yielding $|2^{k_j}\omega_{k_j}| \leq c$ for any k_j, which is a contradiction to (2.18). (Here c is a generic constant.) ◻

We are now ready to provide the main theorem of this section.

THEOREM 2.8 (smoothing factor in nonstationary subdivision schemes). *Consider a nonstationary subdivision scheme $\{S_{a^{[k]}}\}$ satisfying Condition A. If*

$$a^{[k]}(z) = \frac{1}{2}(1+r_k z)c^{[k]}(z), \quad k > K \in \mathbb{Z}_+,$$

where $\{S_{c^{[k]}}\}$ is of compact support and $C^{L+\nu}$ with $L \in \mathbb{Z}_+$ and $\nu \in (0,1)$, then $\{S_{a^{[k]}}\}$ is $C^{L+1+\mu}$ for some $\mu \in (0,1)$.

Proof. From Lemmas 2.6 and 2.5, the proof is immediate. ◻

For interpolatory schemes, we have the stronger result.

THEOREM 2.9. *Let $\{S_{a^{[k]}}\}$ be a nonstationary interpolatory subdivision scheme satisfying Condition A. Assume that $\{S_{a^{[k]}}\}$ is asymptotically equivalent to a stationary subdivision scheme $\{S_a\}$. Then if $\{S_a\}$ is $C^{L+\nu}$ with $L \in \mathbb{Z}_+$ and $\nu \in (0,1)$, $\{S_{a^{[k]}}\}$ is $C^{L+\mu}$ for some $\mu \in (0,1)$.*

Proof. Assume that $\{S_a\}$ is $C^{L+\nu}$. Since S_a is interpolatory, $a(z) = 2^{-L}(1+z)^L c(z)$ with $\{S_c\}$ a C^ν with $\nu \in (0,1)$ [5]. From the fact that $\{S_a\}$ and $\{S_{a^{[k]}}\}$ are asymptotically equivalent, we conclude that $L < M$, and we can write

$$a^{[k]}(z) = \prod_{n=1}^{L}\frac{1}{2}(1+r_{k,n}z)c^{[k]}(z)$$

with $\{S_{c^{[k]}}\}$ asymptotically equivalent to $\{S_c\}$. By [8], the scheme $\{S_{c^{[k]}}\}$ is Hölder continuous with some positive exponent. From Condition A and Lemmas 2.5 and 2.6, we conclude that $\{S_{a^{[k]}}\}$ is $C^{L+\mu}$ with $\mu \in (0,1)$. ◻

In what follows, we use the results of this section to analyze a new family of interpolatory schemes.

3. Subdivision schemes based on Gaussian interpolation and their analysis.

3.1. Construction. Radial basis function (RBF) interpolation is a very strong and convenient tool for interpolation in the multivariate setting [4, 12, 13, 14, 15, 17]. In this section, we apply RBF interpolation in the univariate setting to construct interpolatory subdivision schemes. Given data $(x_j, f(x_j))$, $j = 1, \dots, n$, where $X := \{x_1, \dots, x_N\}$ is a subset of $\mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$, we consider interpolants to the data of the form

$$\mathcal{R}_{f,X}(x) := \sum_{n=1}^{N} \alpha_n G(x - x_n), \tag{3.1}$$

where G is the Gaussian function

$$G(x) = e^{-|x|^2/\lambda^2},$$

with λ a parameter (λ can serve as a shape parameter in the resulting subdivision scheme). The coefficients $\alpha_1, \dots, \alpha_N$ are determined by the interpolation condition

$$\mathcal{R}_{f,X}(x_j) = f(x_j), \quad j = 1, \dots, N. \tag{3.2}$$

It is well known [13] that the linear system (3.2) is nonsingular for any choice of X consisting of distinct points. The interpolant $\mathcal{R}_{f,X}$ in (3.1) has a Lagrange-type representation:

$$\mathcal{R}_{f,X}(x) := \sum_{n=1}^{N} u_n(x) f(x_n), \quad u_n(x_\ell) = \delta_{n,\ell}, \tag{3.3}$$

where u_n are the Lagrange functions from the space $G_X := \operatorname{span}\{G(\cdot - x_1), \dots, G(\cdot - x_N)\}$. The coefficients $u_n(x)$, $n = 1, \dots, N$, can be obtained as the solution of the linear system

$$\sum_{n=1}^{N} u_n(x) G(x_n - x_\ell) = G(x - x_\ell), \quad \ell = 1, \dots, N. \tag{3.4}$$

We study interpolatory subdivision schemes based on interpolation at symmetric $2m$-points to the inserted point. By (3.4), and since $G(x) = G(-x)$, the subdivision schemes considered are nonstationary and uniform in the sense that their refinement rules depend on the level of refinement but are the same everywhere on the same level. Let

$$X_{k,j} := \{(j+n)2^{-k} : n = -m+1, \dots, m\},$$

which is the local set of symmetric $2m$-points around $(j + 2^{-1})2^{-k}$. Then, the value f_{2j+1}^{k+1} is defined by the Gaussian-based interpolation to the data $\{(j+n)2^{-k}, f_{j+n}^k) : n = -m+1, \dots, m\}$, denoted by $\mathcal{R}_{k,j}$. Thus,

$$\begin{aligned} f_{2j+1}^{k+1} &= \mathcal{R}_{k,j}(2^{-k}(j + 2^{-1})) \\ &= \sum_{n=-m+1}^{m} u_n^{[k,j]}(2^{-k}(j+2^{-1})) f^k((j+n)2^{-k}) \end{aligned}$$

with the Lagrange function $u_n^{[k,j]}$ as in (3.4). Here and in what follows, we use the notation

$$X_0 := X_{0,0} := \{-m+1, \dots, m\}. \tag{3.5}$$

It is easy to verify from (3.4) that the $u_n^{[k,j]}(2^{-k}(j+2^{-1}))$ with $n \in X_0$ are independent of the location j. Thus, we can define

$$a_{1-2n}^{[k]} := u_n^{[k,j]}(2^{-k}(j+2^{-1})), \quad j \in \mathbb{Z}, \tag{3.6}$$

and the mask at level k of the $2m$-point Gaussian-based interpolatory subdivision scheme by

$$a_{1-2n}^{[k]} := u_n^{[k,0]}(2^{-k-1}), \quad a_{2n}^{[k]} = \delta_{n,0}, \quad n \in X_0,\ k \in \mathbb{Z}_+. \tag{3.7}$$

Note that by construction,

$$\sum_{n \in X_0} a_{1-2n}^{[k]} G((n-\ell)2^{-k}) = G((2^{-1}-\ell)2^{-k}), \quad \ell \in X_0. \tag{3.8}$$

We denote the nonstationary scheme with mask defined in (3.7) by $\{S^G_{a^{[k]}}\}$. To study the convergence and smoothness of $\{S^G_{a^{[k]}}\}$, we use the results of section 2 and compare $\{S^G_{a^{[k]}}\}$ with the $2m$-point Deslauriers–Dubuc interpolatory subdivision scheme, which we denote by $\{S_a\}$.

The $2m$-point Deslauriers–Dubuc interpolatory subdivision scheme defines the values at the inserted point by using polynomial interpolation of degree $2m-1$ through the symmetric $2m$-points. Define the Lagrange polynomials on the set X_0 in (3.5) by

$$L_n(x) = \prod_{\substack{\ell \neq n \\ \ell \in X_0}} \frac{x-\ell}{n-\ell}, \quad n \in X_0. \tag{3.9}$$

It is obvious that $L_n(\ell) = \delta_{n,\ell}$ with $\ell \in X_0$. Then, the mask of the $2m$-point Deslauriers–Dubuc interpolatory subdivision scheme is given by

$$a_{2n} = \delta_{0,n}, \quad a_{1-2n} := L_n(2^{-1}), \quad n \in X_0. \tag{3.10}$$

One should keep in mind that S_a reproduces polynomials of degree $\leq 2m-1$. In particular, for any $k \in \mathbb{Z}_+$,

$$p(2^{-k-1}) = \sum_{n \in X_0} a_{1-2n} p(n2^{-k}), \quad p \in \Pi_{<2m}, \tag{3.11}$$

where $\Pi_{<n}$ stands for the space consisting of all univariate algebraic polynomials of degree less than n.

3.2. Analysis of convergence. The goal of this section is to prove that the $2m$-point Gaussian-based interpolatory subdivision scheme $\{S^G_{a^{[k]}}\}$ is asymptotically equivalent to the $2m$-point Deslauriers–Dubuc interpolatory scheme $\{S_a\}$, which implies that $\{S^G_{a^{[k]}}\}$ is convergent [8].

THEOREM 3.1. *Let $\{a_n^{[k]}\}$ be the mask at level k of $\{S^G_{a^{[k]}}\}$, and let $\{a_n\}$ be the mask of $\{S_a\}$. Then, there exists a constant $c_{2m} > 0$ such that*

$$\max_{n \in \mathbb{Z}} |a_n^{[k]} - a_n| \leq c_{2m} 2^{-2k}, \quad k \geq K \in \mathbb{Z}_+.$$

Proof. Since $a_{2n} = a^{[k]}_{2n} = \delta_{n,0}$, we need to estimate only the difference $a_{1-2n} - a^{[k]}_{1-2n}$. Recall from (3.7) that $a^{[k]}_{1-2n} = u^{[k,0]}_n(2^{-k-1})$, $n \in X_0$, with $u^{[k,0]}_n$ the Lagrange function of the Gaussian-based interpolation on $X_{k,0} = \{\ell 2^{-k} : \ell \in X_0\}$, satisfying

$$u^{[k,0]}_n(2^{-k}\ell) = \delta_{n,\ell}, \quad \ell \in X_0. \tag{3.12}$$

Further, since $u^{[k,0]}_n(x) \in \operatorname{span}\{G(\cdot - \ell 2^{-k}) : \ell \in X_0\}$, there exist constants $\alpha^{[k]}_\ell$, $\ell \in X_0$, such that

$$u_{k,n}(x) := u^{[k,0]}_n(2^{-k}x) = \sum_{\ell \in X_0} \alpha^{[k]}_\ell G(2^{-k}(x-\ell)), \tag{3.13}$$

yielding $u_{k,n}(\ell) = \delta_{n,\ell}$ for any $\ell \in X_0$. Thus, $u_{k,n}(x)$ can be considered as the RBF interpolant to the data $\{\delta_{n,\ell} : \ell \in X_0\}$ on X_0 by $G(2^{-k}\cdot)$. On the other hand, the mask of Deslauriers–Dubuc scheme is given by $a_{1-2n} = L_n(2^{-1})$, where the function $L_n(x)$ is also a polynomial interpolant to the data $\{\delta_{n,\ell} : \ell \in X_0\}$ on X_0, which means

$$u_{k,n}(\ell) = L_n(\ell) = \delta_{n,\ell}, \quad \ell \in X_0.$$

Recently, it was proved in [16] that the Gaussian interpolant of the form $u_{k,n}(x)$ converges uniformly to the polynomial interpolant $L_n(x)$ as $k \to \infty$, with the convergence rate $O(2^{-2k})$. In particular,

$$|u_{k,n}(2^{-1}) - L_n(2^{-1})| = O(2^{-2k}).$$

Thus, since $a^{[k]}_{1-2n} = u_{k,n}(2^{-1})$ and $a_{1-2n} = L_n(2^{-1})$, we arrive at the required conclusion. □

COROLLARY 3.2. *The scheme $\{S^G_{a^{[k]}}\}$ is convergent and is C^0.*

3.3. Analysis of smoothness. First, we give a simple proof of C^1 smoothness based on a result from [8].

Result A. The scheme $\{S_{a^{[k]}}\}$ is in C^γ if a scheme $\{S_a\}$ is in C^γ and

$$\sum_{k \in \mathbb{Z}_+} 2^{\gamma k} \|S_a - S_{a^{[k]}}\|_\infty < \infty. \tag{3.14}$$

THEOREM 3.3. *Let $\{S^G_{a^{[k]}}\}$ be the $2m$-point Gaussian-based interpolatory subdivision scheme. Then, if $m \geq 2$, $\{S^G_{a^{[k]}}\}$ is at least C^1.*

Proof. Let S_a be the $2m$-point Deslauriers–Dubuc interpolatory scheme; it is clear from Theorem 3.1 that

$$\sum_{k \in \mathbb{Z}_+} 2^k \|S_a - S^G_{a^{[k]}}\|_\infty < \infty.$$

Since S_a is at least C^1 for $m \geq 2$, the C^1 smoothness of $\{S^G_{a^{[k]}}\}$ is an immediate consequence of Result A. □

Next, we show that if the $2m$-point Deslauriers–Dubuc interpolatory scheme $\{S_a\}$ is $C^{L+\nu}$ with $L \in \mathbb{Z}_+$ and $\nu \in (0,1)$, then the nonstationary $2m$-point Gaussian-based interpolatory subdivision scheme $\{S^G_{a^{[k]}}\}$ is $C^{L+\mu}$ for some $\mu \in (0,1)$. It implies that both $2m$-point schemes have the same integer smoothness as the $2m$-point Deslauriers–Dubuc interpolatory scheme. This is proved by Theorem 2.9 and by verifying first that $\{S^G_{a^{[k]}}\}$ satisfies Condition A of section 2. Our proof relies on the following property of the Gaussian function [11]:

$$\det\big(G^{(\ell+n)}(0)\big) \neq 0, \quad \ell, n = 0, \ldots, 2m-1, \tag{3.15}$$

and on the three auxiliary lemmas.

LEMMA 3.4. *Let $T_{G^{(\ell)}}$ be the Taylor polynomial of $G^{(\ell)}$ of degree $2m-1$ around zero, i.e.,*

$$T_{G^{(\ell)}}(x) := \sum_{n=0}^{2m-1} \frac{x^n}{n!}\, G^{(\ell+n)}(0). \tag{3.16}$$

Then, $T_{G^{(\ell)}}$, $\ell = 0, \ldots, 2m-1$, are linearly independent.

Proof. It is sufficient to prove that for any distinct points $t_0, \ldots, t_{2m-1}$, the $2m \times 2m$ matrix $\mathbf{T}$ with entries

$$\mathbf{T}(\ell, n) = T_{G^{(\ell)}}(t_n), \quad \ell, n = 0, \ldots, 2m-1,$$

is nonsingular. We see that the matrix $\mathbf{T}$ can be decomposed as

$$\mathbf{T} = \mathbf{B} \cdot \mathbf{V},$$

where

$$\mathbf{B}(\ell, n) = G^{(\ell+n)}(0) \quad \text{and} \quad \mathbf{V}(\ell, n) = t_n^\ell / \ell!.$$

Since both matrices $\mathbf{B}$ and $\mathbf{V}$ are invertible, the nonsingularity of $\mathbf{T}$ is immediate. $\square$

To prove the next lemma, we recall that the $(n-1)$th order *divided difference* of a function $f \in C^{n-1}$ at the points $(-m+1), \ldots, (-m+n)$ is given by

$$(n-1)!f[-m+1, \ldots, -m+n] = \sum_{\alpha=1}^{n-1} c_{n,\alpha} f(-m+\alpha) = f^{(n-1)}(\xi) \tag{3.17}$$

with $\xi \in [-m+1, -m+n]$, and where

$$c_{n,\alpha} := (n-1)! \prod_{\substack{j \neq \alpha \\ j=1}}^{n} \frac{1}{\alpha - j}, \quad \alpha = 1, \ldots, n. \tag{3.18}$$

LEMMA 3.5. *Let $\mathbf{P}_k$ be the $2m \times 2m$ matrix with entries*

$$\mathbf{P}_k(\ell, n) := \mathbf{P}_{k,x}(\ell, n) := G^{(\ell)}(x - (n-m+1)2^{-k}), \tag{3.19}$$

where $\ell, n = 0, \ldots, 2m-1$. Then there exist $\eta > 0$ and $K \in \mathbb{Z}_+$ such that for any $x \in [-\eta, \eta]$,

$$\det \mathbf{P}_k = O(2^{-km(2m-1)}) \quad \textit{and} \quad \|\mathbf{P}_k^{-1}\|_\infty = O(2^{(2m-1)k}), \quad k \geq K.$$

Here $\|\mathbf{A}\|_\infty$ indicates the ∞-norm of the matrix $\mathbf{A}$.

Proof. Denote by $\mathbf{p}_n$, $n = 1, \ldots, 2m$, the column vectors of the matrix $\mathbf{P}_k$. Since the determinant of a matrix is invariant under elementary column operations, we perform the following column operations:

$$\mathbf{p}'_n := \mathbf{p}_n + \sum_{\alpha=1}^{n-1} c_{n,\alpha}\mathbf{p}_\alpha, \quad n = 2m, 2m-1, \ldots, 1,$$

with $c_{n,\alpha}$ defined as in (3.17). Defining $\mathbf{P}'_k$ to be the matrix with columns $(\mathbf{p}'_1, \ldots, \mathbf{p}'_{2m})$, we observe that $\det(\mathbf{P}_k) = \det(\mathbf{P}'_k)$. Further, applying (3.17), we get

$$\mathbf{P}'_k(\ell, n) = 2^{-nk} G^{(\ell+n)}(x - \xi_{\ell,n}2^{-k}), \tag{3.20}$$

with $0 \le \ell, n \le 2m-1$, and $\xi_{\ell,n} \in [-m+1, -m+1+n]$. Thus, from (3.15), we can deduce that there exist $\eta > 0$ and $K \in \mathbb{Z}_+$ such that $\det \mathbf{P}_k = O(2^{-km(2m-1)}) \neq 0$ for any $x \in [-\eta, \eta]$ and $k \ge K$. Further, a direct calculation from (3.20) easily leads to the estimate $\|\mathbf{P}_k^{-1}\|_\infty = O(2^{(2m-1)k})$ as $k \to \infty$. ∎

For any $\beta = 0, \ldots, 2m-1$, define the function

$$\Phi_{k,\beta}(x) := \sum_{n \in X_0} g_\beta^{[k]}(n) G(x - 2^{-k}n) \tag{3.21}$$

so that the coefficient vector $\mathbf{g}_\beta^{[k]} := (g_\beta^{[k]}(n) : n \in X_0)$ is obtained by solving the linear system

$$\Phi_{k,\beta}^{(\ell)}(2^{-k-1}) = \delta_{\beta,\ell}(-1)^\ell \ell! \tag{3.22}$$

which can be written in the matrix form

$$\mathbf{P}_k \cdot \mathbf{g}_\beta^{[k]} = \mathbf{c},$$

where $\mathbf{P}_k = \mathbf{P}_{k,2^{-k-1}}$ with $\mathbf{P}_{k,x}$ as defined in (3.19) and $\mathbf{c}(\ell) := \delta_{\beta,\ell}(-1)^\ell \ell!$ with $\ell = 0, \ldots, 2m-1$. The following estimates are central to the proof that $\{S_{a^{[k]}}^G\}$ satisfies Condition A.

LEMMA 3.6. *For all* $\beta = 0, \ldots, 2m-1$,

$$\begin{aligned} &\text{(i) } \|\mathbf{g}_\beta\|_\infty = O(2^{k(2m-1)}) \quad \textit{as} \quad k \to \infty, \\ &\text{(ii) } |\Phi_{k,\beta}^{(\ell)}(2^{-k-1})| \le c, \quad \ell = 2m, \ldots, 4m-2. \end{aligned} \tag{3.23}$$

Proof. Since $\mathbf{g}_\beta^{[k]} = \mathbf{P}_k^{-1} \cdot \mathbf{c}$, the estimate $\|\mathbf{g}_\beta^{[k]}\|_\infty = O(2^{k(2m-1)})$ follows immediately from Lemma 3.5. Next, recall that $T_{G^{(\ell)}}$ indicates the Taylor polynomial of degree $2m-1$ of $G^{(\ell)}$ around zero with $\ell = 2m, \ldots, 4m-2$. By Lemma 3.4, there exist some suitable constants $\gamma_{\ell,0}, \ldots, \gamma_{\ell,2m-1}$ such that $T_{G^{(\ell)}} =: \sum_{\alpha=0}^{2m-1} \gamma_{\ell,\alpha} T_{G^{(\alpha)}}$. Thus, we get from (3.21),

$$\begin{aligned} \Phi_{k,\beta}^{(\ell)}(2^{-k-1}) &= \sum_{n \in X_0} g_\beta^{[k]}(n) G^{(\ell)}(2^{-k-1} - n2^{-k}) \\ &= \sum_{n \in X_0} g_\beta^{[k]}(n) \big(T_{G^{(\ell)}}(2^{-k-1} - n2^{-k}) + O(2^{-2mk})\big) \\ &= \sum_{\alpha=0}^{2m-1} \gamma_{\ell,\alpha} \sum_{n \in X_0} g_\beta^{[k]}(n) \big[T_{G^{(\alpha)}}(2^{-k-1} - n2^{-k}) + O(2^{-2mk})\big]. \end{aligned}$$

Note that

$$T_{G^{(\alpha)}}(2^{-k-1} - n2^{-k}) = G^{(\alpha)}(2^{-k-1} - n2^{-k}) + O(2^{-2mk}), \quad \alpha = 0, \dots, 2m-1.$$

Applying this identity to (3.21), we get, in view of (3.23),

$$\begin{aligned}\left|\sum_{n\in X_0} g_\beta^{[k]}(n) T_{G^{(\ell)}},(2^{-k-1} - n2^{-k})\right| &\le \left|\Phi_{k,\beta}^{(\ell)}(2^{-k-1})\right| + O(2^{-k}) \\ &\le \ell! + O(2^{-k}),\end{aligned}$$

as a consequence of (3.22). Also, by (i), the property $|\Phi_{k,\beta}^{(\ell)}(2^{-k-1})| \le c$ for $\ell = 2m, \dots, 4m-2$ is obvious. ∎

Now, we are ready to prove that $\{S^G_{a^{[k]}}\}$ satisfies Condition A with $M = 2m$.

THEOREM 3.7. *Let $a^{[k]}(z) = \sum_{n\in\mathbb{Z}} a_n^{[k]} z^n$ be the Laurent polynomial at level k associated with $\{S^G_{a^{[k]}}\}$. Then, for any $\beta = 0, \dots, 2m-1$, we have*

$$|D^\beta a^{[k]}(-1)| \le c2^{-k(2m-\beta)}, \quad k \ge K,$$

for some $K \in \mathbb{Z}_+$.

Proof. Since

$$D^\beta a^{[k]}(-1) = \sum_{\ell=0}^{\beta} \gamma_{\beta,\ell} \sum_{n\in\mathbb{Z}} a_n^{[k]} n^\ell (-1)^n$$

for some constants $\gamma_{\beta,\ell}$, $\ell = 0, \dots, \beta$, it is sufficient to prove that for any $\beta = 0, \dots, 2m-1$,

$$s_\beta := \sum_{n\in\mathbb{Z}} (-1)^n n^\beta a_n^{[k]} = O(2^{-k(2m-\beta)}), \quad k \to \infty,$$

in order to conclude Condition A. Recalling that $a_{2n}^{[k]} = \delta_{n,0}$, observe that

$$2^{-\beta(k+1)} s_\beta = \delta_{\beta,0} - \sum_{n\in X_0} a_{1-2n}^{[k]} \big((-n + 2^{-1})2^{-k}\big)^\beta. \tag{3.24}$$

Invoking (3.8) and (3.22), we get

$$\delta_{\beta,0} = \Phi_{k,\beta}(2^{-k-1}) = \sum_{n\in X_0} a_{1-2n}^{[k]} \Phi_{k,\beta}(n2^{-k}). \tag{3.25}$$

This together with (3.24) lead to

$$2^{-\beta(k+1)} s_\beta = \sum_{n\in X_0} a_{1-2n}^{[k]} \bigg(\Phi_{k,\beta}(n2^{-k}) - ((-n + 2^{-1})2^{-k})^\beta\bigg). \tag{3.26}$$

In the following, we replace $\Phi_{k,\beta}(n2^{-k})$ with its Taylor polynomial of degree $4m-2$ plus the remainder term. The Taylor expansion of $\Phi_{k,\beta}$ around 2^{-k-1} of degree $4m-2$ is

$$\Phi_{k,\beta}(n2^{-k}) = \sum_{\ell=0}^{4m-2} \big((n - 2^{-1})2^{-k}\big)^\ell \frac{\Phi_{k,\beta}^{(\ell)}(2^{-k-1})}{\ell!} + R_{k,\beta,m}, \tag{3.27}$$

where the remainder $R_{k,\beta,m}$ is given by

$$R_{\Phi_\beta,4m-1} = \big((n-2^{-1})2^{-k}\big)^{4m-1}\frac{\Phi_{k,\beta}^{(4m-1)}(\xi 2^{-k})}{(4m-1)!}$$

with ξ a point between 2^{-1} and n. Noting that by (3.22)

$$\sum_{\ell=0}^{2m-1}\big((n-2^{-1})2^{-k}\big)^{\ell}\frac{\Phi_{k,\beta}^{(\ell)}(2^{-k-1})}{\ell!} = \Big((-n+2^{-1})2^{-k}\Big)^{\beta},$$

we get from (3.25) and (3.26)

$$2^{-\beta(k+1)}s_\beta = \sum_{n\in X_0} a_{1-2n}^{[k]}\left(\sum_{\ell=2m}^{4m-2}\big((n-2^{-1})2^{-k}\big)^{\ell}\frac{\Phi_{k,\beta}^{(\ell)}(2^{-k-1})}{\ell!} + R_{k,\beta,m}\right).$$

By Lemma 3.6, $|\Phi_{k,\beta}^{(\ell)}(2^{-k-1})| \le c$ for any $\ell = 2m,\dots,4m-2$. Consequently,

$$\left|\sum_{n\in\mathbb{Z}}(-1)^n n^\beta a_n^{[k]}\right| \le c2^{-k(2m-\beta)}\Big(1+2^{-k(2m-1)}\|\Phi_{k,\beta}^{(4m-1)}\|_{L_\infty[-\eta,\eta]}\Big). \tag{3.28}$$

Since $\|\mathbf{g}_\beta\|_\infty = O(2^{k(2m-1)})$ and $G^{(4m-1)}$ is bounded,

$$2^{-k(2m-1)}\|\Phi_\beta^{(4m-1)}\|_{L_\infty[-\eta,\eta]} \le c,$$

which completes the proof of the theorem. ◻

We are now ready to state and prove the main theorem of this section.

THEOREM 3.8. *If the $2m$-point Deslauriers–Dubuc interpolatory scheme $\{S_a\}$ is $C^{L+\nu}$ with $L\in\mathbb{Z}_+$ and $\nu\in(0,1)$, then the nonstationary $2m$-point Gaussian-based interpolatory subdivision scheme $\{S_{a^{[k]}}^G\}$ is $C^{L+\mu}$ for some $\mu\in(0,1)$.*

Proof. Due to Theorem 3.1, both $2m$-point schemes $\{S_{a^{[k]}}^G\}$ and $\{S_a\}$ are asymptotically equivalent. As a consequence of Theorems 3.7 and 2.9, this theorem is immediate. ◻

COROLLARY 3.9. *The nonstationary $2m$-point Gaussian-based interpolatory subdivision scheme has the same integer smoothness as the $2m$-point Deslauriers–Dubuc interpolatory scheme.*

Remark. In this section, the Gaussian function has been used. We believe that the same results can be obtained by using other smooth RBFs such as multiquadrics.

4. Examples. In this section, we illustrate the performance of the 4-point Gaussian-based interpolatory subdivision schemes with some numerical examples. Recall that the Gaussian function is of the form

$$G(x) = e^{-|x|^2/\lambda^2}, \quad \lambda > 0.$$

Here λ serves as a shape parameter. Having tried several alternatives for the parameter λ, we found out that good choices of λ are in the range $0 < \lambda^{-1} \le 1.0$. The solid curves in Figure 4.1 are generated by the 4-point Gaussian-based interpolatory subdivision scheme using the parameter $\lambda^{-1} = .5$. The dotted curves in Figure 4.1 are generated by the 4-point Deslauriers–Dubuc scheme. It is known that the 4-point Deslauriers–Dubuc scheme has the smoothness C^1. Hence, due to Theorem 3.8, the

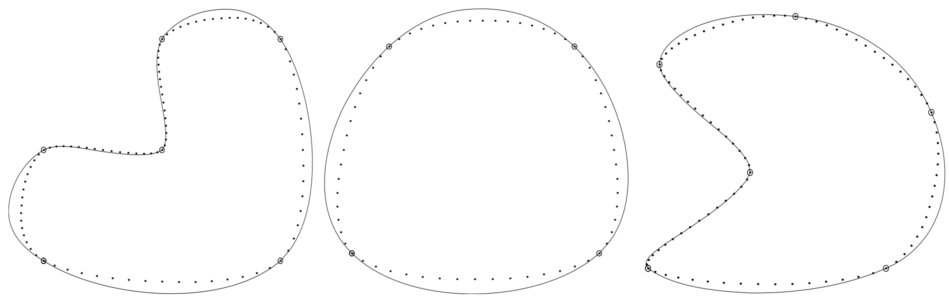


FIG. 4.1. *Interpolating curves generated by the* 4*-point Gaussian-based interpolation with* $\lambda^{-1} =$ .5 *(solid lines) and the* 4*-point DD scheme (dotted lines).*

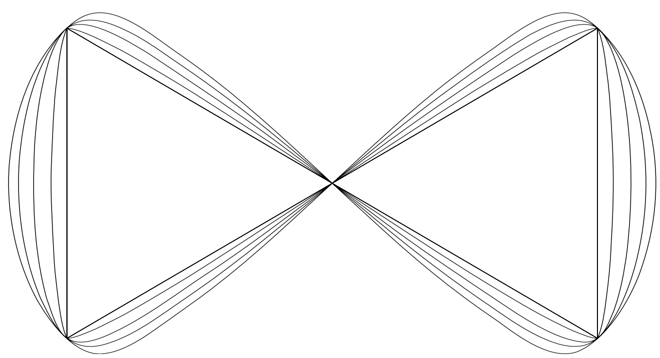


FIG. 4.2. *The effect of the tension parameters in the* 4*-point Gaussian-based scheme and the classical* 4*-point scheme.*

4-point Gaussian-based interpolatory subdivision scheme is also C^1. Also, Figure 4.2 compares the Gaussian-based 4-point scheme to the classical 4-point scheme to the mask $[-\theta, 1/2 + \theta, 1/2 + \theta, -\theta]$. In Figure 4.2, the curves, from the closer to the control polygon and outward, correspond to the values of the tension parameters $\theta = 0.03, 1/16$, and $\lambda^{-1} = 0.5, 0.8$.

In addition, the Gaussian-based schemes have an advantage over polynomial-based schemes, especially in signal processing. The following example shows the superiority of the 4-point Gaussian-based scheme over the 4-point Deslauriers–Dubuc scheme. We approximate oscillatory signals of the form [3]

$$f(t) = \cos(2\pi F t + \beta \sin(2\pi F_s t)).$$

We choose $F = 0.5$, $\beta = 5.75$, $F_s = 0.0062$ and use 311 data samples in the domain $[0, 30]$ as initial data for the subdivision. The solid curve in Figure 4.3 indicates the approximation errors by the 4-point Gaussian-based schemes as a function of λ. The dotted line is the error by the Deslauriers–Dubuc 4-point scheme. We find that by choosing a suitable parameter λ in the Gaussian $g(x) := e^{-x^2/\lambda^2}$ (around $\lambda = 2\pi$ in this example), the 4-point Gaussian-based scheme provides much better accuracy than the polynomial-based 4-point scheme. For this reason, a future project would be to find an algorithm for choosing the appropriate λ for a given signal.

We regard the study of the 4-point Gaussian-based scheme as a first step towards the design and analysis of interpolatory bivariate Gaussian-based nonstationary

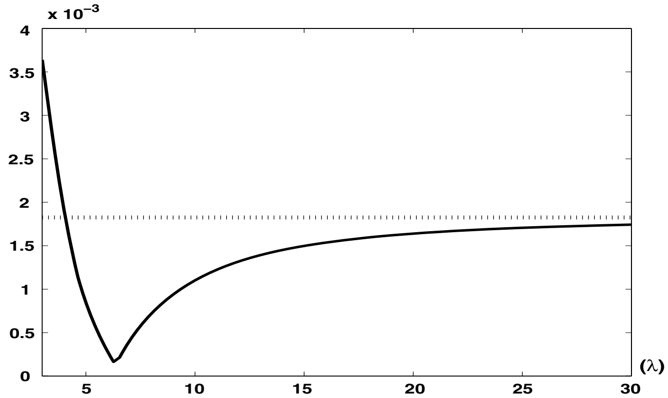


FIG. 4.3. *Approximation errors by the* 4*-point Gaussian-based scheme as a function of* λ*. The dotted line indicates the error by the* 4*-point Deslauriers–Dubuc scheme.*

schemes extending the stationary butterfly scheme. However, this requires much heavier analysis, especially at extraordinary points.

Acknowledgments. We are grateful to the anonymous referees for their helpful suggestions. The third author also would like to thank Tel-Aviv University for its hospitality during his visit as a postdoctoral research associate from October 2000 to July 2001. We also thank Yaron Lipman for suggesting a shorter proof of Lemma 2.4(b).

REFERENCES

[1] A. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary Subdivision*, Mem. Amer. Math. Soc., 93 (1991), no. 453.
[2] A. Cohen and N. Dyn, *Nonstationary subdivision schemes and multiresolution analysis*, SIAM J. Math. Anal., 27 (1996), pp. 1745–1769.
[3] K. Clarke and D. Hess, *Communication Circuits: Analysis and Design*, Addison Wesley, Reading, MA, 1978.
[4] N. Dyn, *Interpolation and Approximation by Radial and Related Functions*, in Approximation Theory VI, C. K. Chui, L. L. Schumaker, and J. Ward, eds., Academic Press, New York, 1989, pp. 211–234.
[5] N. Dyn, *Subdivision Schemes in Computer-Aided Geometric Design*, in Advances in Numerical Analysis, Vol. II: Wavelets, Subdivision Algorithms and Radial Basis Functions, W. A. Light, ed., Oxford University Press, New York, 1992, pp. 36–104.
[6] G. Deslauriers and S. Dubuc, *Symmetric iterative interpolation*, Constr. Approx., 5 (1989), pp. 49–68.
[7] N. Dyn, J. A. Gregory, and D. Levin, *A four-point interpolatory subdivision scheme for curve design*, Comput. Aided Geom. Design, 4 (1987), pp. 257–268.
[8] N. Dyn and D. Levin, *Analysis of asymptotically equivalent binary subdivision schemes*, J. Math. Anal. Appl., 193 (1995), pp. 594–621.
[9] N. Dyn and D. Levin, *The subdivision experience*, in Curves and Surfaces II, P. J. Laurent, A. Le Méhauté, and L. L. Schumaker, eds., AKPeters, Boston, 1991, pp. 1–17.
[10] N. Dyn, D. Levin, and A. Luzzatto, *Exponential reproducing subdivision scheme*, Found. Comput. Math., 3 (2003), pp. 187–206.
[11] S. Karlin and W. J. Studden, *Tchebycheff Systems: With Applications in Analysis and Statistics*, Interscience Publishers, New York, 1966.
[12] C. A. Micchelli, *Interpolation of scattered data: Distances, matrices, and conditionally positive functions*, Constr. Approx., 2 (1986), pp. 11–22.

[13] M. J. D. Powell, *The theory of radial basis functions approximation in* 1990, in Advances in Numerical Analysis, Vol. II: Wavelets, Subdivision Algorithms and Radial Basis Functions, W. A. Light, ed., Oxford University Press, New York, 1992, pp. 105–210.
[14] W. R. Madych and S. A. Nelson, *Multivariate interpolation and conditionally positive functions* I, Approx. Theory Appl., 4 (1988), pp. 77–89.
[15] W. R. Madych and S. A. Nelson, *Multivariate interpolation and conditionally positive functions* II, Math. Comp., 54 (1990), pp. 211–230.
[16] R. Schaback, *Multivariate interpolation by polynomials and radial basis functions*, Constr. Approx., 21 (2005), pp. 293–317.
[17] Z. Wu and R. Schaback, *Local error estimates for radial basis function interpolation of scattered data*, IMA J. Numer. Anal., 13 (1993), pp. 13–27.
[18] J. Warren and H. Weimer, *Subdivision Methods for Geometric Design*, Academic Press, New York, 2002.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 489–506

ELLIPTIC DIFFERENTIAL EQUATIONS WITH COEFFICIENTS MEASURABLE WITH RESPECT TO ONE VARIABLE AND VMO WITH RESPECT TO THE OTHERS*

DOYOON KIM† AND N. V. KRYLOV†

Abstract. We prove the unique solvability of second order elliptic equations in nondivergence form in Sobolev spaces. The coefficients of the second order terms are measurable in one variable and VMO in other variables. From this result, we obtain the weak uniqueness of the Martingale problem associated with the elliptic equations.

Key words. second order equations, vanishing mean oscillation, Martingale problem

AMS subject classifications. 35J15, 60J60

DOI. 10.1137/050646913

1. Introduction.

We study the L_p-theory of the elliptic differential equation

$$a^{jk}(x)u_{x^jx^k}(x) + b^j(x)u_{x^j}(x) + c(x)u(x) = f(x) \quad \text{in} \quad \mathbb{R}^d, \tag{1.1}$$

where $a^{jk}(x)$ are allowed to be only measurable with respect to one coordinate, say $x^1 \in \mathbb{R}$, where $x = (x^1, x') \in \mathbb{R}^d$, $x' \in \mathbb{R}^{d-1}$.

It is well known that if the coefficients a^{jk} are only measurable, then there could not exist a *unique* solution to the above equation even in a very generalized sense (see [11, 13]). We are interested in more regular solutions. In 1967 Ural'tseva (see [7] or the original paper [17]) constructed an example of an equation in $\mathbb{R}^d$ for $d \geq 3$ with the coefficients depending only on the first two coordinates for which there is no unique solvability in W_p^2 with $p \geq d$ (for any $d \geq 3$ and $p \in (1, d)$ this was known before).

Thus in order to have the unique solvability of the equation in W_p^2, we have to impose some (regularity) conditions on the coefficients a^{jk}. The most classical case is when a^{jk} are uniformly continuous. We can also have piecewise continuous or VMO coefficients. For details, see [1, 2, 4, 6, 9, 10].

In this paper, we show that there exists a unique solution to the above equation in W_p^2, $p \in (2, \infty)$, under the assumption that $a^{jk}(x^1, x')$ are measurable in $x^1 \in \mathbb{R}$ and VMO in $x' \in \mathbb{R}^{d-1}$; see Assumptions 2.1 and 2.2. If the coefficients a^{jk} are independent of $x' \in \mathbb{R}^{d-1}$ (more generally, uniformly continuous in $x' \in \mathbb{R}^{d-1}$; see Remark 2.6), then the equation is uniquely solvable in W_2^2 as well. In addition, we show that one can easily solve the equation with the Dirichlet, Neumann, or oblique derivative boundary condition in a half space, say $\mathbb{R}^d_+ = \{(x^1, x') : x^1 > 0, x' \in \mathbb{R}^{d-1}\}$, using the results for equations in the whole space.

The class of coefficients we are dealing with is considerably more general than those previously known, as long as $p \in [2, \infty)$. It actually contains almost all types of discontinuous coefficients that have been investigated so far. For example, it contains

*Received by the editors December 6, 2005; accepted for publication (in revised form) January 9, 2007; published electronically July 4, 2007. The work of the second author was partially supported by NSF grant DMS-0140405.
http://www.siam.org/journals/sima/39-2/64691.html
†School of Mathematics, University of Minnesota, 127 Vincent Hall, Minneapolis, MN 55455 (dykim@math.umn.edu, krylov@math.umn.edu).

the class of piecewise continuous coefficients investigated in [4, 8, 9]. It also contains VMO coefficients with which elliptic equations were investigated in [1, 2, 6]; see also the monograph [10], which treats elliptic and parabolic equations with discontinuous coefficients including oblique derivative problems with VMO coefficients. Although we slightly touch on the oblique derivative problem, we do not say anything about many important issues of equations with VMO coefficients, which are discussed, for instance, in [14, 15, 12].

The highlight of our assumptions on the coefficients a^{jk} would be the following: no assumptions on the regularity of the coefficients with respect to one variable as far as they are uniformly bounded and elliptic. Having only measurable coefficients (as functions of $x^1 \in \mathbb{R}$), we obtain the L_2-estimate for the equation by using the usual Fourier transforms. Based upon this estimate, we establish the L_p-estimate, $p \in (2, \infty)$, using the approach initiated by the second author of this paper (for example, see [6]). In this approach we make use of a pointwise estimate of sharp functions of second order derivatives of the solution. As noted in [6], thanks to this method, we do not need any integral representations of the solution nor commutators, which were used, for example, in [1, 2]. Especially, we deal with VMO coefficients in a rather straightforward manner.

One good motivation to consider the above equation in the whole space is to prove weak uniqueness of stochastic processes associated with the elliptic equation. As is shown in [6, 16], we can say that weak uniqueness of the processes holds true once we find a unique solution of the elliptic equation in W^2_p, $p \geq d$. More details are in [6, 16].

The paper is organized as follows. In section 2 we state our main results. The unique solvability of the equation in W^2_2 is investigated in section 3. In section 4, we present some auxiliary results which are used in section 5, where we finally prove the W^2_p-estimate $p \in (2, \infty)$ for the equation.

2. Main results. We are considering the elliptic differential equation (1.1), where the coefficients a^{jk}, b^j, and c satisfy the assumptions below.

ASSUMPTION 2.1. *The coefficients a^{jk}, b^j, and c are measurable functions defined on $\mathbb{R}^d$, $a^{jk} = a^{kj}$. There exist positive constants $\delta \in (0, 1)$ and K such that*

$$|b^j(x)| \leq K, \qquad |c(x)| \leq K,$$

$$\delta|\vartheta|^2 \leq \sum_{j,k=1}^{d} a^{jk}(x)\vartheta^j\vartheta^k \leq \delta^{-1}|\vartheta|^2$$

for any $x \in \mathbb{R}^d$ and $\vartheta \in \mathbb{R}^d$.

To state another assumption on the coefficients, especially $a = (a^{jk})$, we introduce some notations. Let $B'_r(x') = \{y' \in \mathbb{R}^{d-1} : |x' - y'| < r\}$ and $Q_r(x) = Q_r(x^1, x') = (x^1 - r, x^1 + r) \times B'_r(x')$. Denote

$$\operatorname{osc}_{x'}(a, Q_r(x)) = r^{-1}|B'_r|^{-2} \int_{x^1-r}^{x^1+r} \int_{y', z' \in B'_r(x')} |a(t, y') - a(t, z')| \, dy' \, dz' \, dt,$$

$$a^{\#(x')}_R = \sup_{x \in \mathbb{R}^d} \sup_{r \leq R} \operatorname{osc}_{x'}(a, Q_r(x)),$$

where $|B'_r|$ is the $d-1$-dimensional volume of $B'_r(0)$. We write $a \in VMO_{x'}$ if

$$\lim_{R\to 0} a_R^{\#(x')} = 0.$$

We see that $a \in VMO_{x'}$ if a is independent of x'.

Assumption 2.2. *There is a continuous function $\omega(t)$ defined on $[0,\infty)$ such that $\omega(0) = 0$ and $a_R^{\#(x')} \le \omega(R)$ for all $R \in [0,\infty)$.*

Remark 2.3. It will be seen from our proofs that in Assumption 2.2 the requirement that $\omega(0) = 0$ can be replaced with $\omega(0) \le (4N_1)^{-\nu(d+2)}$, where $N_1 = N_1(d,\delta,p)$ and $\nu = \nu(p)$ are the constants entering (5.6).

As usual, we mean by $W_p^k(\mathbb{R}^d)$, $k = 0,1,\dots$, the Sobolev spaces on $\mathbb{R}^d$. Set $W_p^k = W_p^k(\mathbb{R}^d)$, $L_p = L_p(\mathbb{R}^d)$, and

$$Lu(x) = a^{jk}(x)u_{x^jx^k}(x) + b^j(x)u_{x^j}(x) + c(x)u(x).$$

Here are our main results.

Theorem 2.4. *Let $p \in (2,\infty)$. Then there exists a constant $\lambda_0 \ge 0$, depending only on d, δ, K, p, and the function ω, such that, for any $\lambda > \lambda_0$ and $f \in L_p$, there exists a unique $u \in W_p^2$ satisfying $Lu - \lambda u = f$.*

Furthermore, there is a constant N, depending only on d, δ, K, p, and the function ω such that, for any $\lambda \ge \lambda_0$ and $u \in W_p^2$,

$$\lambda\|u\|_{L_p} + \sqrt{\lambda}\|u_x\|_{L_p} + \|u_{xx}\|_{L_p} \le N\|Lu - \lambda u\|_{L_p}.$$

This theorem obviously covers the case in which the coefficients a^{jk} are independent of $x' \in \mathbb{R}^{d-1}$. However, in that case we can allow $p = 2$, which is detailed in the theorem below. Throughout the paper, we write $N = N(d,\dots)$ if N is a constant depending only on $d,\dots$.

The following theorem can be basically found in [3]. We give it a different proof that seems to be somewhat shorter and more general.

Theorem 2.5. *Let the coefficients a^{jk} be independent of $x' \in \mathbb{R}^{d-1}$. Then there exists a constant $\lambda_0 = \lambda_0(d,\delta,K) \ge 0$ such that, for any $\lambda > \lambda_0$ and $f \in L_2$, there exists a unique $u \in W_2^2$ satisfying $Lu - \lambda u = f$.*

In addition, there is a constant $N = N(d,\delta,K)$ such that, for any $\lambda \ge \lambda_0$ and $u \in W_2^2$,

$$\lambda\|u\|_{L_2} + \sqrt{\lambda}\|u_x\|_{L_2} + \|u_{xx}\|_{L_2} \le N\|Lu - \lambda u\|_{L_2}. \tag{2.1}$$

Remark 2.6. Theorem 2.4 leads to the weak uniqueness of solutions of stochastic differential equations associated with the operator L. For details, see [16, 6]. Theorem 2.5 clearly remains true under the assumption that $a^{jk}(x^1,x')$ are uniformly continuous as functions of $x' \in \mathbb{R}^{d-1}$ uniformly in $x^1 \in \mathbb{R}$.

Three more results deal with the equation $Lu - \lambda u = f$ in the half space

$$\mathbb{R}^d_+ = \{x \in \mathbb{R}^d : x^1 > 0\}.$$

Their proofs show the advantage of having the solvability in $\mathbb{R}^d$ of equations whose coefficients are only measurable in one direction. In what follows, we denote by $\overset{o}{W}{}_p^2(\mathbb{R}^d_+)$ the collection of all $u \in W_p^2(\mathbb{R}^d_+)$ satisfying $u(0,x') \equiv 0$.

Theorem 2.7. *Let $p \in [2,\infty)$. If $p = 2$, then suppose, additionally, that the assumption in Theorem 2.5 is satisfied. Then there exists a constant $\lambda_0 =$*

$\lambda_0(d,\delta,K,p,\omega) \geq 0$ such that, for any $\lambda > \lambda_0$ and $f \in L_p(\mathbb{R}^d_+)$, there exists a unique $u \in \overset{o}{W}{}^2_p(\mathbb{R}^d_+)$ satisfying $Lu - \lambda u = f$.

Furthermore, there is a constant $N = N(d,\delta,K,p,\omega)$ such that, for any $\lambda \geq \lambda_0$ and $u \in \overset{o}{W}{}^2_p(\mathbb{R}^d_+)$,

$$\lambda\|u\|_{L_p(\mathbb{R}^d_+)} + \sqrt{\lambda}\|u_x\|_{L_p(\mathbb{R}^d_+)} + \|u_{xx}\|_{L_p(\mathbb{R}^d_+)} \leq N\|Lu - \lambda u\|_{L_p(\mathbb{R}^d_+)}. \tag{2.2}$$

Proof. We introduce a new operator $\hat{L}v = \hat{a}^{jk}v_{x^jx^k} + \hat{b}^j v_{x^j} + \hat{c}v$, the coefficients of which are as follows. First, we view the coefficients a^{jk}, b^j, and c as functions defined only on $\mathbb{R}^d_+$. Then we define $\hat{a}^{jk}$, $\hat{b}^j$, and $\hat{c}$ to be the odd or even extensions of the original coefficients. Specifically, if $j = k = 1$ or $j,k \in \{2,\dots,d\}$, then (even extension)

$$\hat{a}^{jk}(x) = \begin{cases} a^{jk}(x^1,x') & \text{if } x^1 \geq 0, \\ a^{jk}(-x^1,x') & \text{if } x^1 < 0. \end{cases}$$

If $j = 2,\dots,d$, then (odd extension)

$$\hat{a}^{1j}(x) = \hat{a}^{j1}(x) = \begin{cases} a^{1j}(x^1,x') & \text{if } x^1 \geq 0, \\ -a^{1j}(-x^1,x') & \text{if } x^1 < 0. \end{cases}$$

Similarly, the coefficient $\hat{b}^1(x)$ is the odd extension of $b^1(x)$, and the coefficients $\hat{b}^j(x)$, $j = 2,\dots,d$, and $\hat{c}(x)$ are the even extensions of $b^j(x)$ and $c(x)$, respectively.

Now we notice that the coefficients of $\hat{L}$ satisfy Assumptions 2.1 and 2.2 with $2\,\omega$. Then by Theorems 2.4 and 2.5, we can find a constant $\lambda_0 = \lambda_0(d,\delta,K,p,\omega)$ such that, for any $\lambda > \lambda_0$, there exists a unique $u \in W^2_p$ satisfying $\hat{L}u - \lambda u = \hat{f}$, where $\hat{f} \in L_p$ is the odd extension of $f \in L_p(\mathbb{R}^d_+)$. Obviously, $-u(-x^1,x') \in W^2_p$ also satisfies the same equation, so by uniqueness we have $u(x^1,x') = -u(-x^1,x')$. This implies that u, as a function defined on $\mathbb{R}^d_+$, is in the space $\overset{o}{W}{}^2_p(\mathbb{R}^d_+)$. Since $Lu - \lambda u = f$ in $\mathbb{R}^d_+$, the function u is a solution to the Dirichlet boundary problem.

To prove uniqueness and the estimate (2.2), we use the estimates in Theorems 2.4 and 2.5 and the fact that the odd extension of an element in $\overset{o}{W}{}^2_p(\mathbb{R}^d_+)$ is in W^2_p. The theorem is now proved. ∎

In the same way, only this time taking the even extension of f, one gets the solvability of the Neumann problem.

THEOREM 2.8. *Let $p \in [2,\infty)$. If $p = 2$, then suppose, additionally, that the assumption in Theorem 2.5 is satisfied. Then there exists a constant $\lambda_0 = \lambda_0(d,\delta,K,p,\omega) \geq 0$ such that, for any $\lambda > \lambda_0$ and $f \in L_p(\mathbb{R}^d_+)$, there exists a unique $u \in W^2_p(\mathbb{R}^d_+)$ satisfying $Lu - \lambda u = f$ and $u_{x^1} = 0$ on $\partial\mathbb{R}^d_+$.*

Furthermore, there is a constant $N = N(d,\delta,K,p,\omega)$ such that, for any $\lambda \geq \lambda_0$ and $u \in W^2_p(\mathbb{R}^d_+)$ satisfying $u_{x^1} = 0$ on $\partial\mathbb{R}^d_+$,

$$\lambda\|u\|_{L_p(\mathbb{R}^d_+)} + \sqrt{\lambda}\|u_x\|_{L_p(\mathbb{R}^d_+)} + \|u_{xx}\|_{L_p(\mathbb{R}^d_+)} \leq N\|Lu - \lambda u\|_{L_p(\mathbb{R}^d_+)}.$$

While the Neumann problem is solved without any effort, oblique derivative problems need some, still simple, manipulations.

Let ℓ be a constant vector field $\ell = (\ell^1,\dots,\ell^d)$, where $\ell^1 > 0$. Set $s = 1 - 1/p$ and recall that $g \in W^s_p(\mathbb{R}^{d-1})$ if

$$\|g\|_{W^s_p(\mathbb{R}^{d-1})} = \|g\|_{L_p(\mathbb{R}^{d-1})} + [g]_s < \infty,$$

where

$$[g]_s^p = \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}^{d-1}} \frac{|g(x') - g(y')|^p}{|x' - y'|^{d-1+sp}} \, dx' \, dy'.$$

THEOREM 2.9. *Let $p \in [2, \infty)$. If $p = 2$, then suppose, additionally, that the assumption in Theorem 2.5 is satisfied. Then there exists a constant $\lambda_0 = \lambda_0(d, \delta, K, p, \omega, \ell) \geq 0$ such that, for any $\lambda > \lambda_0$, $f \in L_p(\mathbb{R}^d_+)$, and $g \in W_p^{1-1/p}(\mathbb{R}^{d-1})$, there exists a unique $u \in W_p^2(\mathbb{R}^d_+)$ satisfying $Lu - \lambda u = f$ and $\ell^j \, u_{x^j} = g$ on $\partial \mathbb{R}^d_+$.*

Furthermore, there is a constant $N = N(d, \delta, K, p, \omega, \ell)$ such that, for any $\lambda \geq \lambda_0$ and $u \in W_p^2(\mathbb{R}^d_+)$,

$$\lambda \|u\|_{L_p(\mathbb{R}^d_+)} + \sqrt{\lambda} \|u_x\|_{L_p(\mathbb{R}^d_+)} + \|u_{xx}\|_{L_p(\mathbb{R}^d_+)}$$

$$\leq N \left(\|Lu - \lambda u\|_{L_p(\mathbb{R}^d_+)} + (\lambda \vee 1)^{s/2} \|g\|_{L_p(\mathbb{R}^{d-1})} + [g]_s \right), \tag{2.3}$$

where $\lambda \vee 1 = \max\{\lambda, 1\}$, $s = 1 - 1/p$, and $g(x') = \ell^j \, u_{x^j}(0, x')$.

Proof. We may assume that $\ell^1 = 1$. Then introduce new coordinates in $\mathbb{R}^d_+$ by $y^1 = x^1, y' = x' - x^1 \ell'$. It is easy to check that under this change of variables the condition $\ell^j \, u_{x^j} = g$ becomes $u_{y^1} = g$ and the operator L will be transformed into a different one but yet satisfying the same assumptions as L does (with somewhat different constants δ and K and the same ω). It follows that in the rest of the proof we may assume that $\ell = (1, 0, \ldots, 0)$ is the first unit basis vector.

Next, we reduce the case of general g to that of $g = 0$. One knows (see, for instance, Theorem 2.9.1 of [18]) that the trace operator $u(x) \to u_{x^1}(0, x')$ is a bounded operator from $W_p^2(\mathbb{R}^d_+)$ onto $W_p^s(\mathbb{R}^{d-1})$, for each $g \in W_p^s(\mathbb{R}^{d-1})$ there is a function $v \in W_p^2(\mathbb{R}^d_+)$ such that $v = 0$ and $v_{x^1} = g(x')$ on $\partial \mathbb{R}^d_+$, and for a constant N independent of g,

$$\|v\|_{W_p^2(\mathbb{R}^d_+)} \leq N \|g\|_{W_p^s(\mathbb{R}^{d-1})}. \tag{2.4}$$

It follows by using dilations $(v(x), g(x') \to v(\sqrt{\lambda} x), \sqrt{\lambda} g(\sqrt{\lambda} x'))$ that for any $g \in W_p^s(\mathbb{R}^{d-1})$ and $\lambda > 0$, we can find $v \in W_p^2(\mathbb{R}^d_+)$ satisfying $v = 0$, $v_{x^1} = g$ on $\partial \mathbb{R}^d_+$, and

$$\lambda \|v\|_{L_p(\mathbb{R}^d_+)} + \sqrt{\lambda} \|v_x\|_{L_p(\mathbb{R}^d_+)} + \|v_{xx}\|_{L_p(\mathbb{R}^d_+)} \leq N \left(\lambda^{s/2} \|g\|_{L_p(\mathbb{R}^{d-1})} + [g]_s \right), \tag{2.5}$$

where N depends only on d and p. This implies that if the first assertion of the theorem is true for $g = 0$, then by replacing f with $f - Lv + \lambda v$, finding u, and introducing $w = u + v$, we will have $w \in W_p^2(\mathbb{R}^d_+)$, $Lw - \lambda w = f$ in $\mathbb{R}^d_+$, and $w_{x^1} = g$ on $\partial \mathbb{R}^d_+$.

Furthermore, if estimate (2.3) holds true provided that $u_{x^1} = 0$ on $\partial \mathbb{R}^d_+$, then take an arbitrary $u \in W_p^2(\mathbb{R}^d_+)$, introduce $g := u_{x^1}$ on $\partial \mathbb{R}^d_+$, find v as above, but corresponding to $\lambda \vee 1$ in place of λ, and write that by the assumption

$$\|(u - v)_{xx}\|_{L_p(\mathbb{R}^d_+)} \leq N \|(L - \lambda)(u - v)\|_{L_p(\mathbb{R}^d_+)},$$

$$\|u_{xx}\|_{L_p(\mathbb{R}^d_+)} \leq N \|(L - \lambda) u\|_{L_p(\mathbb{R}^d_+)} + N \|v_{xx}\|_{L_p(\mathbb{R}^d_+)}$$

$$+N\|v_x\|_{L_p(\mathbb{R}^d_+)} + N(1+\lambda)\|v_x\|_{L_p(\mathbb{R}^d_+)}.$$

The last expression is majorated by the right-hand side of (2.3) owing to (2.5) (recall that we take there $\lambda \vee 1$ in place of λ). Similarly, one estimates the remaining terms on the left-hand side of (2.3). Thus, in the rest of the proof we may assume that $g = 0$.

In that case take the operator $\hat{L}$ from the proof of Theorem 2.7 and consider the equation

$$\hat{L}w - \lambda w = \hat{f} \tag{2.6}$$

in $\mathbb{R}^d$, where $\hat{f}(x) := f(|x^1|, x')$. Using Theorems 2.4 and 2.5, we find a unique solution $w \in W^2_p$ to (2.6) for $\lambda > \lambda_0$, where $\lambda_0 = \lambda_0(d, \delta, K, p, \omega, \ell)$ is a constant corresponding to the operator $\hat{L}$. Obviously, $v(x) = u(-x^1, x')$ is also a solution of (2.6). By uniqueness, $v = u$, $u(x) = u(-x^1, x')$ implying that $u_{x^1} = 0$ on $\partial\mathbb{R}^d_+$, and since $u \in W^2_p(\mathbb{R}^d_+)$, we have proved the existence of the desired solution.

To complete the proof, we now prove only (2.3), which implies uniqueness. As we agreed upon above we need only consider the case that $g = u_{x^1} = 0$ on $\partial\mathbb{R}^d_+$. Take such a $u \in W^2_p(\mathbb{R}^d_+)$ and set $v(x) = u(|x^1|, x')$. Since $u_{x^1} = 0$ on $\partial\mathbb{R}^d_+$, we have $v \in W^2_p$ and for $\lambda > \lambda_0 = \lambda_0(d, \delta, K, p, \omega, \ell)$, we have

$$\lambda\|v\|_{L_p} + \sqrt{\lambda}\|v_x\|_{L_p} + \|v_{xx}\|_{L_p} \le N\|f\|_{L_p},$$

where $f = \hat{L}v - \lambda v$. Obviously, $f(x) = f(|x^1|, x')$ and $v = u$ on $\mathbb{R}^d_+$. Hence, the above estimate implies (2.3) (recall that $g = 0$).

We have proved (2.3) for $\lambda > \lambda_0$. For $\lambda = \lambda_0$ we get (2.3) by continuity. □

Remark 2.10. Let $\ell(x') = (\ell^1(x'), \dots, \ell^d(x'))$ be a bounded vector field defined on $\mathbb{R}^{d-1}$ such that $\ell(x') \in C^{1-1/p+\varepsilon}(\mathbb{R}^{d-1})$, $\varepsilon > 0$, and $\ell^1(x') \ge \kappa > 0$. Then using the freezing coefficients, partition of unity, and the method of continuity, we can replace the constant vector field ℓ by $\ell(x')$ in the above theorem. Details can be found in [10].

Remark 2.11. A result similar to Theorem 2.9 holds if we replace the boundary condition $\ell^j u_{x^j} = g$ with $\ell^j u_{x^j} + \sigma u = g$, where σ is a constant. Indeed, again assuming that $\ell^1 = 1$, it is easy to find an infinitely differentiable bounded function $h(x^1)$ having bounded derivatives and bounded away from zero such that $h'(0) = -\sigma h(0)$. Then for $v = u/h$ we have $\ell^j v_{x^j} = g/h$ on $\partial\mathbb{R}^d_+$ and $Lu - \lambda u = h(\bar{L}v - \lambda v)$, where $\bar{L}\phi := h^{-1}L(h\phi)$ is an elliptic operator satisfying our hypotheses with a slightly modified K.

3. Proof of Theorem 2.5. Thanks to the method of continuity and the denseness of $C_0^\infty(\mathbb{R}^d)$ in W^2_2, it suffices to prove the a priori estimate (2.1) for $u \in C_0^\infty(\mathbb{R}^d)$ and a^{jk} that are sufficiently smooth. In addition, on the account of possibly increasing λ_0 one sees that it suffices to prove (2.1) for $b \equiv 0$, $c \equiv 0$, and $\lambda_0 = 0$. In that case set

$$f = Lu - \lambda u. \tag{3.1}$$

For functions $\phi(x^1, x')$ we denote by $\tilde{\phi}(x^1, \xi)$, $\xi \in \mathbb{R}^{d-1}$, its Fourier transform with respect to x'. By taking the Fourier transforms of both sides of (3.1), we obtain

$$\mathsf{a}\tilde{u}_{x^1x^1} + i2\mathsf{b}\tilde{u}_{x^1} - \mathsf{c}\tilde{u} = \tilde{f},$$

$$\tilde{u}_{x^1x^1} + i2\hat{\mathsf{b}}\tilde{u}_{x^1} - \hat{\mathsf{c}}\tilde{u} = \tilde{g}, \tag{3.2}$$

where $\mathrm{i} = \sqrt{-1}$ and

$$\mathsf{a}(x^1) = a^{11}(x^1), \quad \mathsf{b}(x^1,\xi) = \sum_{j=2}^{d} a^{1j}(x^1)\xi^j, \quad \hat{\mathsf{b}} = \mathsf{a}^{-1}\mathsf{b},$$

$$\mathsf{c}(x^1,\xi) = \sum_{j,k=2}^{d} a^{jk}(x^1)\xi^j\xi^k + \lambda, \quad \hat{\mathsf{c}} = \mathsf{a}^{-1}\mathsf{c}, \quad \tilde{g} = \mathsf{a}^{-1}\tilde{f}.$$

LEMMA 3.1. *We have*

$$\delta \le \mathsf{a} = a^{11} \le \delta^{-1}, \quad |\mathsf{b}(x^1,\xi)| \le \delta^{-1}|\xi|,$$

$$\delta^{-1}(|\xi|^2 + \lambda) \ge \mathsf{c}(x^1,\xi) \ge \delta|\xi|^2 + \lambda, \tag{3.3}$$

and

$$\mathsf{a}(x^1)\mathsf{c}(x^1,\xi) - \mathsf{b}^2(x^1,\xi) \ge \delta^2(|\xi|^2 + \lambda).$$

Proof. The first and last inequalities in (3.3) are obvious. The second one follows by Cauchy's inequality $(a^{jk}\eta^j\xi^k)^2 \le (a^{jk}\eta^j\eta^k)(a^{jk}\xi^j\xi^k)$, where $\xi^1 = \eta^j = 0$, $j \ge 2$, $\eta^1 = 1$. Next, from Assumption 2.1 we have

$$\delta(t^2 + |\xi|^2) \le \mathsf{a}(x^1)\,t^2 + 2\,\mathsf{b}(x^1,\xi)\,t + \mathsf{c}(x^1,\xi) - \lambda$$

for all $t \in \mathbb{R}$ and $\xi \in \mathbb{R}^{d-1}$. In particular,

$$\big(\mathsf{a}(x^1) - \delta\big)\,t^2 + 2\,\mathsf{b}(x^1,\xi)\,t + \mathsf{c}(x^1,\xi) - \delta|\xi|^2 - \lambda \ge 0.$$

This implies that

$$\mathsf{b}^2(x^1,\xi) - \big(\mathsf{a}(x^1) - \delta\big)\big(\mathsf{c}(x^1,\xi) - \delta|\xi|^2 - \lambda\big) \le 0.$$

From this and (3.3) the result follows. □

LEMMA 3.2. *For any $\xi \in \mathbb{R}^d$*

$$(|\xi|^2 + \lambda)\int_{\mathbb{R}} |\tilde{u}_{x^1}|^2\,dx^1 + (|\xi|^4 + \lambda|\xi|^2 + \lambda^2)\int_{\mathbb{R}} |\tilde{u}|^2\,dx^1 \le N(\delta)\int_{\mathbb{R}} |\tilde{f}|^2\,dx^1, \tag{3.4}$$

$$\int_{\mathbb{R}} |\tilde{u}_{x^1x^1}|^2\,dx^1 \le N(\delta)\int_{\mathbb{R}} |\tilde{f}|^2\,dx^1. \tag{3.5}$$

Proof. Estimate (3.5) is a direct consequence of (3.2) (allowing one to express $\tilde{u}_{x^1x^1}$ through $\tilde{f}$, $\tilde{u}_{x^1}$, and $\tilde{u}$), (3.3), and (3.4).

While proving (3.4) we define a function $\phi(x^1,\xi)$ by $\phi(0,\xi) = 0$ and $\phi_{x^1} = \hat{\mathsf{b}}$ and set $\rho = \tilde{u}e^{\mathrm{i}\phi}$. Then from (3.2) we see that

$$\rho_{x^1x^1} + (\hat{\mathsf{b}}^2 - \mathrm{i}\,\phi_{x^1x^1} - \hat{\mathsf{c}})\rho = \tilde{g}e^{\mathrm{i}\phi}.$$

Multiply both sides by $\bar{\rho}$ and integrate the result with respect to x^1. Integrating by parts shows that

$$-\int_{\mathbb{R}} |\rho_{x^1}|^2\,dx^1 + \int_{\mathbb{R}} (\hat{\mathsf{b}}^2 - \mathrm{i}\,\phi_{x^1x^1} - \hat{\mathsf{c}})|\tilde{u}|^2\,dx^1 = \int_{\mathbb{R}} \tilde{g}\bar{\tilde{u}}\,dx^1.$$

Taking the real parts of both sides and multiplying by $|\xi|^2+\lambda$, we have

$$\int_{\mathbb{R}}(|\xi|^2+\lambda)|\rho_{x^1}|^2\,dx^1+\int_{\mathbb{R}}(\hat{\mathsf{c}}-\hat{\mathsf{b}}^2)(|\xi|^2+\lambda)|\tilde u|^2\,dx^1$$

$$=-\int_{\mathbb{R}}(|\xi|^2+\lambda)\Re(\tilde g\bar{\tilde u})\,dx^1.$$

Note that for any $\varepsilon>0$

$$-(|\xi|^2+\lambda)\Re(\tilde g\bar{\tilde u})\le\varepsilon(|\xi|^2+\lambda)^2|\tilde u|^2+\varepsilon^{-1}|\tilde g|^2.$$

From this and Lemma 3.1 we obtain

$$\int_{\mathbb{R}}(|\xi|^2+\lambda)|\rho_{x^1}|^2\,dx^1+\int_{\mathbb{R}}(\delta^4-\varepsilon)(|\xi|^2+\lambda)^2|\tilde u|^2\,dx^1\le\varepsilon^{-1}\int_{\mathbb{R}}|\tilde g|^2\,dx^1.$$

By choosing an appropriate $\varepsilon>0$ (e.g., $\varepsilon=\delta^4/2$), we arrive at

$$\int_{\mathbb{R}}(|\xi|^2+\lambda)|\rho_{x^1}|^2\,dx^1+\int_{\mathbb{R}}(|\xi|^4+\lambda|\xi|^2+\lambda^2)|\tilde u|^2\,dx^1\le N(\delta)\int_{\mathbb{R}}|\tilde f|^2\,dx^1.$$

It only remains to observe that in light of (3.3)

$$|\tilde u_{x^1}|=|\rho_{x^1}-\mathsf{i}\mathsf{b}\mathsf{a}^{-1}\tilde u e^{i\phi}|\le|\rho_{x^1}|+N(\delta)|\xi||\tilde u|.\qquad\square$$

Now we can finish the proof of Theorem 2.5. As we pointed out in the beginning of the section, we only need to prove (2.1) for $u\in C_0^\infty(\mathbb{R}^d)$, smooth a^{ij}, $b\equiv0$, $c\equiv0$, and $\lambda_0=0$.

In that case it suffices to add (3.4) and (3.5), integrate over $\mathbb{R}^{d-1}$, and use Parseval's identity. The theorem is proved.

Remark 3.3. We have just proved that if $b^j=c=0$, then

$$\lambda\|u\|_{L_2}+\sqrt{\lambda}\|u_x\|_{L_2}+\|u_{xx}\|_{L_2}\le N\|Lu-\lambda u\|_{L_2}$$

for $u\in W_2^2$ and $\lambda\ge0$, where N depends only on δ.

4. Auxiliary results. Here we state and prove a series of observations which are needed in the proof of Theorem 2.4. First, we introduce some notation. As usual, we set $B_r(x_0)=\{x\in\mathbb{R}^d:|x-x_0|<r\}$ and $B_r=B_r(0)$. By $|B_r|$ we mean the d-dimensional volume of B_r. We denote by $|u|_0$ the supremum of $|u|$ over $\mathbb{R}^d$.

Throughout this section, we assume that

$$Lu(x)=L_0u(x)=a^{jk}(x^1)u_{x^jx^k}(x).$$

Our first auxiliary result is the following lemma.

LEMMA 4.1. *There exists $N=N(d,\delta)$ such that, for any $u\in W_2^2(B_R)$ with $u|_{\partial B_R}=0$, we have*

$$R^2\int_{B_R}|u_x|^2\,dx+\int_{B_R}|u|^2\,dx\le N\,R^4\int_{B_R}|Lu|^2\,dx.\tag{4.1}$$

Proof. Assume that (4.1) is true when $R=1$. For a given $u\in W_2^2(B_R)$ with $u|_{\partial B_R}=0$, we set

$$L_R=a^{jk}(Rx)\frac{\partial^2}{\partial x^j\partial x^k}\quad\text{and}\quad v(x)=R^{-2}u(Rx).$$

Then $v \in W_2^2(B_1)$ and $L_R v(x) = (Lu)(Rx)$ in B_1. Since L_R satisfies the same ellipticity condition as L does, we have

$$\int_{B_R} |u|^2\,dx = R^{d+4}\int_{B_1}|v|^2\,dx$$

$$\le NR^{d+4}\int_{B_1}|L_R v|^2\,dx = NR^4\int_{B_R}|Lu|^2\,dx.$$

Also

$$\int_{B_R} |u_x|^2\,dx = R^{d+2}\int_{B_1}|v_x|^2\,dx$$

$$\le NR^{d+2}\int_{B_1}|L_R v|^2\,dx = NR^2\int_{B_R}|Lu|^2\,dx.$$

This shows that we need only prove the lemma for $R = 1$.

In that case we can divide L by a^{11} and may assume that $a^{11} \equiv 1$. Then we integrate uLu over B_1 using integration by parts to find

$$\delta\int_{B_1}|u_x|^2\,dx \le \int_{B_1} a^{jk}u_{x^j}u_{x^k}\,dx = -\int_{B_1} uLu\,dx$$

$$\le \left(\int_{B_1} u^2\,dx\right)^{1/2}\left(\int_{B_1}(Lu)^2\,dx\right)^{1/2}.$$

We estimate the integral of u^2 through that of $|u_x|^2$ by using Poincaré's inequality and obtain the needed estimate for u_x. This is the only estimate we need to prove since u is estimated by u_x, again owing to Poincaré's inequality. $\square$

The following lemma is almost identical to a theorem in [5]. For completeness, we also present a proof.

LEMMA 4.2. *Let $0 < r < R$. There exists $N = N(d,\delta)$ such that, for $w \in W_2^2(B_R)$,*

$$\|w\|_{W_2^2(B_r)} \le N\left(\|Lw - w\|_{L_2(B_R)} + (R-r)^{-2}\|w\|_{L_2(B_R)}\right). \tag{4.2}$$

Proof. Set

$$R_0 = r, \quad R_m = r + (R-r)\sum_{k=1}^{m}\frac{1}{2^k}, \quad m = 1,2,\ldots,$$

$$B(m) = \{x \in \mathbb{R}^d : |x| \le R_m\}, \quad m = 0,1,\ldots.$$

Also, let $\zeta_m \in C_0^\infty(\mathbb{R}^d)$ be such that $\zeta_m(x) = 1$ in $B(m)$, $\zeta_m(x) = 0$ outside of $B(m+1)$, and

$$|(\zeta_m)_x|_0 \le N\frac{2^{m+1}}{R-r}, \quad |(\zeta_m)_{xx}|_0 \le N\frac{2^{2m+2}}{(R-r)^2},$$

where N depends only on d. To construct them take an infinitely differentiable function $g(t)$, $t \in (-\infty, \infty)$, such that $g(t) = 1$ for $t \leq 1$, $g(t) = 0$ for $t \geq 2$, and $0 \leq g \leq 1$. After this define

$$\zeta_m(x) = g(2^{m+1}(R-r)^{-1}(|x| - R_m) + 1).$$

Now we make use of the L_2-estimate of $\zeta_m w$, which is from Remark 3.3, and obtain (observe that $N2^{2m+2} = N_1 2^{2m}$ with $N_1 = 4N$)

$$\mathcal{A}_m := \|\zeta_m w\|_{W_2^2} \leq N\|(L-1)(\zeta_m w)\|_{L_2}$$

$$\leq N\mathcal{B} + N\|\zeta_{mx} w_x\|_{L_2} + N\frac{2^{2m}}{(R-r)^2}\mathcal{C}, \tag{4.3}$$

where N depends only on d and δ and

$$\mathcal{B} := \|(L-1)w\|_{L_2(B_R)} \quad \text{and} \quad \mathcal{C} := \|w\|_{L_2(B_R)}.$$

By interpolation inequalities

$$\|\zeta_{mx} w_x\|_{L_2} = \|\zeta_{mx}(\zeta_{m+1}w)_x\|_{L_2} \leq N\frac{2^m}{R-r}\|(\zeta_{m+1}w)_x\|_{L_2}$$

$$\leq \varepsilon\mathcal{A}_{m+1} + N\varepsilon^{-1}\frac{2^{2m}}{(R-r)^2}\mathcal{C},$$

where $\varepsilon > 0$ is arbitrary and N depends only on d. Thus, (4.3) yields

$$\mathcal{A}_m \leq N\mathcal{B} + \varepsilon\mathcal{A}_{m+1} + N\varepsilon^{-1}\frac{2^{2m}}{(R-r)^2}\mathcal{C}$$

with ε perhaps different from the one above but still arbitrary. We set $\varepsilon = 1/8$ and get

$$\varepsilon^m\mathcal{A}_m \leq N\varepsilon^m\mathcal{B} + \varepsilon^{m+1}\mathcal{A}_{m+1} + N\varepsilon^m\frac{2^{2m}}{(R-r)^2}\mathcal{C},$$

$$\mathcal{A}_0 + \sum_{m=1}^{\infty}\varepsilon^m\mathcal{A}_m \leq N\mathcal{B} + \sum_{m=1}^{\infty}\varepsilon^m\mathcal{A}_m + N\frac{1}{(R-r)^2}\mathcal{C}.$$

Here the series of $\varepsilon^m\mathcal{A}_m = 8^{-m}\mathcal{A}_m$ converges since

$$\mathcal{A}_m \leq N2^{2m}(R-r)^{-2}\|w\|_{W_2^2(B_R)}.$$

Therefore, after taking care of similar terms we see that $\mathcal{A}_0$ is less than the right-hand side of (4.2). Since its left-hand side is obviously less than $\mathcal{A}_0$, the lemma is proved. □

Remark 4.3. Using the dilation argument as in the proof of Lemma 4.1, we have

$$\lambda\|w\|_{L_2(B_r)} + \sqrt{\lambda}\|w_x\|_{L_2(B_r)} + \|w_{xx}\|_{L_2(B_r)}$$

$$\leq N \left(\|Lw - \lambda w\|_{L_2(B_R)} + (R-r)^{-2} \|w\|_{L_2(B_R)} \right)$$

for any $\lambda > 0$, where N depends only on d and δ. In particular, by letting $\lambda \to 0$, we have

$$\|w_{xx}\|_{L_2(B_r)} \leq N \left(\|Lw\|_{L_2(B_R)} + (R-r)^{-2} \|w\|_{L_2(B_R)} \right). \tag{4.4}$$

In the next few lemmas, we investigate some properties of a solution h of the equation $Lh = 0$. Recall that the coefficients a^{jk} of the operator L do not depend on $x' \in \mathbb{R}^{d-1}$.

LEMMA 4.4. *Let $\gamma = (\gamma^1, \dots, \gamma^d)$ be a multi-index such that $\gamma^1 = 0, 1, 2$. Also let $0 < r < R \leq 4$. If h is a sufficiently smooth function defined on B_4 such that $Lh = 0$ in B_4, then we have*

$$\int_{B_r} |D^\gamma h|^2 \, dx \leq N \int_{B_R} |h|^2 \, dx,$$

where $N = N(d, \delta, \gamma, R, r)$.

Proof. Set $\gamma' = (0, \gamma^2, \dots, \gamma^d)$ and notice that

$$L(D^{\gamma'} h) = 0, \quad \text{that is} \quad (L-1) D^{\gamma'} h = -D^{\gamma'} h \quad \text{in} \quad B_4.$$

Then by Lemma 4.2

$$\|D^\gamma h\|_{L_2(B_r)} \leq N \left(\|D^{\gamma'} h\|_{L_2(B_{r_1})} + (r_1 - r)^{-2} \|D^{\gamma'} h\|_{L_2(B_{r_1})} \right),$$

where $r < r_1 < R$. If $|\gamma'| = 0$, then we are done. Otherwise, we can consider a multi-index γ'' having at least one component less by one than the corresponding component of γ'. Then, $L(D^{\gamma''} h) = 0$ and

$$\|D^{\gamma'} h\|_{L_2(B_{r_1})} \leq N \left(\|D^{\gamma''} h\|_{L_2(B_{r_2})} + (r_2 - r_1)^{-2} \|D^{\gamma''} h\|_{L_2(B_{r_2})} \right),$$

where $r < r_1 < r_2 < R$. We repeat this argument as many times as we need. The lemma is proved. $\square$

Denote by h_x a generic derivative h_{x^j}, $j = 1, \dots, d$, and $h_{x'}$ a generic derivative h_{x^j}, $j = 2, \dots, d$. Thus, for example, $h_{xx'}$ can be $h_{x^j x^k}$, where $j \in \{1, 2, \dots, d\}$ and $k \in \{2, \dots, d\}$.

LEMMA 4.5. *Let h be a sufficiently smooth function h defined on B_4 such that $Lh = 0$ in B_4. Then we have*

$$\sup_{B_1} |h_{xxx'}|^2 \leq N \int_{B_3} |h|^2 \, dx,$$

where $N = N(d, \delta)$.

Proof. Imagine that we have

$$\sup_{B_1} |h_{xx}| \leq N(d, \delta) \|h\|_{L_2(B_{5/2})}. \tag{4.5}$$

Then, using the fact that $Lh_{x'} = 0$, we would obtain

$$\sup_{B_1} |h_{x'xx}| \leq N \|h_{x'}\|_{L_2(B_{5/2})}$$

and it would only remain to appeal to Lemma 4.4.

Therefore, it suffices to prove (4.5). To do that, we first fix an integer k such that $k-(d-1)/2>0$. Then, due to the Sobolev embedding theorem, we can find a constant N such that, for each $-1 \le x^1 \le 1$,

$$\sup_{|x'|\le 1} |h_{x'x^1x^1}(x^1,x')| \le N\|h_{x'x^1x^1}(x^1,\cdot)\|_{W_2^k(B_1')}$$

and

$$\sup_{|x'|\le 1} |h_{x'x^1}(x^1,x')| \le N\|h_{x'x^1}(x^1,\cdot)\|_{W_2^k(B_1')},$$

where $B_1' = \{x' \in \mathbb{R}^{d-1} : |x'| \le 1\}$. Set g to be either $h_{x'x^1x^1}$ or $h_{x'x^1}$. Then

$$\int_{-1}^1 \sup_{|x'|\le 1} |g(x^1,x')|^2\,dx^1 \le N\int_{-1}^1 \|g(x^1,\cdot)\|^2_{W_2^k(B_1')}\,dx^1$$

$$\le N \sum_{\substack{|\gamma|\le k+3\\ 1\le\gamma^1\le 2}} \|D^\gamma h\|^2_{L_2(B_2)}.$$

From this and Lemma 4.4 we have

$$\int_{-1}^1 \sup_{|x'|\le 1} |h_{x'x^1x^1}|^2\,dx^1 + \int_{-1}^1 \sup_{|x'|\le 1} |h_{x'x^1}|^2\,dx^1 \le N\|h\|^2_{L_2(B_{5/2})}, \tag{4.6}$$

where N depends only on d and δ. Now we notice that, for $x^1, y^1 \in [-1,1]$,

$$\sup_{|x'|\le 1} |h_{x'x^1}(x^1,x')| - \sup_{|x'|\le 1} |h_{x'x^1}(y^1,x')|$$

$$\le \sup_{|x'|\le 1} \big|h_{x'x^1}(x^1,x') - h_{x'x^1}(y^1,x')\big| \le \int_{x^1}^{y^1} \sup_{|x'|\le 1} |h_{x'x^1x^1}(t,x')|\,dt$$

$$\le |x^1-y^1|^{1/2}\left(\int_{-1}^1 \sup_{|x'|\le 1} |h_{x'x^1x^1}(t,x')|^2\,dt\right)^{1/2}.$$

This and (4.6) imply

$$\sup_{|x'|\le 1} |h_{x'x^1}(x^1,x')| \le N\|h\|_{L_2(B_{5/2})}\,|x^1-y^1|^{1/2} + \sup_{|x'|\le 1} |h_{x'x^1}(y^1,x')|.$$

Take integrals of both sides with respect to y^1, and take a supremum over x^1. Then

$$\sup_{x\in B_1} |h_{x'x^1}(x)| \le N\|h\|_{L_2(B_{5/2})} + \int_{-1}^1 \sup_{|x'|\le 1} |h_{x'x^1}(y^1,x')|\,dy^1 \le N\|h\|_{L_2(B_{5/2})},$$

where the last inequality follows from (4.6), and N depends only on d and δ. Similarly, we follow the same steps as above with $h_{x'x'x^1}$ and $h_{x'x'}$ in place of $h_{x'x^1x^1}$ and $h_{x'x^1}$, respectively. Therefore, we have

$$\sup_{x\in B_1} |h_{x'x}(x)| \le N(d,\delta)\|h\|_{L_2(B_{5/2})}.$$

Finally, using the fact that $a^{11}h_{x^1x^1} = -\sum_{jk>1} a^{jk}h_{x^jx^k}$, we finish the proof of (4.5). ◻

Denote by $(u)_{B_r(x_0)}$ the average value of a function u over $B_r(x_0)$, that is,

$$(u)_{B_r(x_0)} = \fint_{B_r(x_0)} u(x)\,dx = \frac{1}{|B_r|}\int_{B_r(x_0)} u(x)\,dx.$$

Let $u \in C_0^\infty(\mathbb{R}^d)$ and $f := Lu$. Assume that $a^{jk}(x^1)$ are infinitely differentiable as functions of $x^1 \in \mathbb{R}$. Then we can find a sufficiently smooth function h defined on B_4 such that

$$Lh = 0 \quad \text{in} \quad B_4, \quad h = u \quad \text{on} \quad \partial B_4.$$

For this solution h, we establish the following inequality.

Lemma 4.6. *There exists a constant $N = N(d,\delta)$ such that*

$$\sup_{B_1} |h_{xxx'}|^2 \le N\int_{B_4} |f|^2\,dx + N\int_{B_4} |u_{xx}|^2\,dx.$$

Proof. Define

$$\tilde{u} := u - u_{B_4} - (u_{x^i})_{B_4}x^i \quad \text{in} \quad B_4, \quad \tilde{h} := h - u_{B_4} - (u_{x^i})_{B_4}x^i \quad \text{in} \quad B_4.$$

Then

$$L\tilde{u} = f, \quad L\tilde{h} = 0 \quad \text{in} \quad B_4, \quad \text{and} \quad \tilde{h} = \tilde{u} \quad \text{on} \quad \partial B_4.$$

By Lemma 4.5 we see that

$$\sup_{B_1} |h_{xxx'}|^2 = \sup_{B_1} |\tilde{h}_{xxx'}|^2 \le N\int_{B_3} |\tilde{h}|^2\,dx.$$

Let η be a function in $C_0^\infty(\mathbb{R}^d)$ such that $\eta(x) = 0$ in B_3 and $\eta(x) = 1$ at ∂B_4. Then $\tilde{h} - \eta\tilde{u} \in W_2^2(B_4)$ and $\tilde{h} - \eta\tilde{u} = 0$ on ∂B_4. Therefore, by Lemma 4.1

$$\int_{B_3} |\tilde{h}|^2\,dx = \int_{B_3} |\tilde{h} - \eta\tilde{u}|^2\,dx \le N(d,\delta)\int_{B_4} |L(\eta\tilde{u})|^2\,dx.$$

Note that

$$L(\eta\tilde{u}) = \eta Lu + 2a^{ij}\eta_{x^i}\tilde{u}_{x^j} + \tilde{u}L\eta$$

$$= \eta f + 2a^{ij}\eta_{x^i}(u_{x^j} - (u_{x^j})_{B_4}) + (u - u_{B_4} - (u_{x^i})_{B_4}x^i)L\eta.$$

Hence we have

$$\int_{B_4} |L(\eta\tilde{u})|^2\,dx \le N\int_{B_4} (|f|^2 + |u_{x^j} - (u_{x^j})_{B_4}|^2)\,dx$$

$$+N\int_{B_4} |u - u_{B_4} - (u_{x^i})_{B_4}x^i|^2\,dx \le N\int_{B_4} |f|^2\,dx + N\int_{B_4} |u_{xx}|^2\,dx,$$

where the last inequality follows from Lemmas 3.1 and 3.2 in [6], and N depends only on d and δ. The lemma is now proved. ∎

LEMMA 4.7. *Let $\kappa \geq 4$ and $r > 0$. Also let $a^{jk}(x^1)$ be infinitely differentiable. For a given $u \in C_0^\infty(\mathbb{R}^d)$, we find a smooth function h defined on $B_{\kappa r}$ such that $Lh = 0$ in $B_{\kappa r}$ and $h = u$ on $\partial B_{\kappa r}$. Then there exists a constant $N = N(d, \delta)$ such that*

$$\fint_{B_r} |h_{xx'} - (h_{xx'})_{B_r}|^2 \, dx \leq N\kappa^{-2} \left[\left(|Lu|^2\right)_{B_{\kappa r}} + \left(|u_{xx}|^2\right)_{B_{\kappa r}} \right]. \tag{4.7}$$

Proof. Using the dilation argument as in the proof of Lemma 4.1, we see that we need to prove only the case $r = 1$. In that case we first observe that by using the same dilation argument and Lemma 4.6, we have

$$\sup_{B_{\kappa/4}} |h_{xxx'}|^2 \leq N\kappa^{-2} \left[\left(|Lu|^2\right)_{B_\kappa} + \left(|u_{xx}|^2\right)_{B_\kappa} \right],$$

where N depends only on d and δ. Now we need only observe that $\kappa/4 \geq 1$, $r = 1$, and the left-hand side of the inequality (4.7) is not greater than a constant times $\sup_{B_1} |h_{xxx'}|^2$. The lemma is now proved. ∎

Using the results obtained above, we will finally arrive at the following lemma.

LEMMA 4.8. *There exists a constant $N = N(d, \delta)$ such that, for any $\kappa \geq 4$, $r > 0$, and $u \in C_0^\infty(\mathbb{R}^d)$, we have*

$$\fint_{B_r} |u_{xx'} - (u_{xx'})_{B_r}|^2 \, dx \leq N\kappa^d \left(|Lu|^2\right)_{B_{\kappa r}} + N\kappa^{-2} \left(|u_{xx}|^2\right)_{B_{\kappa r}}. \tag{4.8}$$

Proof. We can assume that $a^{jk}(x^1)$ are infinitely differentiable. In that case, we find a sufficiently smooth h defined on $B_{\kappa r}$ such that $Lh = 0$ in $B_{\kappa r}$ and $h = u$ on $\partial B_{\kappa r}$. Note that $L(u - h) = Lu$ in $B_{\kappa r}$ and $u - h = 0$ on $\partial B_{\kappa r}$. From Lemma 4.7 we have

$$\fint_{B_r} |h_{xx'} - (h_{xx'})_{B_r}|^2 \, dx \leq N\kappa^{-2} \left[\left(|Lu|^2\right)_{B_{\kappa r}} + \left(|u_{xx}|^2\right)_{B_{\kappa r}} \right]. \tag{4.9}$$

On the other hand, from estimate (4.4) we have

$$\int_{B_r} |u_{xx'} - h_{xx'}|^2 \, dx \leq N \left(\int_{B_{\kappa r}} |Lu|^2 \, dx + r^{-4}(\kappa - 1)^{-4} \int_{B_{\kappa r}} |u - h|^2 \, dx \right).$$

Moreover, by Lemma 4.1

$$\int_{B_{\kappa r}} |u - h|^2 \, dx \leq N \, (\kappa r)^4 \int_{B_{\kappa r}} |Lu|^2 \, dx.$$

Hence

$$\fint_{B_r} |u_{xx'} - h_{xx'}|^2 \, dx \leq N\kappa^d \left(|Lu|^2\right)_{B_{\kappa r}}.$$

This and (4.9) prove the inequality (4.8) with $(h_{xx'})_{B_r}$ in place of $(u_{xx'})_{B_r}$. Now we need only notice that

$$\fint_{B_r} |u_{xx'} - (u_{xx'})_{B_r}|^2 \, dx \leq \fint_{B_r} |u_{xx'} - (h_{xx'})_{B_r}|^2 \, dx.$$

The lemma is now proved. ∎

5. Proof of Theorem 2.4. In this section we suppose that all assumptions of Theorem 2.4 are satisfied. Recall that

$$Lu(x) = a^{jk}(x)u_{x^jx^k}(x) + b^j(x)u_{x^j}(x) + c(x)u(x),$$

and introduce

$$L_0u(x) = a^{jk}(x)u_{x^jx^k}(x).$$

We use the maximal and sharp functions given by

$$Mg(x) = \sup_{r>0} \fint_{B_r(x)} |g(y)|\,dy,$$

$$g^{\#}(x) = \sup_{r>0} \fint_{B_r(x)} |g(y) - (g)_{B_r(x)}|\,dy.$$

THEOREM 5.1. *Let μ, $\nu \in (1,\infty)$, $1/\mu + 1/\nu = 1$, and $R \in (0,\infty)$. Then there exists a constant $N = N(d,\delta,\mu)$ such that, for any $u \in C_0^\infty(B_R)$, we have*

$$(u_{xx'})^{\#} \le N\left(a_R^{\#(x')}\right)^{\alpha}\left[M(|u_{xx}|^{2\mu})\right]^{\beta}$$

$$+N\left[M(|L_0u|^2)\right]^{1/(d+2)}\left[M(|u_{xx}|^2)\right]^{d/(2d+4)}, \tag{5.1}$$

where $\alpha = \nu^{-1}(d+2)^{-1}$, $\beta = 2^{-1}\mu^{-1}$.

Proof. Fix $\kappa \ge 4$, $r \in (0,\infty)$, and $x_0 = (x_0^1, x_0') \in \mathbb{R}^d$. Introduce

$$\bar a^{jk}(x^1) = \frac{1}{|B'_{\kappa r}|}\int_{B'_{\kappa r}(x'_0)} a^{jk}(x^1,y')\,dy' \quad \text{if} \quad \kappa r < R,$$

$$\bar a^{jk}(x^1) = \frac{1}{|B'_R|}\int_{B'_R} a^{jk}(x^1,y')\,dy' \quad \text{if} \quad \kappa r \ge R,$$

$$\mathcal{A} = M(|L_0u|^2)(x_0), \qquad \mathcal{B} = M(|u_{xx}|^2)(x_0), \quad \mathcal{C} = \left(M(|u_{xx}|^{2\mu})(x_0)\right)^{1/\mu}.$$

Set $\bar L_0 u = \bar a^{jk}(x^1)u_{x^jx^k}$. Then Lemma 4.8 along with the fact that $\kappa \ge 4$ allows us to obtain

$$\fint_{B_r(x_0)} |u_{xx'} - (u_{xx'})_{B_r(x_0)}|^2\,dx$$

$$\le N\kappa^d\left(|\bar L_0 u|^2\right)_{B_{\kappa r}(x_0)} + N\kappa^{-2}\left(|u_{xx}|^2\right)_{B_{\kappa r}(x_0)} \tag{5.2}$$

for $\kappa \ge 4$, where N depends only on d and δ. Note that

$$\int_{B_{\kappa r}(x_0)} |\bar L_0 u|^2\,dx \le 2\int_{B_{\kappa r}(x_0)} |\bar L_0 u - L_0 u|^2\,dx + 2\int_{B_{\kappa r}(x_0)} |L_0 u|^2\,dx \tag{5.3}$$

and

$$\int_{B_{\kappa r}(x_0)} |\bar{L}_0 u - L_0 u|^2\,dx = \int_{B_{\kappa r}(x_0)\cap B_R} |\bar{L}_0 u - L_0 u|^2\,dx$$

$$\leq \left(\int_{B_{\kappa r}(x_0)\cap B_R} |\bar{a} - a|^{2\nu}\,dx\right)^{1/\nu} \left(\int_{B_{\kappa r}(x_0)} |u_{xx}|^{2\mu}\,dx\right)^{1/\mu} := I^{1/\nu} J^{1/\mu}.$$

If $\kappa r < R$, we have

$$I \leq N \int_{x_0^1-\kappa r}^{x_0^1+\kappa r} \int_{B'_{\kappa r}(x'_0)} |\bar{a}(x^1) - a(x^1, x')|\,dx'\,dx^1$$

$$\leq N(\kappa r)^d a_{\kappa r}^{\#(x')} \leq N(\kappa r)^d a_R^{\#(x')}.$$

In case $\kappa r \geq R$

$$I \leq N \int_{-R}^{R} \int_{B'_R} |\bar{a}(x^1) - a(x^1, x')|\,dx'\,dx^1$$

$$\leq N R^d a_R^{\#(x')} \leq N(\kappa r)^d a_R^{\#(x')}.$$

Hence

$$\int_{B_{\kappa r}(x_0)} |\bar{L}_0 u - L_0 u|^2\,dx \leq N(\kappa r)^{d/\nu} \left(a_R^{\#(x')}\right)^{1/\nu} \left(\int_{B_{\kappa r}(x_0)} |u_{xx}|^{2\mu}\,dx\right)^{1/\mu}.$$

From this and (5.3) it follows that

$$\left(|\bar{L}_0 u|^2\right)_{B_{\kappa r}(x_0)} \leq N\left[\left(a_R^{\#(x')}\right)^{1/\nu} \left(|u_{xx}|^{2\mu}\right)_{B_{\kappa r}(x_0)}^{1/\mu} + \left(|L_0 u|^2\right)_{B_{\kappa r}(x_0)}\right].$$

This and (5.2) allow us to have

$$\fint_{B_r(x_0)} |u_{xx'} - (u_{xx'})_{B_r(x_0)}|^2\,dx \leq N\kappa^d \left(a_R^{\#(x')}\right)^{1/\nu} \left(|u_{xx}|^{2\mu}\right)_{B_{\kappa r}(x_0)}^{1/\mu}$$

$$+N\kappa^d \left(|L_0 u|^2\right)_{B_{\kappa r}(x_0)} + N\kappa^{-2} \left(|u_{xx}|^2\right)_{B_{\kappa r}(x_0)}$$

$$\leq N\kappa^d \left(a_R^{\#(x')}\right)^{1/\nu} \mathcal{C} + N\kappa^d \mathcal{A} + N\kappa^{-2}\mathcal{B}$$

for all $r > 0$ and $\kappa \geq 4$. In addition, the above inequality is also true for $0 < \kappa < 4$ since then

$$\fint_{B_r(x_0)} |u_{xx'} - (u_{xx'})_{B_r(x_0)}|^2\,dx \leq \fint_{B_r(x_0)} |u_{xx'}|^2\,dx \leq \mathcal{B} \leq 16\kappa^{-2}\mathcal{B}.$$

By taking the supremum with respect to $r > 0$ and then minimizing with respect to $\kappa > 0$, we have

$$\left[u^{\#}_{xx'}(x_0)\right]^2 \le N \left(\left(a_R^{\#(x')}\right)^{1/\nu} \mathcal{C} + \mathcal{A} \right)^{\frac{2}{d+2}} \mathcal{B}^{\frac{d}{d+2}}$$

$$\le N \left(a_R^{\#(x')}\right)^{\frac{2}{\nu(d+2)}} \mathcal{C}^{\frac{2}{d+2}} \mathcal{B}^{\frac{d}{d+2}} + N\, \mathcal{A}^{\frac{2}{d+2}}\, \mathcal{B}^{\frac{d}{d+2}},$$

where $N = N(d, \delta, \mu)$. Notice that $\mathcal{B} \le \mathcal{C}$. Thus, by replacing $\mathcal{B}$ with $\mathcal{C}$ in the first term on the right-hand side, we finish the proof. □

COROLLARY 5.2. *For $p > 2$, there exist constants $R = R(d, \delta, p, \omega)$ and $N = N(d, \delta, p)$ such that, for any $u \in C_0^\infty(B_R)$, we have*

$$\|u_{xx}\|_{L_p} \le N \|L_0 u\|_{L_p}. \tag{5.4}$$

Proof. Choose real numbers $\mu > 1$ such that $p > 2\mu$. Then we use the inequality (5.1) along with the Fefferman–Stein theorem on sharp functions and the Hardy–Littlewood maximal function theorem. We also use Hölder's inequality to have (note that $p/2\mu > 1$ and $p/2 > 1$)

$$\|u_{xx'}\|_{L_p} \le N \left(a_R^{\#(x')}\right)^{\frac{1}{\nu(d+2)}} \|u_{xx}\|_{L_p} + N \|L_0 u\|_{L_p}^{\frac{2}{d+2}} \|u_{xx}\|_{L_p}^{\frac{d}{d+2}}, \tag{5.5}$$

where $1/\mu + 1/\nu = 1$, and N depends only on d, δ, and p. Since

$$u_{x^1x^1} = \frac{1}{a^{11}} L_0 u - \sum_{jk>1} \frac{a^{jk}}{a^{11}} u_{x^j x^k},$$

by using (5.5) we arrive at

$$\|u_{xx}\|_{L_p} \le N_1 \left(a_R^{\#(x')}\right)^{\frac{1}{\nu(d+2)}} \|u_{xx}\|_{L_p} + N \|L_0 u\|_{L_p} + N \|L_0 u\|_{L_p}^{\frac{2}{d+2}} \|u_{xx}\|_{L_p}^{\frac{d}{d+2}}. \tag{5.6}$$

We now invoke Assumption 2.2 by which we can choose a sufficiently small R such that

$$N_1 \left(a_R^{\#(x')}\right)^{\frac{1}{\nu(d+2)}} \le 1/2.$$

Then we have

$$\frac{1}{2} \|u_{xx}\|_{L_p} \le N \|L_0 u\|_{L_p} + N \|L_0 u\|_{L_p}^{\frac{2}{d+2}} \|u_{xx}\|_{L_p}^{\frac{d}{d+2}},$$

which implies (5.4). □

Proof of Theorem 2.4. Since we have an L_p-estimate for functions with small compact support, we can just follow the standard argument, which can be found in [6]. □

Acknowledgments. The authors are sincerely grateful to Hongjie Dong who pointed out an omission in the first draft of the article and to the referees of the paper for careful reading and many useful suggestions.

REFERENCES

[1] F. Chiarenza, M. Frasca, and P. Longo, *Interior $W^{2,p}$ estimates for nondivergence elliptic equations with discontinuous coefficients*, Ricerche Mat., 40 (1991), pp. 149–168.

[2] F. Chiarenza, M. Frasca, and P. Longo, *$W^{2,p}$-solvability of the Dirichlet problem for nondivergence elliptic equations with VMO coefficients*, Trans. Amer. Math. Soc., 336 (1993), pp. 841–853.

[3] G. Chiti, *A $W^{2,2}$ bound for a class of elliptic equations in nondivergence form with rough coefficients*, Invent. Math., 33 (1976), pp. 55–60.

[4] Doyoon Kim, *Second order elliptic equations in $\mathbb{R}^d$ with piecewise continuous coefficients*, Potential Anal., 26 (2007), pp. 182–212.

[5] N. V. Krylov, *The first boundary value problem for elliptic equations of second order*, Differencial′nye Uravnenija, 3 (1967), pp. 315–326 (in Russian).

[6] N. V. Krylov, *Parabolic and elliptic equations with VMO coefficients*, Comm. Partial Differential Equations, to appear, available online at http://arxiv.org/pdf/math.AP/0511731.

[7] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasilinear Equations of Elliptic Type*, 2nd ed., Revised, Izdat. "Nauka," Moscow, 1973 (in Russian).

[8] A. Lorenzi, *On elliptic equations with piecewise constant coefficients*, Applicable Anal., 2 (1972), pp. 79–96.

[9] A. Lorenzi, *On elliptic equations with piecewise constant coefficients*, II, Ann. Scuola Norm. Sup. Pisa, 26 (1972), pp. 839–870.

[10] A. Maugeri, D. K. Palagachev, and L. G. Softova, *Elliptic and Parabolic Equations with Discontinuous Coefficients*, Math. Res. 109, Wiley-VCH Verlag, Berlin, 2000.

[11] N. S. Nadirashvili, *Nonuniqueness in the Martingale problem and the Dirichlet problem for uniformly elliptic operators*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 24 (1997), pp. 537–549.

[12] D. Palagachev and L. Softova, *A priori estimates and precise regularity for parabolic systems with discontinuous data*, Discrete Contin. Dyn. Syst., 13 (2005), pp. 721–742.

[13] M. V. Safonov, *Nonuniqueness for second-order elliptic equations with measurable coefficients*, SIAM J. Math. Anal., 30 (1999), pp. 879–895.

[14] L. Softova, *Quasilinear parabolic operators with discontinuous ingredients*, Nonlinear Anal., 52 (2003), pp. 1079–1093.

[15] L. Softova, *$W^{2,1}_p$-solvability for the parabolic Poincaré problem*, Comm. Partial Differential Equations, 29 (2004), pp 1783–1798.

[16] D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, in Grundlehren Math. Wiss. [Fundamental Principles of Mathematical Sciences] 233, Springer-Verlag, Berlin, 1979.

[17] N. N. Ural'tseva, *On Impossibility of W^2_p Estimates for Multidimensional Elliptic Equations with Discontinuous Coefficients*, Zap. Nauchn. Seminarov LOMI im. Steklova, t. 5, "Nauka," Leningrad, 1967 (in Russian).

[18] M. H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1978.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 507–536

A NEW CLASS OF ENTROPY SOLUTIONS OF THE BUCKLEY–LEVERETT EQUATION*

C. J. VAN DUIJN†, L. A. PELETIER‡, AND I. S. POP†

Abstract. We discuss an extension of the Buckley–Leverett (BL) equation describing two-phase flow in porous media. This extension includes a third order mixed derivatives term and models the dynamic effects in the pressure difference between the two phases. We derive existence conditions for traveling wave solutions of the extended model. This leads to admissible shocks for the original BL equation, which violate the Oleinik entropy condition and are therefore called *nonclassical*. In this way we obtain nonmonotone weak solutions of the initial-boundary value problem for the BL equation consisting of constant states separated by shocks, confirming results obtained experimentally.

Key words. conservation laws, dynamic capillarity, two-phase flows in porous media, shock waves, pseudoparabolic equations

AMS subject classifications. 35L65, 35L67, 35K70, 76S05, 76T05

DOI. 10.1137/05064518X

1. Introduction.

We consider the first order initial-boundary value problem

$$\text{(BL)}\begin{cases} \dfrac{\partial u}{\partial t}+\dfrac{\partial f(u)}{\partial x}=0 & \text{in} \quad Q=\{(x,t): x>0,\, t>0\},\\ u(x,0)=0 & x>0,\\ u(0,t)=u_B & t>0,\end{cases} \tag{1.1}$$

where u_B is a constant such that $0 \le u_B \le 1$. The nonlinearity $f:\mathbf{R}\to\mathbf{R}$ is given by

$$f(u)=\frac{u^2}{u^2+M(1-u)^2} \qquad \text{if} \qquad 0\le u\le 1, \tag{1.2}$$

whilst $f(u)=0$ if $u<0$ and $f(u)=1$ if $u>1$. Here, $M>0$ is a fixed constant. The function $f(u)$ is shown in Figure 1.

Equation (1.1), with the given flux function f, arises in two-phase flow in porous media, and problem (BL) models oil recovery by water-drive in one-dimensional horizontal flow. In this context, $u:\overline{Q}\to[0,1]$ denotes water saturation, f the water fractional flow function, and M the water/oil viscosity ratio. In petroleum engineering, (1.1) is known as the *Buckley–Leverett (BL) equation* [5]. It is a prototype for first order conservation laws with convex-concave flux functions.

It is well known that first order equations such as (1.1) may have solutions with discontinuities, or shocks. The value (u_ℓ) to the left of the shock, the value (u_r) to

*Received by the editors November 15, 2005; accepted for publication (in revised form) January 29, 2007; published electronically July 4, 2007. The authors are grateful for the support from the Centrum voor Wiskunde en Informatica in Amsterdam (CWI), as well as from the Dutch government through the national program BSIK: knowledge and research capacity, in the ICT project BRICKS (http://www.bsik-bricks.nl), theme MSV1.

http://www.siam.org/journals/sima/39-2/64518.html

†Department of Mathematics and Computer Science, Technische Universiteit Eindhoven, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (c.j.v.duijn@tue.nl, i.pop@tue.nl).

‡Mathematical Institute, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands (peletier@math.leidenuniv.nl).

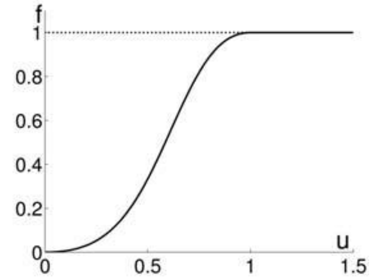


FIG. 1. *Nonlinear flux function for Buckley–Leverett ($M = 2$).*

the right, and the speed s of the shock with trace $x = x(t)$ are related through the *Rankine–Hugoniot condition*,

$$\text{(1.3)} \qquad \text{(RH)} \qquad \frac{dx}{dt} = s = \frac{f(u_\ell) - f(u_r)}{u_\ell - u_r}.$$

We will denote shocks by their values to the left and to the right: $\{u_\ell, u_r\}$.

If a function u is such that (1.1) is satisfied away from the shock curve, and the Rankine–Hugoniot condition is satisfied across the curve, then u satisfies the identity

$$\text{(1.4)} \qquad \int_Q \Big\{ u\frac{\partial \varphi}{\partial t} + f(u)\frac{\partial \varphi}{\partial x} \Big\} = 0 \qquad \text{for all} \qquad \varphi \in C_0^\infty(Q).$$

Functions $u \in L^\infty(Q)$ which satisfy (1.4) are called weak solutions of (1.1). Clearly, for any $u_B \in [0, 1]$, a weak solution of problem (BL) is given by the shock wave

$$\text{(1.5)} \qquad u(x,t) = S(x,t) \overset{\text{def}}{=} \begin{cases} u_B & \text{for} \quad x < st \\ 0 & \text{for} \quad x > st \end{cases} \qquad \text{where} \qquad s = \frac{f(u_B)}{u_B}.$$

Experiments of two-phase flow in porous media reveal complex infiltration profiles, which may involve *overshoot*; i.e., profiles may not be monotone [13]. Our main objective is to understand the shape of these profiles and to determine how the shape depends on the boundary value u_B and the flux function $f(u)$.

Equation (1.1) usually arises as the limit of a family of extended equations of the form

$$\text{(1.6)} \qquad \frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = \mathcal{A}_\varepsilon(u), \qquad \varepsilon > 0,$$

in which $\mathcal{A}_\varepsilon(u)$ is a singular regularization term involving higher order derivatives. It is often referred to as a *viscosity term*. Weak solutions of problem (BL) are called *admissible* when they can be constructed as limits, as $\varepsilon \to 0$, of solutions u_ε of (1.6), i.e., for which $\mathcal{A}_\varepsilon(u_\varepsilon) \to 0$ as $\varepsilon \to 0$ in some weak sense. We return to this limit in section 6. This raises the question of which of the shock waves $S(x,t)$ defined in (1.5) are admissible. We shall see that this depends on the operator $\mathcal{A}_\varepsilon$. To obtain criteria for admissibility we shall use families of traveling wave solutions.

A classical viscosity term is

$$\mathcal{A}_\varepsilon(u) = \varepsilon \frac{\partial^2 u}{\partial x^2},$$

and with this term, (1.6) becomes

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = \varepsilon \frac{\partial^2 u}{\partial x^2}. \tag{1.7}$$

Seeking a traveling wave solution, we put

$$u = u(\eta) \qquad \text{with} \qquad \eta = \frac{x - st}{\varepsilon}, \tag{1.8}$$

and we find that $u(\eta)$ satisfies the following two-point boundary value problem:

$$\begin{cases} -su' + \big(f(u)\big)' = u'' \qquad \text{in} \qquad \mathbf{R}, & \text{(1.9a)} \\ u(-\infty) = u_\ell, \qquad u(\infty) = u_r, & \text{(1.9b)} \end{cases}$$

where primes denote differentiation with respect to η. An elementary analysis shows that problem (1.9) has a solution if and only if f and the limiting values u_ℓ and u_r satisfy (i) the Rankine–Hugoniot condition (1.3), and (ii) the *Oleinik entropy condition* [29]:

$$\text{(E)} \qquad \frac{f(u_\ell) - f(u)}{u_\ell - u} \geq \frac{f(u_\ell) - f(u_r)}{u_\ell - u_r} \qquad \text{for } u \text{ between } u_\ell \text{ and } u_r. \tag{1.10}$$

Shocks $\{u_\ell, u_r\}$ which satisfy (E) are called *classical* shocks.

Note that in the limit as $\varepsilon \to 0^+$, traveling waves converge to the shock $\{u_\ell, u_r\}$.

Applying (RH) and (E) to the flux function (1.2) we find that the function $S(x,t)$ defined in (1.5) is an admissible shock wave if and only if

$$s = \frac{f(u_B)}{u_B} \quad \text{(RH)} \qquad \text{and} \qquad u_B \leq \alpha \quad \text{(E)}, \tag{1.11}$$

where α is the unique root of

$$f'(u) = \frac{f(u)}{u}.$$

It is found to be given by

$$\alpha = \sqrt{\frac{M}{M+1}} \in (0,1).$$

If $u_B > \alpha$, then the weak solution is composed of a rarefaction wave in the region where $u > \alpha$ and a shock which spans the range $0 < u < \alpha$. Thus, for any $u_B \in (0,1]$ the weak solution $u(x,t)$ is, at any given time t, a nonincreasing function of x, in contrast to the experimental data for infiltration in porous media.

For gaining a better understanding of the data, it is natural to go back to the origins of (1.1). With S_i $(i = o, w)$ being the saturations of the two phases, oil and water, conservation of mass yields

$$\phi \frac{\partial S_i}{\partial t} + \frac{\partial q_i}{\partial x} = 0, \qquad i = o, w, \tag{1.12}$$

where q_i denotes the specific discharge of oil/water and ϕ the porosity of the medium. By Darcy's law, q_i is proportional to the gradient of the phase pressure P_i:

$$q_i = -k \frac{k_{ri}(S_i)}{\mu_i} \frac{\partial P_i}{\partial x}, \tag{1.13}$$

where k denotes the absolute permeability and k_{ri} and μ_i the relative permeability and the viscosity of water, respectively, oil. The capillary pressure P_c expresses the difference in the pressures of the two phases:

$$P_c = P_o - P_w. \tag{1.14}$$

This quantity is commonly found to depend on one phase saturation, say S_w. In addition to this, studies like [28] and [30] show that P_c does not only depend on S_w, but also involves hysteretic and dynamic effects. Hassanizadeh and Gray [19, 20] have defined the dynamic capillary pressure as

$$P_c = p_c(S_w) - \phi\tau\frac{\partial S_w}{\partial t}, \tag{1.15}$$

where $p_c(S_w)$ is the *static* capillary pressure and τ a positive constant. Assuming that the medium is completely saturated,

$$S_w + S_o = 1,$$

and we obtain, upon combining (1.12)–(1.15), the single equation for the water saturation $u = S_w$:

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = -\frac{\partial}{\partial x}\left\{H(u)\frac{\partial}{\partial x}\left(J(u) - \tau\frac{\partial u}{\partial t}\right)\right\}, \tag{1.16}$$

in which the functions f, H, and J are related to k_{ri} and p_c. Other noneqilibrium models are considered in [3]. Restricting, for simplicity, to linear terms on the right-hand side of (1.16), we obtain, after a suitable scaling, the pseudoparabolic equation

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = \varepsilon\frac{\partial^2 u}{\partial x^2} + \varepsilon^2\tau\frac{\partial^3 u}{\partial x^2 \partial t}. \tag{1.17}$$

Thus, in addition to the classical second order term εu_{xx}, we find a third order term $\varepsilon^2\tau u_{xxt}$, its relative importance being determined by the parameter τ. We show that the value of τ is critical in determining the type of profile the solution of problem (BL) will have.

The right-hand side of (1.17) resembles the regularization $\mathcal{A}_\varepsilon(u) = \varepsilon u_{xx} + \varepsilon^2\delta u_{xxx}$, which has received considerable attention (cf. [23] and the monograph [26] and the references cited therein). We mention in particular the seminal paper [24] in which $f(u) = u^3$. There, for $\delta > 0$ an explicit function $\varphi(u;\delta)$ is derived such that the shock $\{u_\ell, u_r\}$ is admissible if and only if $u_r = \varphi(u_\ell;\delta)$. Properties of this *kinetic function* φ, such as monotonicity with respect to u_ℓ, have been studied in a series of papers (see [6] and the references cited there).

Other regularizations have been studied in [7] and [8] where a fourth order viscosity term was introduced motivated by thin film flow ($\mathcal{A}_\varepsilon(u) = -(u^3 u_{xxx})_x$, and the flux function is $f(u) = u^2 - u^3$) and in [18], where fourth order regularizations are used, motivated by problems in image processing. Traveling waves for dynamic capillarity models, but for a convex flux function, are investigated in [11].

In this paper we focus on the relation between u_ℓ and the parameter τ. With β being defined in Proposition 1.1 (see also Figure 2) we establish the existence of a function $\tau(u_\ell)$ defined for $\alpha < u_\ell < \beta$ such that (1.17) has a traveling wave solution with $u_r = 0$, if and only if $\tau = \tau(u_\ell)$. We shall show that this function is *monotone*, *continuous*, and has limits

$$\tau(u_\ell) \to \tau_* > 0 \quad \text{as} \quad u_\ell \searrow \alpha \qquad \text{and} \qquad \tau(u_\ell) \to \infty \quad \text{as} \quad u_\ell \nearrow \beta.$$

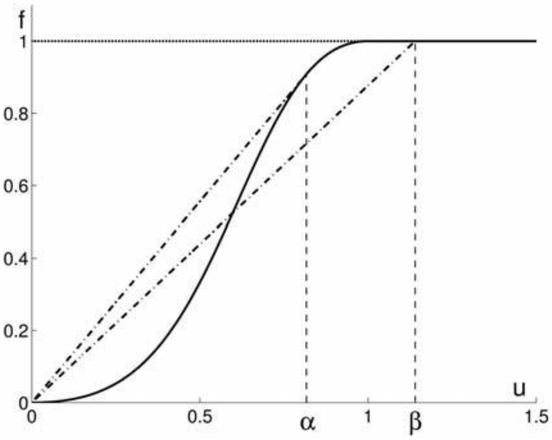


FIG. 2. *Critical values of u when $M = 2$: $\alpha \approx 0.816$ and $\beta \approx 1.147$.*

Thus τ serves as a bifurcation parameter: for $0 < \tau \leq \tau_*$ the situation will be much like in the classical case (E), but for $\tau > \tau_*$ the situation changes abruptly and new types of shock waves become admissible. Note that in the framework of [26] we have $0 = \varphi(u_\ell; \tau(u_\ell))$.

The properties of the function $\tau(u_\ell)$ will be based on three existence, uniqueness, and nonexistence theorems, Theorems 1.1, 1.2, and 1.3, for traveling waves of (1.17). Substituting (1.8) into (1.17) we obtain

$$-su' + \big(f(u)\big)' = u'' - s\tau u''' \qquad \text{in} \qquad \mathbf{R}.$$

When we integrate this equation over (η, ∞), we obtain the second order boundary value problem

$$\text{(1.18a)}\atop\text{(1.18b)} \qquad (\mathrm{TW}) \begin{cases} -s(u - u_r) + \{f(u) - f(u_r)\} = u' - s\tau u'' & \text{in} \qquad \mathbf{R}, \\ u(-\infty) = u_\ell, \qquad u(\infty) = u_r, \end{cases}$$

where $s = s(u_\ell, u_r)$ is given by the Rankine–Hugoniot condition (1.3).

We consider two cases:

$$\text{(I)} \;\; u_r = 0, \quad u_\ell > 0 \qquad \text{and} \qquad \text{(II)} \;\; u_r > u_\ell > 0.$$

Case I. $u_r = 0$. We first establish an upper bound for u_ℓ.

PROPOSITION 1.1. *Let u be a solution of problem* (TW) *such that $u_r = 0$. Then, $u_\ell < \beta$, where β is the value of u for which the equal area rule holds:*

$$\text{(1.19)} \qquad \int_0^\beta \left\{ f(u) - \frac{f(\beta)}{\beta} u \right\} du = 0.$$

In Figure 2 we indicate the different critical values of u in a graph of $f(u)$ when $M = 2$.

Proof. When we put $u_r = 0$ into (1.18a), multiply by u', and integrate over $\mathbf{R}$, we obtain the inequality

$$\int_0^{u_\ell} \left\{ f(u) - \frac{f(u_\ell)}{u_\ell} u \right\} du = -\int_{\mathbf{R}} (u')^2(\eta)\, d\eta < 0,$$

from which it readily follows that $u_\ell < \beta$. ☐

Next, we turn to the questions of existence and uniqueness. Note that if $u_\ell \in (\alpha, \beta)$, then

$$s = s(u_\ell, 0) = \frac{f(u_\ell)}{u_\ell} > f'(u_\ell) \geq f'(0) \qquad \text{for} \qquad u_\ell > \alpha,$$

and traveling waves, if they exist, lead to an admissibility condition for *fast undercompressive waves*. For convenience we write $s(u_\ell, 0) = s(u_\ell)$.

In the theorems below we first show that for each $\tau > 0$, there exists a unique value of $u_\ell \geq \alpha$, denoted by $\overline{u}(\tau)$, for which there exists a solution of problem (TW) such that $u_r = 0$.

THEOREM 1.1. *Let $M > 0$ be given. Then there exists a constant $\tau_* > 0$ such that the following hold:*

(a) *For every $0 \leq \tau \leq \tau_*$, problem* (TW) *has a unique solution with $u_\ell = \alpha$ and $u_r = 0$.*

(b) *For each $\tau > \tau_*$ there exists a unique constant $\overline{u}_\ell(\tau) \in (\alpha, \beta)$ such that problem* (TW) *has a unique solution with $u_\ell = \overline{u}_\ell(\tau)$ and $u_r = 0$.*

(c) *The function $\overline{u} : [0, \infty) \to [\alpha, \beta)$ defined by*

$$\overline{u}(\tau) = \begin{cases} \alpha & \textit{for} \quad 0 \leq \tau \leq \tau_*, \\ \overline{u}_\ell(\tau) & \textit{for} \quad \tau > \tau_* \end{cases} \tag{1.20}$$

is continuous, strictly increasing for $\tau \geq \tau_*$, and $\overline{u}(\infty) = \beta$.

The solutions in parts (a) *and* (b) *are strictly decreasing.*

We shall refer to $\overline{u} = \overline{u}(\tau)$ as the *plateau value* of u. In what follows, we shall often denote the speed $s(\overline{u})$ of the shock $\{\overline{u}, 0\}$ by $\overline{s}$.

Next, suppose that $u_\ell \neq \overline{u}(\tau)$. To deal with this case we need to introduce another critical value of u, which we denote by $\underline{u}(\tau)$.

- For $\tau \in [0, \tau_*]$ we put $\underline{u}(\tau) = \alpha$.
- For $\tau > \tau_*$ we define $\underline{u}(\tau)$ as the unique zero in the interval $(0, \overline{u}(\tau))$ of

$$f(r) - \frac{f(\overline{u})}{\overline{u}} r = 0, \qquad 0 < r < \overline{u}.$$

Plainly, if $\tau > \tau_*$, then

$$0 < \underline{u}(\tau) < \alpha < \overline{u}(\tau) < \beta \qquad \text{for} \quad \tau > \tau_*.$$

In Figure 3, we show graphs of the functions $\overline{u}(\tau)$ and $\underline{u}(\tau)$. They are computed numerically for $M = 2$ by means of a shooting technique that is explained in section 3. In this case we found

$$\tau_* \approx 0.61.$$

The following theorem states that if $u_r = 0$ and $u_\ell \in (0, \overline{u})$, then traveling waves exist if and only if $u_\ell < \underline{u}(\tau)$.

THEOREM 1.2. *Let $M > 0$ and $\tau > 0$ be given, and let $\underline{u} = \underline{u}(\tau)$ and $\overline{u} = \overline{u}(\tau)$.*

(a) *For any $u_\ell \in (0, \underline{u})$, there exists a unique solution of problem* (TW) *such that $u_r = 0$. We have $s(u_\ell) < \overline{s}$.*

(b) *Let $\tau > \tau_*$. Then for any $u_\ell \in (\underline{u}, \overline{u})$, there exists no solution of problem* (TW) *such that $u_r = 0$.*

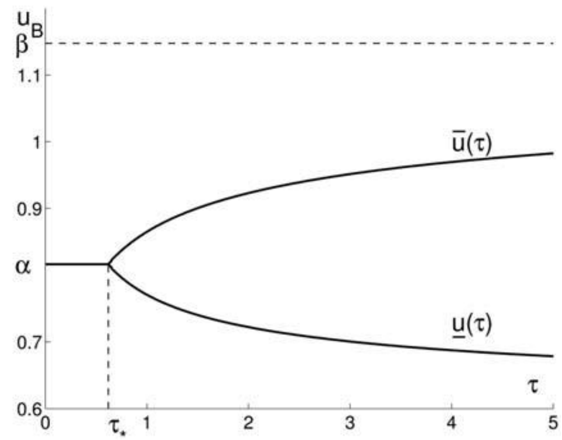


FIG. 3. *The functions $\overline{u}(\tau)$ and $\underline{u}(\tau)$ computed for $M = 2$.*

The solution in part (a) *may exhibit a damped oscillation as it tends to u_ℓ.*

Case II. $u_r > 0$. The results of Case I raise the question as to how to deal with solutions of problem (BL) when $u_B \in (\underline{u}, \overline{u})$, and by Theorem 1.2 there is no traveling wave solution with $u_r = 0$. In this situation we use two traveling waves in succession: one from u_B to the plateau value $\overline{u}$, and one from $\overline{u}$ down to $u = 0$. The existence of the latter has been established in Theorem 1.1. In the next theorem we deal with the former, in which $u_r = \overline{u}$.

THEOREM 1.3. *Let $M > 0$ and $\tau > \tau_*$ be given, and let $\underline{u} = \underline{u}(\tau)$ and $\overline{u} = \overline{u}(\tau)$.*

(a) *For any $u_\ell \in (\underline{u}, \overline{u})$, there exists a unique solution of problem* (TW) *such that $u_r = \overline{u}$. We have $s(u_\ell, \overline{u}) < \overline{s}$.*

(b) *For any $u_\ell \in (0, \underline{u})$, there exists no solution of problem* (TW) *such that $u_r = \overline{u}$.*

The solution in part (a) *may exhibit a damped oscillation as it tends to u_ℓ.*

In section 2 we show how these theorems can be used to construct weak solutions of problem (BL), i.e., weak solutions, which are admissible within the context of the regularization proposed in (1.17), and which involve shocks which may be either classical or nonclassical. In section 3 we solve the Cauchy problem for (1.17) numerically, starting from a smoothed step function, i.e., $u(x, 0) = u_B\tilde{H}(-x)$, where $\tilde{H}(x)$ is a regularized Heaviside function and $M = 2$. We find that for different values of the parameters u_B, τ, and ε the solution converges to solutions constructed in section 2 as $t \to \infty$. In sections 4 and 5 we prove Theorems 1.1, 1.2, and 1.3. The proofs rely on phase plane arguments. We conclude this paper with a discussion of the dissipation of the entropy function $u^2/2$ when u is the solution of the Cauchy problem for (1.17) (cf. section 6).

In this paper we have seen that nonmonotone traveling waves such as those observed in [13] may be explained by a regularization that takes into account properties of two-phase flow. It will be interesting to determine to what extent such results as derived in this paper for the simplified equation (1.17), continue to hold for the full equation (1.16) when realistic functions $H(u)$ and $J(u)$ are used. Such equations may be degenerate at $u = 0$ as well as at $u = 1$, and singular behavior, as in the *porous media equation* [2, 4, 27] may be expected. In this connection it is interesting to mention a numerical study of traveling waves of the original, fully nonlinear equations of this model in [14, 15].

2. Entropy solutions of problem (BL). In this section we give a classification of admissible solutions of problem (BL) based on the "extended viscosity model" (1.17), using the results about traveling wave solutions formulated in Theorems 1.1, 1.2, and 1.3. Before doing that we make a few preliminary observations, and we recall the construction based on the classical model (1.7).

Because (1.1) is a first order partial differential equation and u_B is a constant, any solution of problem (BL) depends only on the combination x/t, with shocks, constant states, and rarefaction waves as building blocks [29]. The latter are continuous solutions of the form

$$u(x,t) = r(\zeta) \qquad \text{with} \qquad \zeta = \frac{x}{t}. \tag{2.1}$$

After substitution into (1.1) this yields

$$\frac{dr}{d\zeta}\left(-\zeta + \frac{df}{du}\big(r(\zeta)\big)\right) = 0. \tag{2.2}$$

Hence, the function $r(\zeta)$ satisfies

$$\textit{either} \quad r = \text{constant} \qquad \textit{or} \qquad \frac{df}{du}\big(r(\zeta)\big) = \zeta.$$

When solving problem (BL), we will combine solutions of (2.2) with admissible shocks, i.e., shocks $\{u_\ell, u_r\}$ in which u_ℓ and u_r are such that (1.6), with the a priori selected and physically relevant viscous extension $\mathcal{A}_\varepsilon$, has a traveling wave solution $u(\eta)$ such that $u(\eta) \to u_\ell$ as $\eta \to -\infty$ and $u(\eta) \to u_r$ as $\eta \to +\infty$. Although in the physical context in which the viscous extension employed in (1.17) was derived, $0 \le u_B \le 1$, we shall drop this restriction. It will be convenient to first assume that $0 \le u_B \le \beta$. At the end of this section we discuss the case that $u_B > \beta$.

All solution graphs shown in this section and the next are numerically obtained solutions of (1.17). They are expressed in terms of the independent variable ζ and t, i.e.,

$$u(x,t) = w(\zeta, t),$$

and considered for fixed $\varepsilon > 0$ $(= 1)$ and for large times t. We return to the computational aspects in section 3.

Before discussing the implications of the viscous extension in (1.17), we recall the construction of *classical entropy solutions* of problem (BL). It uses (RH) and the entropy condition (E), which was derived for the diffusive viscous extension used in (1.8). We distinguish two cases:

$$\text{(a)} \qquad 0 \le u_B \le \alpha \qquad \text{and} \qquad \text{(b)} \qquad \alpha < u_B \le \beta.$$

Case (a). $0 \le u_B \le \alpha$. This case was discussed in the introduction, where we found that the entropy solution is given by the shock $\{u_B, 0\}$.

Case (b). $\alpha < u_B \le \beta$. In the introduction we saw that in this case, the shock $\{u_B, 0\}$ is no longer a classical entropy solution. Instead, in this case the entropy solution is a composition of three functions:

$$u(x,t) = v(\zeta) = \begin{cases} u_B & \text{for} \quad 0 \le \zeta \le \zeta_B, \\ r(\zeta) & \text{for} \quad \zeta_B \le \zeta \le \zeta_*, \\ 0 & \text{for} \quad \zeta_* \le \zeta < \infty, \end{cases} \tag{2.3}$$

where ζ_B and ζ_* are determined by

$$\zeta_B = \frac{df}{du}(u_B) \qquad \text{and} \qquad \zeta_* = \frac{df}{du}(\alpha) = \frac{f(\alpha)}{\alpha} = s(\alpha),$$

and $r : [\zeta_B, \zeta_*] \to [\alpha, u_B]$ by the relation

$$\frac{df}{du}\big(r(\zeta)\big) = \zeta \qquad \text{for} \qquad \zeta_B \le \zeta \le \zeta_*. \tag{2.4}$$

Since $f''(u) < 0$ for $u \in [\alpha, u_B]$, (2.4) has a unique solution, and hence $r(\zeta)$ is well defined. Note that if $u_B \ge 1$, then $\zeta_B = 0$, because $f'(u) = 0$ if $u \ge 1$.

Solutions corresponding to Case (b) are shown in Figure 4.

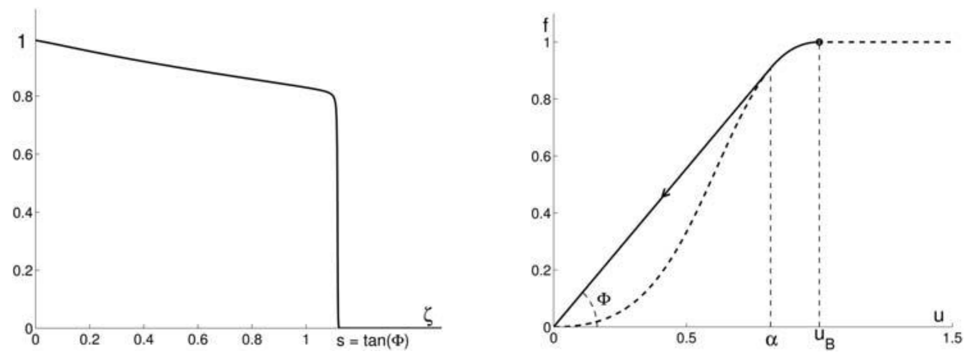


FIG. 4. *Case* (b). *Solution graph (left) and flux function with transitions from* u_B *to* α *and from* α *to* 0 *(right).*

We now turn to the pseudoparabolic equation (1.17) that arises in the context of the two-phase flow model of Hassanizadeh and Gray [19, 20]. For this problem, we define a class of nonclassical entropy solutions in which shocks are admissible if problem (TW) has a traveling wave solution with the required limit conditions.

For given $M > 0$ and $\tau > 0$, the relative values of u_B and $\overline{u}(\tau)$ and $\underline{u}(\tau)$ are now important for the type of solution we are going to get. It is easiest to represent them in the (u_B, τ)-plane. Specifically, we distinguish three regions in this plane:

$$\begin{aligned}
\mathcal{A} &= \{(u_B, \tau) : \tau > 0, \quad \overline{u}(\tau) \le u_B < \beta\},\\
\mathcal{B} &= \{(u_B, \tau) : \tau > \tau_*, \quad \underline{u}(\tau) < u_B < \overline{u}(\tau)\},\\
\mathcal{C} &= \{(u_B, \tau) : \tau > 0, \quad 0 < u_B < \underline{u}(\tau)\}.
\end{aligned}$$

These three regions are shown in Figure 5.

Case I. $(u_B, \tau) \in \mathcal{A}$. If $0 \le \tau \le \tau_*$, i.e., $(u_B, \tau) \in \mathcal{A}_1$, the construction is as in the classical case described above. After a plateau, where $u = u_B$ and $0 \le \zeta = x/t \le \zeta_B$, we find a rarefaction wave $r(\zeta)$ from u_B down to α followed by a classical shock connecting α to the initial state $u = 0$.

If $\tau > \tau_*$, i.e., $(u_B, \tau) \in \mathcal{A}_2$, the solution starts out as before, with a plateau where $u = u_B$ and $0 \le \zeta \le \zeta_B$ and a rarefaction wave $r(\zeta)$ which now takes u down from u_B to $\overline{u} > \alpha$. This takes place over the interval $\zeta_B \le \zeta \le \overline{\zeta}$. By (2.2),

$$\overline{\zeta} = \frac{df}{du}\big(\overline{u}(\tau)\big).$$

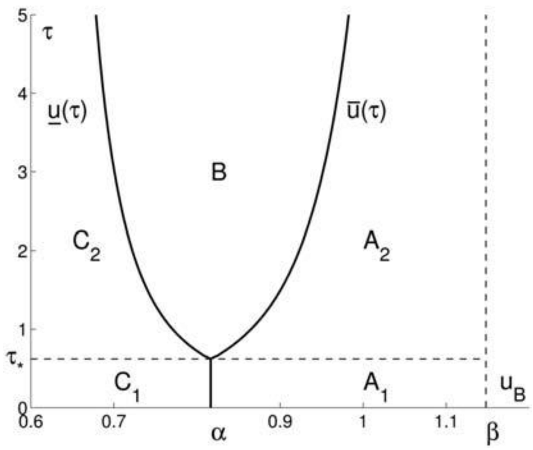


FIG. 5. *The regions $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ in the (u_B, τ)-plane.*

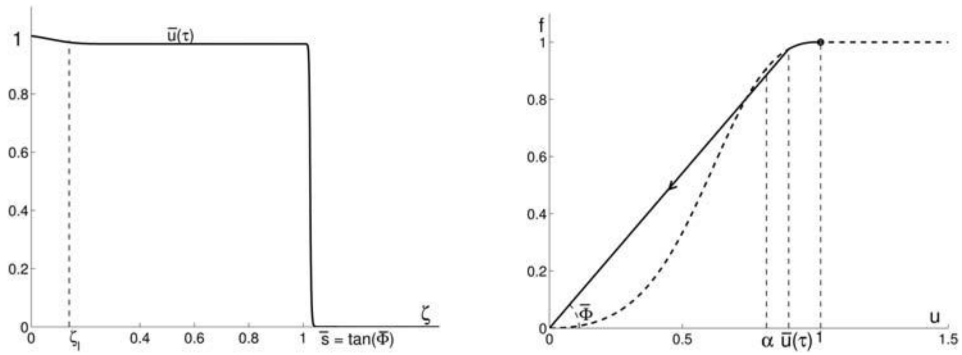


FIG. 6. *Case* I. *Solution graph (left) and flux function (right), with transitions from $u_B = 1$ to $\overline{u}(\tau)$ and from $\overline{u}(\tau)$ to* 0.

Subsequently, u drops down to the initial state $u = 0$ through a shock, $\{\overline{u}, 0\}$, which is admissible by Theorem 1.1. By (RH) the shock moves with speed

$$s = \overline{s} = \frac{f(\overline{u})}{\overline{u}} > \frac{df}{du}(\overline{u}) = \overline{\zeta},$$

because f is concave on (α, ∞). Therefore, the shock outruns the rarefaction wave and a second plateau develops between the rarefaction wave and the shock in which $u = \overline{u}$. Summarizing, we find that the (nonclassical) entropy solution has the form

$$u(x,t) = v(\zeta) = \begin{cases} u_B & \text{for} \quad 0 \le \zeta \le \zeta_B, \\ r(\zeta) & \text{for} \quad \zeta_B \le \zeta \le \overline{\zeta}, \\ \overline{u}(\tau) & \text{for} \quad \overline{\zeta} \le \zeta \le \overline{s}, \\ 0 & \text{for} \quad \overline{s} \le \zeta < \infty. \end{cases} \tag{2.5}$$

A graph of $v(\zeta)$ is given in Figure 6.

Note that if $u_B \ge 1$, then $\zeta_B = 0$. At this point v shocks to the maximum of $\overline{u}(\tau)$ and 1. If $\overline{u}(\tau) \ge 1$, then the rarefaction wave disappears and for $\zeta > 0$ the solution is continued by the shock $\{\overline{u}(\tau), 0\}$.

Case II. $(u_B, \tau) \in \mathcal{B}$. It follows from Theorem 1.2 that there are no traveling wave solutions with $u_\ell = u_B$ and $u_r = 0$, so that the shock $\{u_B, 0\}$ is now not admissible. However, in Theorem 1.3 we have shown that there does exist a traveling wave solution, and hence an admissible shock, with $u_\ell = u_B$ and $u_r = \overline{u}(\tau)$, and speed $s = s(u_B, \overline{u}(\tau))$. This shock is then followed by a second shock from $u = \overline{u}(\tau)$ down to $u = 0$, which is admissible because by Theorem 1.1 there does exist a traveling wave solution which connects $\overline{u}$ and $u = 0$ with speed $\overline{s} > s(u_B, \overline{u}(\tau))$. Thus

$$u(x,t) = v(\zeta) = \begin{cases} u_B & \text{for} \quad 0 \le \zeta \le s(u_B, \overline{u}), \\ \overline{u}(\tau) & \text{for} \quad s(u_B, \overline{u}) \le \zeta \le \overline{s}, \\ 0 & \text{for} \quad \overline{s} \le \zeta < \infty. \end{cases} \tag{2.6}$$

An example of this type of solution is shown in Figure 7. The undershoot in the solution graph is due to oscillations which are also present in the traveling waves.

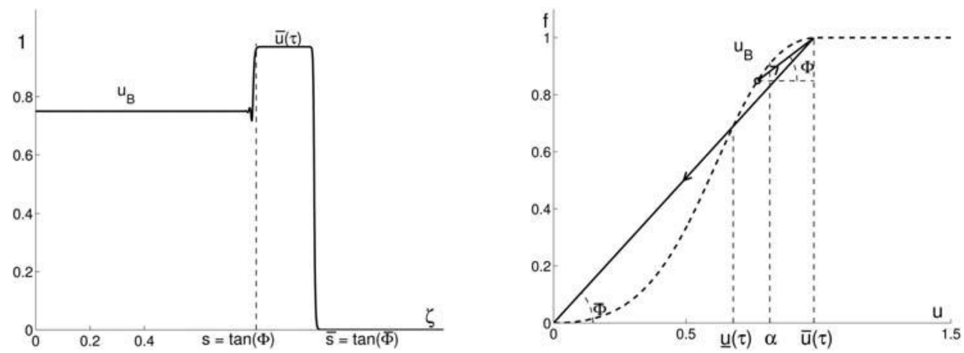


FIG. 7. *Case* II. *Solution graph (left) and flux function (right), with transitions from* $u_B = 0.75$ *to* $\overline{u}(\tau)$ *and from* $\overline{u}(\tau)$ *to* 0.

Remark 2.1. It is readily seen that

$$s(u_B, \overline{u}(\tau)) \nearrow \overline{s} \qquad \text{as} \qquad u_B \searrow \underline{u},$$

while the plateau level $\overline{u}$ remains the same. Thus, in this limit, the plateau

$$\left\{ \left(u, \frac{x}{t}\right) : u = \overline{u}(\tau),\ s(u_B, \overline{u}(\tau)) < \frac{x}{t} < \overline{s} \right\}$$

becomes thinner and thinner and eventually disappears when $u_B = \underline{u}$.

Remark 2.2. If $u_B = 1$ and $\overline{u}(\tau) > 1$, then the first shock degenerates in the sense that

$$s(u_B, \overline{u}(\tau)) = 0 \qquad \text{and} \qquad u(x,t) = \overline{u}(\tau) \quad \text{for all} \quad 0 < \frac{x}{t} < \overline{s}.$$

Case III. $(u_B, \tau) \in \mathcal{C}$. We have seen in Theorem 1.2 that in this case there exists a traveling wave solution with $u_\ell = u_B$ and $u_r = 0$. It may exhibit oscillatory behavior near $u = u_\ell$, and it leads to the classical entropy shock solution $\{u_B, 0\}$. An example of such a solution is shown in Figure 8. Note the overshoot in the solution graph, reflecting oscillations also present in the traveling waves.

We conclude with a remark about the case when $u_B > \beta$. It is readily verified that for such values of u_B the situation is completely analogous to the one for $(u_B, \tau) \in \mathcal{A}$.

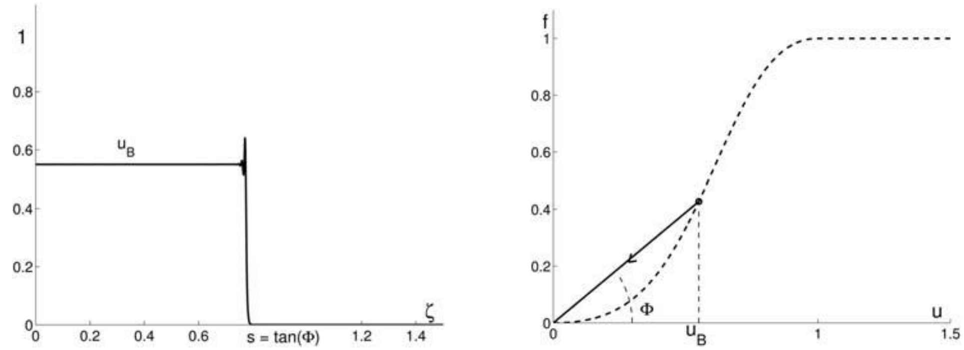


FIG. 8. *Case* III. *Solution graph (left) and flux function (right), with transition from* $u_B = 0.55$ *to* 0.

3. Numerical experiments for large times. In this section we report on the computations carried out for obtaining the numerical results presented in this paper. All computations are done for $M = 2$. We start with the calculation of the diagram in Figure 3. For determining the graphs of $\overline{u}$ and $\underline{u}$ as functions of τ we fix $u_r = 0$. Then, given a $\tau > 0$ and a left state $u_\ell \geq 0$, we look for a strictly decreasing solution $u(\eta)$ of the problem (1.18a) and (1.18b). If such a solution exists, we can invert the function $u(\eta)$ and define the new dependent variable $z(u) = -u'(\eta(u))$, which satisfies

$$s\tau zz' + z = su - f(u)$$

on the open interval $(0, u_\ell)$. Moreover, we have $z > 0$ on $(0, u_\ell)$, and $z(0) = z(u_\ell) = 0$.

Following Theorem 1.1, an $\tau_* > 0$ exists so that solutions z to the given first order equation and boundary conditions are possible for any $\tau \leq \tau_*$, and with $u_\ell = \alpha$. To compute τ_* we fix $u_\ell = \alpha$ and solve the equation in z with $z(0) = 0$. We start with a sufficiently small $\tau > 0$ and increase its value until $z(u_\ell)$ becomes strictly positive. This gives

$$\tau_* \approx 0.61.$$

Further, for $\tau > \tau_*$ there is a unique $u_\ell = \overline{u}(\tau) \in (\alpha, \beta)$ yielding a solution z with the required properties. Moreover, $\overline{u}$ is strictly increasing in τ. For finding the corresponding u_ℓ we solve numerically the equation in z with the initial value $z(0) = 0$. We repeat this procedure for different values of τ, starting close to τ_* and increasing gradually the difference between two successive values of τ as the corresponding u_ℓ approaches β. Accurate computations with different ODE solvers have led to negligible differences in the resulting diagrams. Finally, the function $\underline{u}(\tau)$ follows from a simple construction involving $f(u)$.

Nonstandard shock solutions of a hyperbolic conservation are computed numerically in [21] and [22]. The schemes considered there are applied to the hyperbolic problem, but they actually solve more accurately a regularized problem involving a ∂_{xxx} term. This term vanishes as the discretization parameters are approaching 0.

Here we consider the regularized initial value problem for (1.17) in the domain $S = \mathbf{R} \times \mathbf{R}^+$:

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial f(u)}{\partial x} = \varepsilon \dfrac{\partial^2 u}{\partial x^2} + \varepsilon^2 \tau \dfrac{\partial^3 u}{\partial x^2 \partial t} & \text{in} \quad S, \qquad (3.1a) \\ u(x,0) = u_B \tilde{H}(-x) & \text{for} \quad x \in \mathbf{R}. \qquad (3.1b) \end{cases}$$

Here $\tilde{H}(x)$ is a smooth monotone approximation of the Heaviside function H. We use $\tilde{H}$ instead of H because discontinuities in the initial conditions will persist for all $t > 0$, as shown in [12]. This would require an adapted and more complicated numerical approach for ensuring the continuity in flux and pressure (see, for example, [10], or [9, Chapter 3]). By the above choice we avoid this unnecessary complication.

Important parameters in this problem are M, ε, $\tau > 0$, and $u_B \in (0,1]$. The scaling

$$x \to \frac{x}{\varepsilon}, \qquad t \to \frac{t}{\varepsilon} \tag{3.2}$$

removes the parameter ε from (3.1a). Therefore, we fix $\varepsilon = 1$ and show how for different values of τ and u_B the solution $u(x,t)$ of problem (3.1) converges as $t \to \infty$ to qualitatively different final profiles.

For solving (3.1) numerically we consider a first order time stepping, combined with the finite difference discretization of the terms involving ∂_{xx}. To deal with the first order term we apply a minmod slope limiter method that is based on first order upwinding and Richtmyer's scheme. Specifically, with $k > 0$ and $h > 0$ being the discretization parameters, we define $x_i = ih$ ($i \in \mathbb{Z}$) and $t_n = nk$ ($n \in \mathbb{N}$), and let u_i^n stand for the numerical

approximation of $u(x_i, t_n)$. With

$$\Delta_{h,i}(u) := \frac{1}{h^2}\big(u_{i+1} - 2u_i + u_{i-1}\big),$$

the fully discrete counterpart of (3.1a) in (x_i, t_k) reads

$$u_i^{n+1} - u_i^n + \frac{k}{h}\big(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}}\big) = k\varepsilon\Delta_{h,i}(u^{n+1}) + \varepsilon^2\tau\Delta_{h,i}(u^{n+1} - u^n),$$

where $F^n_{i+\frac{1}{2}}$ is the numerical flux at $x_{i+\frac{1}{2}} = x_i + \frac{h}{2}$ and t_n. As mentioned above, $F^n_{i+\frac{1}{2}}$ is a convex combination of the first order upwind flux and the second order Richtmyer flux:

$$F^n_{i+\frac{1}{2}} = (1 - \Theta_i^n)F^{n,low}_{i+\frac{1}{2}} + \Theta_i^n F^{n,high}_{i+\frac{1}{2}},$$

where

$$\begin{aligned} F^{n,low}_{i+\frac{1}{2}} &= f(u_i^n), \\ F^{n,high}_{i+\frac{1}{2}} &= f(w_i^n), \qquad \text{for } w_i^n = \tfrac{u_i^n + u_{i+1}^n}{2} - \tfrac{k}{2h}\big(f(u_{i+1}^n) - f(u_i^n)\big), \end{aligned}$$

and

$$\Theta_i^n = \max(0, \min(1, \theta_i^n)), \qquad \text{for } \theta_i^n = \begin{cases} 0 & \text{if } u_i^n = u_{i+1}^n, \\ \frac{u_i^n - u_{i-1}^n}{u_{i+1}^n - u_i^n} & \text{otherwise.} \end{cases}$$

To compute the numerical solution, we restrict (3.1) to the sufficiently large spatial interval $(-1000, 5000)$, and define the artificial boundary conditions $u(-1000,t) = u_B$ and $u(5000,t) = 0.0$. The computations are performed for large times ($t > 2000$), as long as the results are not affected by the presence of the boundaries. We apply the discretization scheme mentioned above, yielding a linear tridiagonal system that is solved at each time step. A convergence proof for the numerical scheme is beyond the scope of the present work. Similar numerical schemes for problems of pseudoparabolic type are considered, for example, in [1] and [16], where also the convergence is proven.

In the figures below, we show graphs of solutions at various times t, appropriately scaled in space. Specifically, we show graphs of the function

$$w(\zeta, t) = u(x, t), \qquad \text{where} \qquad \zeta = \frac{x}{t}, \tag{3.3}$$

so that a front with speed s will be located at $\zeta = s$.

We recall that the numerical results are obtained for $M = 2$. In this case $\tau_* \approx 0.61$ (see also Figures 3 and 5). We begin with a simulation where $(u_B, \tau) = (1, 0.2) \in \mathcal{A}_1$. In Figure 9 we show the resulting solution $w(\zeta, t)$ at time $t = 1000$. It is evident that w converges to the classical entropy solution constructed in section 2.

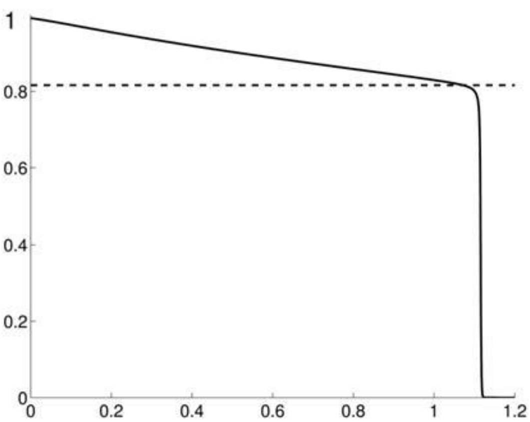


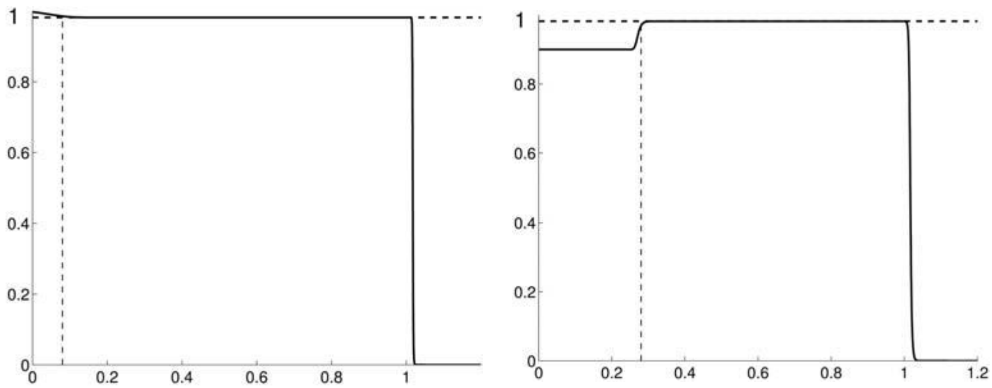
FIG. 9. *Graph of* $w(\zeta, t)$ *at* $t = 1000$ *when* $(u_B, \tau) = (1, 0.2) \in \mathcal{A}_1$. *In this case* $\overline{u}(\tau) = \alpha \approx 0.816$ *and* $s \approx 1.11$.

In the simulations that we present in the remainder of this section we take τ to be fixed above τ_*: $\tau = 5$. Correspondingly, by the ODE method involved in computing the diagram in Figure 3 we obtained $\overline{u}(\tau = 5) \approx 0.98$ and $\underline{u}(\tau = 5) \approx 0.68$. In the first of these experiments, in which we keep $u_B = 1$, we see that for large time the graph consists of three pieces: one in which w gradually decreases from $w = u_B = 1$ to the "plateau" value $w = \overline{u}$, one in which w is constant and equal to $\overline{u}$, and one in which it drops down to $u = 0$; see Figure 10(a). It is clear from the graph that $\overline{u} > \alpha$. The plateau value $\overline{u} \approx 0.98$ computed here is in excellent agreement with the value obtained by the ODE method; see also Figures 3 and 5.

In the next experiment we decrease u_B to $u_B = 0.9$. We are then in the region $\mathcal{B}$. For large times the solution $w(\zeta, t)$ develops two shocks, one where it jumps up from u_B to the plateau at $\overline{u} \approx 0.98$ (the same value as in the previous experiment), and one where it jumps down from $\overline{u}$ to $w = 0$; see Figure 10(b).

In the next experiments we decrease the value of u_B to values around the value $\underline{u} \approx 0.68$. The results are shown in Figure 11, where we have zoomed into the front. We see that, as u_B decreases and approaches the boundary between the regions $\mathcal{B}$ and $\mathcal{C}_2$ in Figure 5, the part of the graph where $w \approx u_B$ grows at the expense of the part where $w \approx \overline{u}$.

Finally, in Figure 12 we show the graph of $w(\zeta, t)$ when $\tau = 5$ and u_B is further



(a) $(u_B, \tau) = (1, 5) \in \mathcal{A}_2$ (b) $(u_B, \tau) = (0.9, 5) \in \mathcal{B}$

FIG. 10. *Graphs of* $w(\zeta, t)$ *at* $t = 1000$ *when* $(u_B, \tau) = (1, 5) \in \mathcal{A}_2$ *(left) and* $(u_B, \tau) = (0.9, 5) \in \mathcal{B}$ *(right). Here* $\overline{u}(\tau) \approx 0.98$ *and* $s \approx 1.02$*, while* $\zeta_\ell \approx 0.08$ *(left) and* $s_B \approx 0.28$ *(right).*

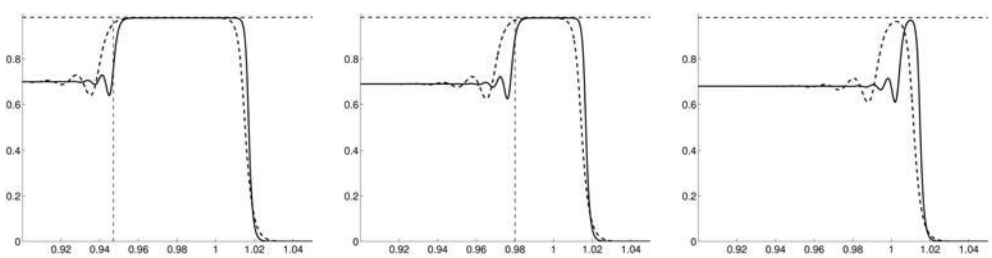


FIG. 11. *Graphs of* $w(\zeta, t)$ *with* $\tau = 5$ *at* $t = 1000$ *(dashed) and* $t = 2000$ *(solid); zoomed view:* $0.9 \leq \zeta \leq 1.05$*. Here* $\underline{u}(\tau) \approx 0.68$ *and* u_B *approaches* $\underline{u}(\tau)$ *from above through* 0.70 *(left),* 0.69 *(middle), and* 0.68 *(right). Then* s_B *increases from* 0.95 *(left) to* 0.98 *(middle) up to* 1.02 *(right). The other values are* $\overline{u}(\tau) \approx 0.98$ *and* $s \approx 1.02$*.*

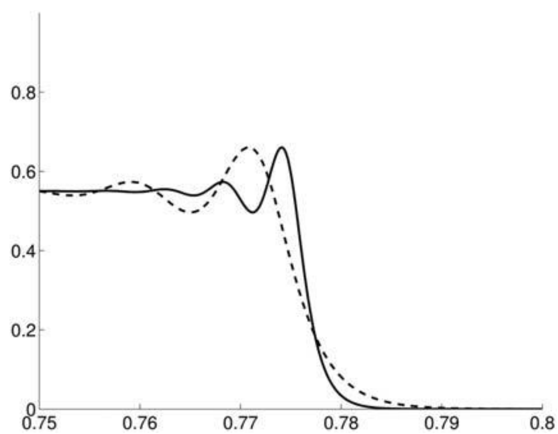


FIG. 12. *Graphs of* $w(\zeta, t)$ *at* $t = 1000$ *(dashed) and* $t = 2000$ *(solid) when* $(u_B, \tau) = (0.55, 5) \in \mathcal{C}_2$*; zoomed view:* $0.75 \leq \zeta \leq 0.8$*. Then* $s \approx 0.78$*.*

reduced to 0.55, so that we are now in $\mathcal{C}_2$. We find that the solution no longer jumps up to a higher plateau, but instead jumps right down after a small oscillation.

Note that the oscillations in Figures 11 and 12 contract around the shock as time progresses. This is due to the scaling, since we have plotted $w(\zeta, t)$ versus $\zeta = x/t$ for different values of time t.

We conclude from these simulations that the entropy solutions constructed in section 2 emerge as limiting solutions of the Cauchy problem (3.1). This suggests that these entropy solutions enjoy certain stability properties. It would be interesting to see whether these same entropy solutions would emerge if the initial value were chosen differently. We leave this question to a future study.

4. Proof of Theorem 1.1. In Theorem 1.1 we considered traveling wave solutions $u(\eta)$ of (1.17) in which the limiting conditions had been chosen so that $u(-\infty) = u_\ell \geq \alpha$ and $u(\infty) = u_r = 0$. Putting $u_r = 0$ in (1.18a) and (1.18b) we find that they are solutions of the problem

$$\text{(TW}_0\text{)}\begin{cases} s\tau u'' - u' - su + f(u) = 0 \qquad \text{for} \qquad -\infty < \eta < \infty, & \text{(4.1a)}\\ u(-\infty) = u_\ell, \qquad u(+\infty) = 0, & \text{(4.1b)}\end{cases}$$

in which the speed s is a priori determined by u_ℓ through

$$s = s(u_\ell) \stackrel{\text{def}}{=} \frac{f(u_\ell)}{u_\ell}. \tag{4.2}$$

The proof proceeds in a series of steps.

Step 1. We choose $u_\ell \in (\alpha, \beta)$ and prove that there exists a unique $\tau > 0$ for which problem (TW$_0$) has a solution, which is also unique. This defines a function $\tau = \tau(u_\ell)$ on (α, β). We then show that $\tau(u_\ell)$ is increasing, continuous, and that

$$\tau(u) \to \infty \qquad \text{if} \qquad u \to \beta.$$

Finally, we write

$$\tau_* \stackrel{\text{def}}{=} \lim_{u \to \alpha^+} \tau(u).$$

Step 2. We show that for any $\tau \in (0, \tau_*]$, problem (TW$_0$) has a solution with $u_\ell = \alpha$.

The proof is concluded by defining the function $\overline{u}_\ell(\tau)$ on (τ_*, ∞) as the inverse of the function $\tau(u_\ell)$ on the interval (α, β). The resulting function $\overline{u}(\tau)$, defined by (1.17) on $\mathbf{R}^+$, then has all the properties required in Theorem 1.1.

4.1. The function $\boldsymbol{\tau(u)}$. As a first result we prove that $\tau(u)$ is well defined on the interval (α, β).

LEMMA 4.1. *For each $u_\ell \in (\alpha, \beta)$ there exists a unique value of τ such that there exists a solution of problem* (TW$_0$). *This solution is unique and decreasing.*

Proof. It is convenient to write (4.1a) in a more conventional form, and introduce the variables

$$\xi = -\eta/\sqrt{s\tau} \qquad \text{and} \qquad \tilde{u}(\xi) = u(\eta).$$

In terms of these variables, problem (TW$_0$) becomes

$$\begin{cases} u'' + cu' - g(u) = 0 \qquad \text{in} \qquad -\infty < \xi < \infty, & \text{(4.3a)}\\ u(-\infty) = 0, \qquad u(+\infty) = u_\ell, & \text{(4.3b)}\end{cases}$$

where

$$c = \frac{1}{\sqrt{s\tau}} \qquad \text{and} \qquad g(u) = su - f(u), \tag{4.4}$$

and the tildes have been omitted. Graphs of $g(u)$ for $M = 2$ and different values of s are shown in Figure 13.

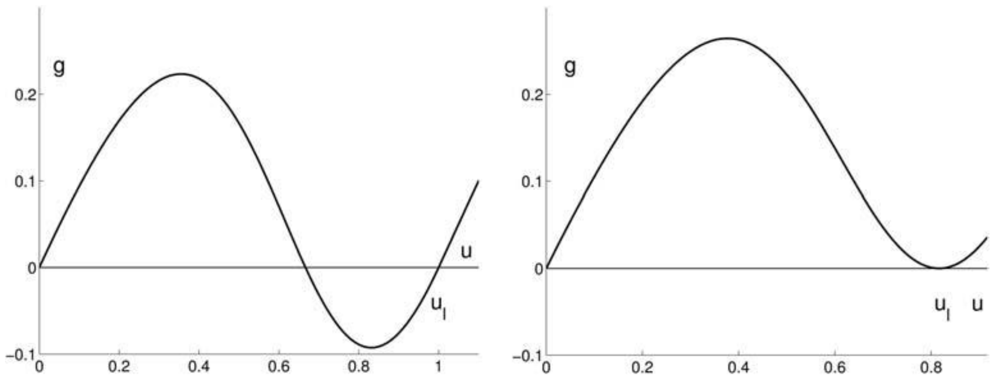


FIG. 13. *The function for $g(u)$ for $M = 2$, and $s = 0.95$ (left) and $s = s(\alpha) = 1.113$ (right).*

We study problem (4.3) in the phase plane and write (4.3a) as the first order system

$$\mathcal{P}(c,s) \begin{cases} u' = v, & (4.5a) \\ v' = -cv + g(u). & (4.5b) \end{cases}$$

For $u_\ell \in (\alpha, \beta)$ the function $g(u)$ has three distinct zeros, which we denote by u_i, $i = 0$, 1, and 2, where

$$u_0 = 0 \qquad \text{and} \qquad u_1 < \alpha < u_2 = u_\ell.$$

Plainly the points $(u, v) = (u_i, 0)$, $i = 0, 1, 2$, are the equilibrium points of (4.5) with associated eigenvalues

$$\lambda_\pm = -\frac{c}{2} \pm \frac{1}{2}\sqrt{c^2 + 4g'(u_i)}. \tag{4.6}$$

Since

$$g'(u_0) > 0, \qquad g'(u_1) < 0, \qquad \text{and} \qquad g'(u_2) > 0,$$

the outer points, $(u_0, 0)$ and $(u_2, 0)$, are saddles and $(u_1, 0)$ is either a stable node or a stable spiral.

Since we are interested in a traveling wave with $u(-\infty) = 0$ and $u(+\infty) = u_\ell$, we need to investigate orbits which connect the points $(0, 0)$ and $(u_\ell, 0)$. The existence of a unique wave speed c for which there exists such a solution of the system $\mathcal{P}(c, s)$, which is unique and decreasing, has been established in [25]; see also [17]. This allows us to define the function $c = c(u_\ell)$ for $\alpha < u_\ell < \beta$.

By definition, $c(u_\ell)$ only takes on positive values. This is consistent with the identity, obtained by multiplying (4.3a) by u' and integrating the result over $\mathbf{R}$:

$$c \int_{\mathbf{R}} \{u'(\xi)\}^2 \, d\xi = \int_{\mathbf{R}} g(u(\xi))\, u'(\xi) \, d\xi = \int_0^{u_\ell} g(t) \, dt \stackrel{\text{def}}{=} G(u_\ell), \tag{4.7}$$

because $G(u_\ell) > 0$ when $0 < u_\ell < \beta$.

Finally, by (4.2) and (4.4), we find that τ is uniquely determined by u_ℓ through the relation

$$\tau(u_\ell) = \frac{1}{s(u_\ell)c^2(u_\ell)}. \tag{4.8}$$

This completes the proof of Lemma 4.1. ∎

Lemma 4.1 allows us to define a function $\tau(u)$ on (α, β), such that if $u_\ell \in (\alpha, \beta)$, then problem ($\mathrm{TW}_0$) has a unique solution $u(\eta)$ if and only if $\tau = \tau(u_\ell)$. In the next lemma we show that the function $\tau(u)$ is strictly increasing on (α, β).

LEMMA 4.2. *Let $u_{\ell,i} = \gamma_i$ for $i = 1, 2$, where $\gamma_1 \in (\alpha, \beta)$, and let $\tau(\gamma_i) = \tau_i$. Then*

$$\gamma_1 < \gamma_2 \qquad \Longrightarrow \qquad \tau_1 < \tau_2.$$

Proof. For $i = 1, 2$ we write

$$s_i = \frac{f(\gamma_i)}{\gamma_i} \qquad \text{and} \qquad g_i(u) = s_i u - f(u).$$

Since

$$\frac{d}{du}\left(\frac{f(u)}{u}\right) = \frac{1}{u}\left(f'(u) - \frac{f(u)}{u}\right) < 0 \qquad \text{for} \qquad \alpha \le u < \beta,$$

it follows that

$$\gamma_1 < \gamma_2 \qquad \Longrightarrow \qquad s_1 > s_2 \quad \text{and} \quad g_1(u) > g_2(u) \quad \text{for} \quad u > 0. \tag{4.9}$$

To prove Lemma 4.2 we return to the formulation used in the proof of Lemma 4.1. Traveling waves correspond to heteroclinic orbits in the (u, v)-plane. Those associated with γ_1 and γ_2 we denote by Γ_1 and Γ_2. They connect the origin to $(\gamma_1, 0)$ and $(\gamma_2, 0)$, respectively.

We shall show that

$$\gamma_1 < \gamma_2 \qquad \Longrightarrow \qquad c_1 = c(\gamma_1) > c(\gamma_2) = c_2. \tag{4.10}$$

We can then conclude from (4.4) that

$$\tau_2 s_2 > \tau_1 s_1 \qquad \Longrightarrow \qquad \tau_2 > \frac{s_1}{s_2}\tau_1 > \tau_1,$$

as asserted.

Thus, suppose to the contrary that $c_1 \le c_2$. We claim that this implies that near the origin the orbit Γ_1 lies below Γ_2. Orbits of the system $\mathcal{P}(c, s)$ leave the origin along the unstable manifold under the angle θ given by

$$\theta = \theta(c, s) \stackrel{\text{def}}{=} \frac{1}{2}\left\{\sqrt{c^2 + 4s} - c\right\}. \tag{4.11}$$

An elementary computation shows that

$$\frac{\partial \theta}{\partial c} < 0 \qquad \text{and} \qquad \frac{\partial \theta}{\partial s} > 0. \tag{4.12}$$

Hence, since $s_1 > s_2$ and we assume that $c_1 \le c_2$, it follows that

$$\theta_1 = \theta(c_1, s_1) > \theta(c_2, s_2) = \theta_2,$$

and hence that the orbit Γ_1 starts out *above* Γ_2.

Since $(\gamma_2, 0)$ lies to the right of the point $(\gamma_1, 0)$ we conclude that Γ_1 and Γ_2 must intersect. Let us denote the first point of intersection by $P = (u_0, v_0)$. Then at P the slope of Γ_1 cannot exceed the slope of Γ_2. The slopes at P are given by

$$\frac{dv}{du}\Big|_{\Gamma_i} = -c_i + \frac{g_i(u_0)}{v_0}, \qquad i = 1, 2.$$

Because $g_1(u) > g_2(u)$ for $u > 0$ by (4.9), it follows that

$$\frac{dv}{du}\Big|_{\Gamma_1} > \frac{dv}{du}\Big|_{\Gamma_2} \qquad \text{at} \quad P,$$

so that, at P, the slope of Γ_1 exceeds the slope of Γ_2, a contradiction. Therefore we find that $c_1 > c_2$, as asserted. $\square$

In the next lemma we show that the function $\tau(u)$ is continuous.

LEMMA 4.3. *The function* $\tau : (\alpha, \beta) \to \mathbf{R}^+$ *is continuous.*

Proof. Because the function $s(\gamma) = \gamma^{-1} f(\gamma)$ is continuous, it suffices to show that the function $c(\gamma)$ is continuous. Since we have shown in the proof of Lemma 4.1 that $c(\gamma)$ is decreasing (cf. (4.10)), we only need to show that it cannot have any jumps.

Suppose to the contrary that it has a jump at γ_0, and let us write

$$\liminf_{\gamma \searrow \gamma_0^+} c(\gamma) = c^+ \qquad \text{and} \qquad \limsup_{\gamma \nearrow \gamma_0^-} c(\gamma) = c^-.$$

Then, since $c(\gamma)$ is decreasing, we may assume that $c^- > c^+$.

Thus, there exist sequences $\{\gamma_n^-\}$ and $\{\gamma_n^+\}$ with corresponding heteroclinic orbits $(u_n^\pm, v_n^\pm)$ and wave speeds $c_n^\pm$, such that

$$c_n^+ \searrow c^+ \qquad \text{and} \qquad c_n^- \nearrow c^- \qquad \text{as} \quad n \to \infty.$$

Since the unstable manifold at $(0, 0)$ and the stable manifold at $(\gamma, 0)$ depend continuously on c, it follows that the corresponding orbits also converge, i.e., that there exist orbits (u^+, v^+) and (u^-, v^-) such that

$$(u_n^\pm, v_n^\pm)(\xi) \to (u^\pm, v^\pm)(\xi) \qquad \text{as} \quad n \to \infty,$$

uniformly on $\mathbf{R}$. This argument yields two heteroclinic orbits, one with speed c^+ and one with speed c^-, which both connect the origin to the point $(\gamma_0, 0)$. Since by Lemma 4.1 there exists only one such orbit, we have a contradiction.

It follows that $c^- = c^+$, and continuity of the function $c(\gamma)$, and hence of $\tau(\gamma)$, has been established. $\square$

In the following lemma we prove the final assertion made in Step 1, which involves the behavior of $\tau(u)$ as $u \to \beta$.

LEMMA 4.4. *We have*

$$\tau(\gamma) \to \infty \qquad \textit{as} \qquad \gamma \to \beta^-.$$

Proof. In view of the definition (4.7) of τ, it suffices to show that $c(\gamma) \to 0$ as $\gamma \to \beta$. Proceeding as in the proof of Lemma 4.3, we find that $c(\gamma)$ and the orbit $\Gamma(c(\gamma))$ converge to c_0 and $\Gamma(c_0) = \{(u_0, v_0)(t) : t \in \mathbf{R}\}$ as $\gamma \to \beta$. Note that

$$c(\gamma) \int_{\mathbf{R}} v^2(\xi;\gamma)\, d\xi = \int_0^\gamma g(t;\gamma)\, dt,$$

where $g(t;\gamma) = s(\gamma)t - f(t)$. If we let $\gamma \to \beta$ in this identity, we obtain

$$c_0 \int_{\mathbf{R}} v_0^2(\xi)\, d\xi = \int_0^\beta g(t;\beta)\, dt = 0. \tag{4.13}$$

Because at the origin the unstable manifold points into the first quadrant when $\gamma = \beta$ (cf. (4.9)), it follows that $v_0 > 0$ on $\mathbf{R}$. Therefore, (4.13) implies that $c_0 = 0$, as asserted. ∎

4.2. Traveling waves with $u_\ell = \alpha$. In Lemmas 4.1 and 4.2 we have shown that $\tau(u)$ is an increasing function on (α, β). Since $\tau(u) > 0$ for all $u \in (\alpha, \beta)$, the limit

$$\tau_* \stackrel{\text{def}}{=} \lim_{u \to \alpha^+} \tau(u)$$

exists. In the following lemmas we show that $\tau_* > 0$ and that for all $\tau \in (0, \tau_*]$, problem (TW$_0$) has a unique solution with $u_\ell = \alpha$.

Let $\mathcal{S} \in \mathbf{R}^+$ denote the set of values of τ for which problem (TW$_0$) has a unique solution with $u_\ell = \alpha$.

LEMMA 4.5. *There exists a constant $\tau_0 > 0$ such that $(0, \tau_0) \subset \mathcal{S}$.*

Proof. We shall show that there exists a wave speed $c_0 > 0$ such that if $c > c_0$, then problem (4.5) has a heteroclinic orbit connecting the origin to the point $(\alpha, 0)$. This then yields Lemma 4.5 when we put

$$\tau_0 = \frac{1}{c_0^2 s(\alpha)}.$$

In (4.6) we saw that the origin is a saddle and that the slope of the unstable manifold is given by

$$\theta(c) = \frac{1}{2}\left\{\sqrt{c^2 + 4s} - c\right\}.$$

Note that

$$\theta(c) < \frac{1}{c} g'(0) = \frac{s}{c}.$$

Hence, near the origin the orbit lies below the isocline $\mathcal{I}_v = \{(u,v) : v = c^{-1} g(u), u \in \mathbf{R}\}$.

Since $u' > 0$ and $v' > 0$ in the lens shaped region

$$\mathcal{L} = \{(u,v) : 0 < u < \alpha,\ 0 < v < c^{-1} g(u),\ u \in \mathbf{R}\},$$

the orbit will leave $\mathcal{L}$ again. To see what happens next, we consider the triangular region Ω_m bounded by the positive u- and v-axis and the line

$$\ell_m \stackrel{\text{def}}{=} \{(u,v) : v = m(\alpha - u)\}, \qquad m > 0.$$

On the axes the vector field points into Ω_m, and on the line ℓ_m it points inwards if

$$\left.\frac{dv}{du}\right|_{\ell_m} = -c + \frac{g(u)}{m(\alpha - u)} < -m. \tag{4.14}$$

Let

$$m_0 = \inf\{m > 0 : g(u) < m(\alpha - u) \text{ on } (0, \alpha)\}.$$

Then

$$-c + \frac{g(u)}{m(\alpha - u)} < -c + \frac{m_0}{m},$$

and (4.14) will hold for values of c and m which satisfy the inequality

$$-c + \frac{m_0}{m} < -m$$

or

$$c > m + \frac{m_0}{m}.$$

To obtain the largest range of values of c for which the vector field points into Ω_m we choose m so that the right-hand side of this inequality becomes smallest; i.e., we put $m = \sqrt{m_0}$. We thus find that for

$$c > c_0 \stackrel{\text{def}}{=} 2\sqrt{m_0}$$

the region $\Omega_{\sqrt{m_0}}$ is invariant, and hence, that the orbit must tend to the point $(\alpha, 0)$. This completes the proof of Lemma 4.5. □

The next lemma gives the structure of the set $\mathcal{S}$.

LEMMA 4.6. *If* $\tau_0 \in \mathcal{S}$, *then* $(0, \tau_0] \subset \mathcal{S}$.

Proof. As in earlier lemmas we prove a related result for problem (4.5). Let $\mathcal{S}^*$ be the set of values of c for which there exists a heteroclinic orbit of problem (4.5) from $(0,0)$ to $(\alpha, 0)$. We show that if $c_0 \in \mathcal{S}^*$, then $[c_0, \infty) \subset \mathcal{S}^*$. Plainly this implies Lemma 4.6 with $\tau_0 = 1/(c_0 s^2)$.

As before, we denote the orbit emanating from the origin by $\Gamma(c)$. Suppose that $c > c_0$. Then, since $\theta'(c) < 0$ it follows that $\theta(c_0) > \theta(c)$, so that near the origin $\Gamma(c_0)$ lies above $\Gamma(c)$. We claim that $\Gamma(c_0)$ and $\Gamma(c)$ will not intersect for $u \in (0, \alpha)$. Accepting this claim for the moment, we conclude that since $\Gamma(c_0)$ tends to $(\alpha, 0)$, the orbit $\Gamma(c)$ must converge to $(\alpha, 0)$ as well.

It remains to prove the claim. Suppose that $\Gamma(c_0)$ and $\Gamma(c)$ do intersect at some $u \in (0, \alpha)$, and let (u_0, v_0) be the first point of intersection. Then

$$\left.\frac{dv}{du}\right|_{\Gamma(c)} \geq \left.\frac{dv}{du}\right|_{\Gamma(c_0)} \quad \text{at} \quad (u_0, v_0). \tag{4.15}$$

But, from the differential equations we deduce that

$$\left.\frac{dv}{du}\right|_{\Gamma(c)} = -c + \frac{g(u_0)}{v_0} < -c_0 + \frac{g(u_0)}{v_0} = \left.\frac{dv}{du}\right|_{\Gamma(c_0)} \quad \text{at} \quad (u_0, v_0),$$

which contradicts (4.15). This proves the claim and so completes the proof of Lemma 4.6. □

We conclude this section by showing that $\tau_* \in \mathcal{S}$, and hence that $\mathcal{S} = (0, \tau_*]$.

LEMMA 4.7. *We have* $\mathcal{S} = (0, \tau_*]$.

Proof. It follows from Lemmas 4.1 and 4.2 that for every $\varepsilon \in (0, \beta - \alpha)$, there exists a $\tau_\varepsilon = \tau(\alpha + \varepsilon) > 0$ such that problem (TW_0) has a unique traveling wave $u_\varepsilon(\eta)$ with speed $s_\varepsilon = s(\alpha + \varepsilon)$, such that

$$u_\varepsilon(-\infty) = \alpha + \varepsilon \qquad \text{and} \qquad u_\varepsilon(\infty) = 0.$$

This wave corresponds to a heteroclinic orbit $\Gamma_\varepsilon = \{(u_\varepsilon(\xi), v_\varepsilon(\xi)) : \xi \in \mathbf{R}\}$ of the system $\mathcal{P}(c_\varepsilon, s_\varepsilon)$, where $c_\varepsilon = 1/\sqrt{s_\varepsilon \tau_\varepsilon}$, which connects the points $(0,0)$ and $(\alpha + \varepsilon, 0)$. It leaves the origin along the stable manifold under an angle $\theta_\varepsilon = \theta(c_\varepsilon, s_\varepsilon)$ and enters the point $(\alpha + \varepsilon, 0)$ along the stable manifold under the angle

$$\psi_\varepsilon = \psi(c_\varepsilon, s_\varepsilon) = \frac{1}{2}\left\{-c_\varepsilon - \sqrt{c_\varepsilon^2 + 4g'(\alpha + \varepsilon)}\right\} \to -c_0 = -\frac{1}{\sqrt{s(\alpha)\tau_*}} \qquad \text{as} \quad \varepsilon \to 0.$$

Reversing time, i.e., replacing ξ by $-\xi$, we can view Γ_ε as the unique orbit emanating from the point $(\alpha + \varepsilon, 0)$ into the first quadrant and entering the origin as $\xi \to \infty$. In the limit, as $\varepsilon \to 0$, we find that

$$u_\varepsilon(\xi) \to u_0(\xi) \quad \text{and} \quad v_\varepsilon(\xi) \to v_0(\xi) \quad \text{as} \quad \varepsilon \to 0 \qquad \text{for} \quad -\infty < \xi \le \xi_0,$$

where ξ_0 is any finite number. We claim that

$$u_0(\xi) \to 0 \quad \text{and} \quad v_0(\xi) \to 0 \qquad \text{as} \quad \xi \to \infty;$$

i.e., $\Gamma_0 \stackrel{\text{def}}{=} \{(u_0(\xi), v_0(\xi)) : \xi \in \mathbf{R}\}$ is a heteroclinic orbit, which connects $(\alpha, 0)$ and the origin $(0,0)$.

Suppose to the contrary that Γ_0 does not enter the origin as $\xi \to \infty$ and possibly does not even exist for all $\xi \in \mathbf{R}$. Then, since

$$\frac{dv}{du} = -c_0 + \frac{g(u)}{v} > -c_0 \qquad \text{if} \qquad 0 < u < \alpha, \quad v > 0,$$

Γ_0 must leave the first quadrant in finite time, either through the u-axis or through the v-axis. This means by continuity that for ε small enough Γ_ε must also leave the first quadrant in finite time. Since Γ_ε is known to enter the origin for every $\varepsilon > 0$, and hence never to leave the first quadrant, we have a contradiction. This proves the claim that Γ_0 is a heteroclinic orbit, which connects $(\alpha, 0)$ and $(0,0)$. ∎

Remark 4.1. It is evident from Lemmas 4.6 and 4.7 that

$$\tau_* \ge \tau_0 = \frac{1}{4m_0 s(\alpha)}, \tag{4.16}$$

where m_0 was defined in (4.15). For $M = 2$, we find that $s(\alpha) \approx 1.11$, $m_0 \approx 0.70$ and hence $\tau_0 \approx 0.32$. Numerically, we find that $\tau_* \approx 0.61$.

5. Proof of Theorems 1.2 and 1.3. For the proofs of Theorems 1.2 and 1.3 we turn to the system $\mathcal{P}(c, s)$ defined in section 4. For convenience we restate it here,

$$\mathcal{P}(c,s)\begin{cases} u' = v, & \text{(5.1a)} \\ v' = -cv + g_s(u), & \text{(5.1b)} \end{cases}$$

where

$$c = \frac{1}{\sqrt{s\tau}} \qquad \text{and} \qquad g_s(u) = su - f(u).$$

Part (a) of Theorem 1.2 is readily seen to be a consequence of the following lemma.

LEMMA 5.1. *Let $\tau > \tau_*$ be given. Then for every $u_\ell \in (0, \underline{u})$, there exists a unique heteroclinic orbit of the system $\mathcal{P}(c, s)$ in which*

$$s = s_\ell = \frac{f(u_\ell)}{u_\ell} \qquad \textit{and} \qquad c = c_\ell = \frac{1}{\sqrt{s_\ell \tau}},$$

which connects $(0,0)$ and $(u_\ell, 0)$.

Proof. Let Γ_ℓ and $\overline{\Gamma}$ denote the orbits of $\mathcal{P}(c_\ell, s_\ell)$ and $\mathcal{P}(\overline{c}, \overline{s})$, where $\overline{c} = c(\overline{u})$ and $\overline{s} = s(\overline{u})$, which enter the first quadrant from the origin. They do this under the angles $\theta(c_\ell, s_\ell)$ and $\theta(\overline{c}, \overline{s})$, respectively. Since $c_\ell > \overline{c}$ and $s_\ell < \overline{s}$, it follows from (4.12) that

$$\theta(c_\ell, s_\ell) < \theta(\overline{c}, \overline{s}).$$

Hence, near the origin, Γ_ℓ lies below $\overline{\Gamma}$. Thus, Γ_ℓ enters the region Ω enclosed between $\overline{\Gamma}$ and the u-axis. Since

$$\frac{dv}{du}\bigg|_{\Gamma_\ell} = -c_\ell + \frac{s_\ell u - f(u)}{v} < -\overline{c} + \frac{\overline{s}u - f(u)}{v} = \frac{dv}{du}\bigg|_{\overline{\Gamma}},$$

it follows that Γ_ℓ cannot leave Ω though its "top" $\overline{\Gamma}$. We define the following subsets of the bottom of Ω:

$$\begin{aligned} S_1 &= \{(u, v) : 0 < u < u_\ell,\ v = 0\}, \\ S_2 &= \{(u, v) : u = u_\ell,\ v = 0\}, \\ S_3 &= \{(u, v) : u_\ell < u < \overline{u},\ v = 0\}. \end{aligned}$$

Inspection of the vector field show that orbits can only leave Ω through S_3. Note that the set S_2 consists of an equilibrium point.

There are two possibilities: *either* Γ_ℓ never leaves Ω, *or* Γ_ℓ leaves Ω, necessarily through the set S_3. In the first case Γ_ℓ is a heteroclinic orbit from $(0,0)$ to $(u_\ell, 0)$, and the proof is complete.

Thus, let us assume that Γ_ℓ leaves Ω at some point $(u, v) = (u_0, 0)$. Consider the energy function

$$\mathcal{H}(u, v) = \frac{1}{2}v^2 - G_{s_\ell}(u),$$

where G_{s_ℓ} is the primitive of g_{s_ℓ} as defined in (4.7), and write $H(\xi) = \mathcal{H}(u(\xi), v(\xi))$, when $(u(\xi), v(\xi))$ is an orbit. Then differentiation shows that

$$H'(\xi) = -c_\ell v^2(\xi) < 0.$$

Since $\mathcal{H}(0,0) = 0$, it follows that

$$\mathcal{H}(u_0, 0) = -G_{s_\ell}(u_0) < 0$$

and that

$$\mathcal{H}(u(\xi), v(\xi)) = \frac{1}{2}v^2 - G_{s_\ell}(u) < -G_{s_\ell}(u_0) \qquad \text{for} \qquad \xi > \xi_0.$$

This means that

$$G_{s_\ell}(u) > G_{s_\ell}(u_0) > 0 \qquad \text{for} \qquad \xi > \xi_0.$$

Let

$$u_1 = \inf\{s \in \mathbf{R} : G_{s_\ell}(s) > G_{s_\ell}(u_0) \text{ on } (s, u_0)\}.$$

Since $G_{s_\ell}(u_0) > 0$ it follows that $u_1 \in (0, u_\ell)$. Therefore

$$\left.\begin{array}{l} 0 < u_1 < u(x) < u_0 \\ v^2(x) < 2\{G_{s_\ell}(u_\ell) - G_{s_\ell}(u_0)\} \end{array}\right\} \qquad \text{for} \qquad x > x_0.$$

From a simple energy argument we conclude that $(u(x), v(x)) \to (u_\ell, 0)$ as $x \to \infty$. This completes the proof of Lemma 5.1. ☐

Part (b) follows from the following result.

LEMMA 5.2. *Let $\tau > \tau_*$ be given. For any $u_\ell \in (\underline{u}(\tau), \overline{u}(\tau))$ there exists no solution of the system $\mathcal{P}(c_\ell, s_\ell)$, with*

$$s_\ell = \frac{f(u_\ell)}{u_\ell} \qquad \textit{and} \qquad c_\ell = \frac{1}{\sqrt{s_\ell \tau}},$$

which connects $(0,0)$ and $(u_\ell, 0)$.

Proof. Let $\overline{\Gamma}$ denote the orbit corresponding to $\overline{c}$ and $\overline{s}$, which connects $(0,0)$ and the point $(\overline{u}, 0)$, and let Γ_ℓ denote the orbit which corresponds to c_ℓ and s_ℓ. Observe that

$$s_\ell > \overline{s} \qquad \text{and} \qquad c_\ell < \overline{c},$$

and hence

$$\theta(c_\ell, s_\ell) > \theta(\overline{c}, \overline{s}).$$

Therefore, near the origin, Γ_ℓ lies above $\overline{\Gamma}$. Hence, to reach the point $(u_\ell, 0)$, the orbit $\overline{\Gamma}_\ell$ has to cross $\overline{\Gamma}$ somewhere, and at the first point of crossing we must have

$$\left.\frac{dv}{du}\right|_{\overline{\Gamma}} \geq \left.\frac{dv}{du}\right|_{\Gamma_\ell}.$$

However, by the equations, we have

$$\left.\frac{dv}{du}\right|_{\overline{\Gamma}} = -\overline{c} + \frac{g_{\overline{s}}(u)}{u} < -c_\ell + \frac{g_{s_\ell}(u)}{u} = \left.\frac{dv}{du}\right|_{\Gamma_\ell},$$

so that we have a contradiction. ☐

This completes the proof of Theorem 1.2.

The proof of Theorem 1.3 is entirely analogous to that of Theorem 1.2, and we omit it.

6. Entropy dissipation. In this section we study the Cauchy problem

$$\text{(CP)}\begin{cases} u_t + (f(u))_x = \mathcal{A}_\varepsilon(u) & \text{in} \quad S = \mathbf{R}\times\mathbf{R}^+, & \text{(6.1a)} \\ u(\cdot,0) = u_0(\cdot) & \text{on} \quad \mathbf{R}, & \text{(6.1b)} \end{cases}$$

where

$$\mathcal{A}_\varepsilon(u) = \varepsilon u_{xx} + \varepsilon^2 \tau u_{xxt} \qquad (\varepsilon > 0). \tag{6.2}$$

With this choice (6.1a) becomes the regularized BL equation (1.17) for which we obtained traveling wave solutions in the previous sections. In (6.1a) and (6.2) we introduce subscripts to denote partial derivatives. Without further justification we assume that problem (CP) has a smooth, nonnegative, and bounded solution u^ε for each $\varepsilon > 0$, and that there exists a limit function $u : S \to [0,\infty)$ such that for each $(x,t) \in S$,

$$u^\varepsilon(x,t) \to u(x,t) \qquad \text{as} \qquad \varepsilon \to 0.$$

In addition we assume the following structural properties:

(i) $\|u^\varepsilon\|_\infty < C$ for some constant $C > 0$, and for each fixed $t > 0$,

$$u^\varepsilon(x,t) \to u_\ell \in \mathbf{R}^+ \qquad \text{as} \qquad x \to -\infty,$$
$$u^\varepsilon(x,t) \to u_r \in \mathbf{R}^+ \qquad \text{as} \qquad x \to +\infty.$$

(ii) The partial derivatives of u^ε vanish as $|x| \to \infty$.

(iii) Let $U(s) = \frac{1}{2}s^2$ for $s \geq 0$, $U_\ell = U(u_\ell)$, and $U_r = U(u_r)$. Then there exists a smooth function $\lambda_\varepsilon : [0,\infty) \to \mathbf{R}$ which is uniformly bounded with respect to $\varepsilon > 0$ in any bounded interval $(0,T)$, such that

$$\int_{\mathbf{R}} \{U(u^\varepsilon(x,t)) - G_\varepsilon(x,t)\}\, dx = 0 \qquad \text{for all} \quad t > 0,$$

where G_ε is the step function

$$G_\varepsilon(x,t) = U_\ell + (U_r - U_\ell)H(x - \lambda_\varepsilon(t)), \qquad (x,t) \in S$$

in which H denotes the Heaviside function.

Note that the traveling waves constructed in this paper all have these properties.

Remark 6.1. The question as to which conditions on u_0 would generate such a solution is left open in this paper. Clearly we need that $u_0 : \mathbf{R} \to \mathbf{R}$ satisfies (i) and (ii), and $U(u_0) - G \in L^1(\mathbf{R})$. Further we require that $u_0' \in L^2(\mathbf{R})$.

The main purpose of this section is to show that $U(u^\varepsilon)$ is an entropy for (6.1a). In doing so we borrow arguments and ideas of LeFloch [26]. For completeness we recall some definitions. We say that the term $\mathcal{A}_\varepsilon(u)$ is *conservative* if

$$\lim_{\varepsilon\to 0^+} \int_S \mathcal{A}_\varepsilon(u^\varepsilon)\varphi = 0 \qquad \text{for all} \quad \varphi \in C_0^\infty(S) \tag{6.3}$$

and we say that $\mathcal{A}_\varepsilon(u)$ is *entropy dissipative* (for an entropy U) if

$$\limsup_{\varepsilon\to 0^+} \int_S \mathcal{A}_\varepsilon(u^\varepsilon)U'(u^\varepsilon)\varphi \leq 0 \qquad \text{for all} \quad \varphi \in C_0^\infty(S), \quad \varphi \geq 0. \tag{6.4}$$

We establish the following theorem.

THEOREM 6.1. *Let u^ε be the solution of problem* (CP)*, and let u^ε satisfy* (i), (ii), *and* (iii). *Then, the regularization $\mathcal{A}_\varepsilon(u)$ defined in* (6.2) *has the following properties:*

(a) $\mathcal{A}_\varepsilon(u)$ *is conservative.*

(b) $\mathcal{A}_\varepsilon(u)$ *is entropy dissipative for the entropy* $U(u) = \frac{1}{2}u^2$.

Proof. Part (a). For any $\varphi \in C_0^\infty(S)$ we obtain after partial integration with respect to x and t,

$$\int_S \mathcal{A}_\varepsilon(u^\varepsilon)\varphi = \varepsilon \int_S u^\varepsilon \varphi_{xx} - \varepsilon^2\tau \int_S u^\varepsilon \varphi_{xxt} \to 0 \qquad \text{as} \quad \varepsilon \to 0.$$

Part (b). To simplify notation, we drop the superscript ε from u^ε. When we multiply (6.1a) by u we obtain

$$\partial_t U(u) + \partial_x F(u) = u\mathcal{A}_\varepsilon(u) = \varepsilon u u_{xx} + \varepsilon^2 \tau u u_{xxt}, \tag{6.5}$$

where

$$F(u) = \int_0^u U'(s) f'(s)\, ds = \int_0^u s f'(s)\, ds = u f(u) - \int_0^u f(s)\, ds. \tag{6.6}$$

An elementary computation shows that

$$\begin{aligned} \varepsilon u u_{xx} &= \varepsilon U_{xx} - \varepsilon u_x^2, \\ \varepsilon^2 \tau u u_{xxt} &= \varepsilon^2 \tau \left(U_{xxt} - \frac{1}{2}(u_x^2)_t - (u_x u_t)_x \right). \end{aligned}$$

Hence

$$\begin{aligned} \int_S \mathcal{A}_\varepsilon(u)\, u\varphi = \varepsilon \int_S U\varphi_{xx} - \varepsilon \int_S u_x^2 \varphi \\ - \varepsilon^2\tau \int_S U\varphi_{xxt} + \frac{1}{2}\varepsilon^2\tau \int_S u_x^2 \varphi_t + \varepsilon^2\tau \int_S u_t u_x \varphi_x. \end{aligned} \tag{6.7}$$

Plainly

$$\varepsilon \int_S U\varphi_{xx} \to 0 \quad \text{and} \quad \varepsilon^2\tau \int_S U\varphi_{xxt} \to 0 \qquad \text{as} \quad \varepsilon \to 0.$$

Since $\varphi \geq 0$, it remains to estimate the last two terms on the right-hand side of (6.7).

For this purpose we establish the following two estimates.

LEMMA 6.1. *Let $T > 0$, and let $S_T = \mathbf{R} \times (0, T]$. Then there exists a constant $C > 0$ such that for all $\varepsilon > 0$,*

$$\varepsilon \int_{S_T} u_x^2 \leq C \tag{6.8}$$

and

$$\varepsilon \int_{S_T} u_t^2 \leq C. \tag{6.9}$$

Proof of (6.8). We write (6.5) as

$$\partial_t U(u) + \partial_x F(u) = \varepsilon U_{xx} - \varepsilon u_x^2 + \varepsilon^2 \tau \left\{ U_{xxt} - \frac{1}{2}(u_x^2)_t - (u_t u_x)_x \right\}. \tag{6.10}$$

Using properties (i)–(iii) and writing $F_\ell = F(u_\ell)$, $F_r = F(u_r)$, we find that

$$\frac{d}{dt}\int_{\mathbf{R}} \{U(x,t) - G_\varepsilon(x,t)\}\,dx - \frac{d\lambda_\varepsilon}{dt}(U_r - U_\ell) \\ + (F_r - F_\ell) + \varepsilon\int_{\mathbf{R}} u_x^2 + \frac{1}{2}\varepsilon^2\tau\frac{d}{dt}\int_{\mathbf{R}} u_x^2 \le 0,$$

or, when we integrate over $(0,T)$

$$-\{\lambda_\varepsilon(t) - \lambda_\varepsilon(0)\}(U_r - U_\ell) + (F_r - F_\ell)t + \varepsilon\int_{S_T} u_x^2 + \frac{1}{2}\varepsilon^2\tau\int_{\mathbf{R}} u_x^2(t) \le \frac{1}{2}\varepsilon^2\tau\int_{\mathbf{R}}(u_0')^2,$$

from which (6.8) immediately follows. $\square$

Proof of (6.9). We multiply (6.1) by u_t. This yields

$$u_t^2 + (f(u))_x u_t = u_t\mathcal{A}_\varepsilon(u) = \varepsilon u_t u_{xx} + \varepsilon^2\tau u_t u_{xxt}. \tag{6.11}$$

Using the identities

$$u_t u_{xx} = (u_x u_t)_x - \frac{1}{2}(u_x^2)_t \qquad \text{and} \qquad u_t u_{xxt} = (u_{xt}u_t)_x - (u_{xt})^2,$$

we find that

$$u_t^2 + \frac{\varepsilon}{2}(u_x^2)_t \le -f'(u)u_t u_x + \varepsilon(u_x u_t)_x + \varepsilon^2\tau(u_{xt}u_t)_x.$$

When we integrate over $\mathbf{R}$ and use Schwarz's inequality and properties (i) and (ii), we obtain

$$\int_{\mathbf{R}} u_t^2 + \frac{\varepsilon}{2}\frac{d}{dt}\int_{\mathbf{R}} u_x^2 \le \frac{1}{2}\int_{\mathbf{R}} u_t^2 + \frac{K^2}{2}\int_{\mathbf{R}} u_x^2,$$

where $K = \max\{|f'(s)| : s > 0\}$. Hence, when we integrate over $(0,t)$,

$$\int_{S_t} u_t^2 \le \varepsilon\int_{\mathbf{R}}(u_0')^2 + K^2\int_{S_t} u_x^2.$$

In view of the first estimate this establishes (6.9) and completes the proof of Lemma 6.1. $\square$

We now return to the proof of Theorem 6.1(b). For each $\varphi \in C_0^\infty(S)$ we choose $T > 0$ so that $\operatorname{supp}\varphi \subset S_T$. Then (6.8) implies that

$$\varepsilon^2\int_{S_T} u_x^2\varphi_t \le \varepsilon^2 K_1\int_{S_T} u_x^2 \le \varepsilon K_1 C \quad \text{with} \quad K_1 = \sup|\varphi_t|, \tag{6.12}$$

and (6.8) and (6.9) together imply that

$$\varepsilon^2\int_{S_T} u_t u_x\varphi_x \le \varepsilon^2 K_2\int_{S_T}|u_t||u_x| \le \varepsilon K_2 C \quad \text{with} \quad K_2 = \sup|\varphi_x|. \tag{6.13}$$

Using (6.12) and (6.13) in (6.7) we conclude that, writing $u = u^\varepsilon$ again,

$$\limsup_{\varepsilon\searrow 0}\int_S \mathcal{A}_\varepsilon(u^\varepsilon)\,u^\varepsilon\varphi \le 0,$$

which is what was claimed in Theorem 6.1. □

It now follows from (6.5) that in the limit as $\varepsilon \to 0$,

$$\partial_t U(u) + \partial_x F(u) \leq 0 \tag{6.14}$$

holds in a weak or distributional sense. This shows that (U, F) is an entropy pair for (1.1).

The inequality in (6.14) indicates entropy dissipation. Across shocks $\{u_\ell, u_r\}$ it can be computed explicitly. Let

$$u(x,t) = \begin{cases} u_\ell & \text{for} \quad x < st, \\ u_r & \text{for} \quad x > st. \end{cases}$$

Then (6.14) implies that

$$-s(U_r - U_\ell) + (F_r - F_\ell) \leq 0.$$

Hence the entropy dissipation is given by

$$E(u_\ell, u_r) \stackrel{\text{def}}{=} -s(U_r - U_\ell) + (F_r - F_\ell). \tag{6.15}$$

We conclude by observing that if $u = u(\eta)$ is a traveling wave satisfying problem (TW), then (6.15) can be written as

$$E(u_\ell, u_r) = \int_{\mathbf{R}} \{-s(U(u))' + (F(u))'\}\, d\eta. \tag{6.16}$$

Applying (6.6) and the definition of U gives

$$E(u_\ell, u_r) = \int_{\mathbf{R}} u \left(-s + \frac{df}{du}\right) u'\, d\eta = \int_{u_\ell}^{u_r} u \left(-s + \frac{df}{du}\right) du. \tag{6.17}$$

Rewriting further

$$-s + \frac{df}{du} = \frac{d}{du}\big(-s(u - u_\ell) + f(u) - f(u_\ell)\big),$$

integrating (6.17) by parts, and using the Rankine–Hugoniot condition yields

$$E(u_\ell, u_r) = \int_{u_r}^{u_\ell} \{f(u) - f(u_\ell) - s(u - u_\ell)\}\, du.$$

In the special case when $u_r = 0$ we have $s = f(u_\ell)/u_\ell$ and thus

$$E(u_\ell, 0) = \int_0^{u_\ell} \{f(u) - su\}\, du.$$

Returning to the proof of Proposition 1.1 we observe that the integral is negative provided $u_\ell < \beta$. Thus this condition acts as an entropy condition in the sense that $E(u_\ell, 0) < 0$ only if $u_\ell < \beta$.

REFERENCES

[1] D. N. Arnold, J. J. Douglas, and V. Thomée, *Superconvergence of a finite element approximation to the solution of a Sobolev equation in a single space variable*, Math. Comp., 36 (1981), pp. 53–63.

[2] D. G. Aronson, *The porous media equation*, in Nonlinear Diffusion Problems (Montecatini Terme, 1985), Lecture Notes in Math. 1224, Springer, Berlin, 1986, pp. 1–46.

[3] G. I. Barenblatt, J. Garcia-Azorero, A. De Pablo, and J. L. Vazquez, *Mathematical model of the non-equilibrium water-oil displacement in porous strata*, Appl. Anal., 65 (1997), pp. 19–45.

[4] G. I. Barenblatt, V. M. Entov, and V. M. Ryzhik, *Theory of Fluid Flow through Natural Rocks*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[5] J. Bear, *Dynamics of Fluids in Porous Media*, American Elsevier, New York, London, Amsterdam, 1972.

[6] N. Bedjaoui and P. G. LeFloch, *Diffusive-dispersive travelling waves and kinetic relations. V. Singular diffusion and nonlinear dispersion*, Proc. Roy. Soc. Edinburgh Sect. A, 134 (2004), pp. 815–843.

[7] A. L. Bertozzi, A. Münch, and M. Shearer, *Undercompressive shocks in thin film flows*, Phys. D, 134 (1999), pp. 431–464.

[8] A. L. Bertozzi and M. Shearer, *Existence of undercompressive traveling waves in thin film equations*, SIAM J. Math. Anal., 32 (2000), pp. 194–213.

[9] C. Cuesta, *Pseudo-Parabolic Equations with Driving Convection Term*, Ph.D. thesis, VU Amsterdam, The Netherlands, 2003.

[10] C. Cuesta and I. S. Pop, *Numerical Schemes for a Pseudo-Parabolic Burgers Equation: Discontinuous Data and Long-Time Behaviour*, CASA Report 07-22, Eindhoven University of Technology, Eindhoven, The Netherlands, 2007.

[11] C. Cuesta, C. J. van Duijn, and J. Hulshof, *Infiltration in porous media with dynamic capillary pressure: Travelling waves*, European J. Appl. Math., 11 (2000), pp. 381–397.

[12] C. Cuesta and J. Hulshof, *A model problem for groundwater flow with dynamic capillary pressure: Stability of travelling waves*, Nonlinear Anal., 52 (2003), pp. 1199–1218.

[13] D. A. DiCarlo, *Experimental measurements of saturation overshoot on infiltration*, Water Resour. Res., 40 (2004), pp. W04215.1–W04215.9.

[14] A. G. Egorov, R. Z. Dautov, J. L. Nieber, and A. Y. Sheshukov, *Stability analysis of traveling wave solution for gravity-driven flow*, in Developments in Water Science, Elsevier, Amsterdam, The Netherlands, 2003, pp. 121–128.

[15] A. G. Egorov, R. Z. Dautov, J. L. Nieber, and A. Y. Sheshukov, *Stability analysis of gravity-driven infiltrating flow*, Water Resour. Res., 39 (2003), p. 1266.

[16] R. E. Ewing, *Time-stepping Galerkin methods for nonlinear Sobolev partial differential equations*, SIAM J. Numer. Anal., 15 (1978), pp. 1125–1150.

[17] P. C. Fife and J. B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Ration. Mech. Anal., 65 (1977), pp. 335–361.

[18] J. B. Greer and A. L. Bertozzi, *Traveling wave solutions of fourth order PDEs for image processing*, SIAM J. Math. Anal., 36 (2004), pp. 38–68.

[19] S. M. Hassanizadeh and W. G. Gray, *Mechanics and thermodynamics of multiphase flow in porous media including interphase boundaries*, Adv. Water Resour., 13 (1990), pp. 169–186.

[20] S. M. Hassanizadeh and W. G. Gray, *Thermodynamic basis of capillary pressure in porous media*, Water Resour. Res., 29 (1993), pp. 3389–3405.

[21] B. T. Hayes and P. G. LeFloch, *Non-classical shocks and kinetic relations: Scalar conservation laws*, Arch. Ration. Mech. Anal., 139 (1997), pp. 1–56.

[22] B. T. Hayes and P. G. LeFloch, *Nonclassical shocks and kinetic relations: Finite difference schemes*, SIAM J. Numer. Anal., 35 (1998), pp. 2169–2194.

[23] B. T. Hayes and M. Shearer, *Undercompressive shocks and Riemann problems for scalar conservation laws with non-convex fluxes*, Proc. Roy. Soc. Edinburgh Sect. A, 129 (1999), pp. 733–754.

[24] D. Jacobs, W. R. McKinney, and M. Shearer, *Travelling wave solutions of the modified Korteweg-de Vries-Burgers equation*, J. Differential Equations, 116 (1995), pp. 448-467.

[25] Ya. I. Kaniel, *On the stabilisation of solutions of the Cauchy problem for equations arising in the theory of combustion*, Mat. Sb. (N.S.), 59 (1962), pp. 245–288. See also Dokl. Akad. Nauk SSSR, 132 (1960), pp. 268–271 (= Soviet Math. Dokl., 1 (1960), pp. 533–536) and Dokl. Akad. Nauk SSSR, 136 (1961), pp. 277–280 (= Soviet Math. Dokl., 2 (1961), pp. 48–51).

[26] P. G. LeFloch, *Hyperbolic Systems of Conservation Laws. The Theory of Classical and Non-*

classical Shock Waves, Lectures Math. ETH Zürich, Birkhäuser Verlag, Basel, 2002.

[27] L. A. Peletier, *The porous media equation*, in Application of Nonlinear Analysis in the Physical Sciences, H. Amman, N. Bazley, and K. Kirchgaessner, eds., Pitman, London, 1981, pp. 229–241.

[28] D. E. Smiles, G. Vachaud, and M. Vauclin, *A test of the uniqueness of the soil moisture characteristic during transient, non-hysteretic flow of water in rigid soil*, Soil Sci. Soc. Am. Proc., 35 (1971), pp. 534–539.

[29] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[30] F. Stauffer, *Time dependence of the relations between capillary pressure, water content and conductivity during drainage of porous media*, in Proceedings of the LAHR Symposium on Scale Effects in Porous Media, Thessaloniki, Greece, 1978, pp. 156–177.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 537–553

CONVERGENCE OF INCREASINGLY FLAT RADIAL BASIS INTERPOLANTS TO POLYNOMIAL INTERPOLANTS*

YEON JU LEE†, GANG JOON YOON‡, AND JUNGHO YOON†

Abstract. In this paper, we study the convergence behavior of interpolants by smooth radial basis functions to polynomial interpolants in $\mathbb{R}^d$ as the radial basis functions are scaled to be increasingly flat. Larsson and Fornberg [*Comput. Math. Appl.*, 49 (2005), pp. 103–130] conjectured a sufficient property for this convergence, and they also conjectured that Bessel radial functions do not satisfy this property. First, in the case of positive definite radial functions, we prove both conjectures by Larsson and Fornberg for the convergence of increasingly flat radial function interpolants. Next, we extend the results to the case of conditionally positive definite radial functions of order $m > 0$.

Key words. radial basis function, interpolation, polynomial, conditionally positive definite function

AMS subject classifications. 41A05, 41A15, 41A25, 41A30, 41A63

DOI. 10.1137/050642113

1. Introduction.

One of the most important issues in many areas of applied mathematics and applications is to construct an approximation to scattered data: Given a set X of scattered points in $\Omega \subset \mathbb{R}^d$ and values $f|_X$ of some underlying function f, the goal is to find a function $s : \Omega \to \mathbb{R}^d$ such that s approximates f in some sense. Radial basis functions (RBFs) provide well-established tools for solving the scattered data approximation or interpolation problem. In addition, they are becoming increasingly popular for the numerical solution of partial differential equations.

A function $\phi : \mathbb{R}^d \to \mathbb{R}$ is radial in the sense that $\phi(x) = \Phi(|x|)$, where $|x| := (x_1^2 + \cdots + x_d^2)^{1/2}$ stands for the usual Euclidean norm. The common choices of ϕ can be divided into two groups, piecewise smooth such as the thin plate spline $\phi(x) = |x|^2 \log |x|$ and infinitely smooth such as multiquadrics $\phi(x) = (|x|^2 + \lambda^2)^{1/2}$. In this paper, we are interested in the infinitely smooth basis functions ϕ so that each function ϕ can be expanded as

$$\phi(x) = \Phi(|x|) = \sum_{n=0}^{\infty} c_n |x|^{2n},$$

which actually means ϕ is analytic. An attractive feature of using an infinitely smooth function is that it can provide spectral approximation order [10, 11, 13, 16, 17]. Typical examples of such basis functions are given as follows: For $d \in \mathbb{N}$ and $\lambda > 0$,

- $\phi(x) := (|x|^2 + \lambda^2)^{m-d/2}$, d odd, $m - d/2 > 0$, (multiquadrics),
- $\phi(x) := (|x|^2 + \lambda^2)^{m-d/2} \log(|x|^2 + \lambda^2)^{1/2}$, $m > d/2$, d even, ('shifted' surface splines).

*Received by the editors October 7, 2005; accepted for publication (in revised form) November 27, 2006; published electronically July 11, 2007.
http://www.siam.org/journals/sima/39-2/64211.html

†Corresponding author: Jungho Yoon. Department of Mathematics, Ewha W. University, Seoul, 120-750, S. Korea (lee08@ewhain.net, yoon@math.ewha.ac.kr). This work has been supported by grant R01-2006-000-10424-0 from Korea Science and Engineering Foundation in Ministry of Science and Technology.

‡School of Mathematics, KIAS, Seoul, 130-722, S. Korea (ykj@kias.re.kr).

- $\phi(x) := (|x|^2 + \lambda^2)^{m-d/2}$, $0 < m < d/2$, (inverse multiquadrics),
- $\phi(x) := e^{-|x|^2/\lambda^2}$, (Gaussians),

These functions are considered as tempered distributions and have generalized Fourier transforms of the form

$$|\cdot|^{2m}\hat{\phi} = F \in L_\infty(\mathbb{R}^d), \quad m \geq 0, \tag{1.1}$$

where, indeed, the function F is nonnegative on $\mathbb{R}^d$ and positive at least on an open subset of $\mathbb{R}^d$. We will see in section 3 that it is an important ingredient for the convergence of increasingly flat RBF interpolants to multivariate polynomial interpolants. Further, the condition (1.1) is the major difference from the Bessel radial function ϕ_d [5], which is defined by

$$\phi_d(x) := \frac{J_{\frac{d}{2}-1}(|x|)}{|x|^{\frac{d}{2}-1}}, \quad d = 1, 2, \ldots, \tag{1.2}$$

where J_α denotes the first kind Bessel function of order α. The Fourier transform of ϕ_d is proportional to a Dirac distribution, that is, $\hat{\phi}_d \approx \delta(|\cdot| - 1)$ (see [7, p. 364]).

The RBFs can be scaled in such a way to be wider by a shape parameter $\epsilon > 0$, i.e., $\phi_\epsilon(x) := \phi(\epsilon x)$. It has been observed independently during the last decades that for smooth data, a very small value of ϵ gives very accurate results for both interpolation problems and solving elliptic partial differential equations. For this, the reader is referred to the recent papers [4, 8]; unfortunately, we have been unable to locate it in any previous references. In this case, the basis function becomes very flat and the condition number of the interpolation system grows rapidly.

In [3], Driscoll and Fornberg introduced a surprising observation that the limit of an RBF interpolant often exists and takes the form of a polynomial. Later, Fornberg, Wright, and Larsson [6] and, in parallel, Schaback [14] proved that the limiting RBF interpolant is a (multivariate) finite order polynomial interpolant, if it exists. In particular, Schaback [14] showed that interpolation with scaled Gaussians always converges to the de Boor–Ron polynomial interpolation when the Gaussian widths increase. Subsequent to carrying out this study, we became aware of another recent paper by Larsson and Fornberg [9]; they generated a matrix in (2.4) (hereafter, denoted by $\mathbf{B}_{p,K}$ with $0 \leq p \leq d$ and $K \geq 0$) from the Taylor expansion of the given RBF and proved that the nonsingularity of these matrices guarantees the convergence of the RBF interpolation to a multivariate polynomial interpolation, if it exists, as the shape parameter $\epsilon \to 0$. However, the nonsingularity of $\mathbf{B}_{p,K}$ is just conjectured and yet to be proved. Thus, the first goal of this paper is to prove this conjecture. Moreover, it is also observed in [9] that this convergence property holds for the commonly well known RBF interpolants, but some different behavior was seen for the Bessel RBF; we will see it also in section 3. There needs to be a clear discussion on the reason for the deviant behavior of the Bessel RBF interpolant. For this reason, our next aim is to provide a detailed proof and conditions which address this issue. The specific contribution of this study is given as follows:

- For any $0 \leq p \leq d$ and $K \geq 0$, the matrix $\mathbf{B}_{p,K}$ is proved to be nonsingular for all positive definite RBFs which satisfy the condition (1.1). It implies that a positive definite RBF interpolant converges to a multivariate polynomial interpolant, if it exists, as the shape parameter $\epsilon \to 0$. In fact, the existence of the limiting interpolant is guaranteed when the set X is unisolvent. However, if the set X is nonunisolvent, it is still an open problem to describe the condition under which the limit exists.

- We provide more detailed discussion on the difference between the commonly used RBFs and the Bessel RBF. It is verified that the matrix $\mathbf{B}_{p,K}$ for the Bessel RBF is singular (as is conjectured by Larsson and Fornberg in [9]) whenever $K \geq p+2$.
- The state-of-the-art studies on the limit of increasing flat RBF have considered the interpolation from the space $\mathrm{span}\{\phi(\varepsilon(x-x_j)) : x_j \in X\}$ [8, 9, 14]. However, usually, when ϕ is conditionally positive definite of order m, some suitable polynomial of degree $m-1$ is added for the construction of interpolation. Thus, this study is concerned with RBF interpolation from the augmented space $\mathrm{span}\{\phi(\varepsilon(x-x_j)) : x_j \in X\} + \Pi_{<m}$. Specifically, the matrix $\mathbf{B}_{p,K}$ is modified for conditionally positive definite RBFs of order $m > 0$ and proved to be nonsingular. Then, we show that the corresponding RBF interpolant tends to a polynomial interpolant, if it exists, as the shape parameter $\epsilon \to 0$.

We use the following notation throughout this paper. For $\alpha, \beta \in \mathbb{Z}_+^d := \{(\gamma_1, \dots, \gamma_d) \in \mathbb{Z}^d : \gamma_k \geq 0\}$, we set $\alpha! := \alpha_1! \cdots \alpha_d!$, $|\alpha|_1 := \sum_{k=1}^d \alpha_k$, and $\alpha^\beta := \alpha_1^{\beta_1} \cdots \alpha_d^{\beta_d}$. The Fourier transform of $f \in L_1(\mathbb{R}^d)$ is defined as

$$\hat{f}(\theta) := \int_{\mathbb{R}^d} f(t) \exp(-i\theta \cdot t)\, dt.$$

Also, for a function $f \in L_1(\mathbb{R}^d)$, we use the notation $f^\vee$ for the inverse Fourier transform. The Fourier transform can be uniquely extended to the space of tempered distributions on $\mathbb{R}^d$. Let $\Pi_{<K}$ denote the space of d-variate algebraic polynomials of degree $< K$ on $\mathbb{R}^d$ and denote

$$K_d := \dim \Pi_{<K+1} = \binom{K+d}{d}. \tag{1.3}$$

2. Larsson and Fornberg's conjectures. In this section, we revisit the work of Larsson and Fornberg in [9] and introduce their conjectures. For this, the multi-index definitions from [9] are introduced.

DEFINITION 2.1. *Let $\alpha =: (\alpha_1, \dots, \alpha_d)$ and $\beta =: (\beta_1, \dots, \beta_d)$ be in $\mathbb{Z}_+^d$. Then we define the following terminologies:*

(a) *The multi-indices α, β are said to have the same parity if the components α_j and β_j with $j = 1, \dots, d$ have the same parity.*

(b) *The polynomial ordering of a sequence of multi-indices is determined in the way that the index α comes before β if $|\alpha|_1 < |\beta|_1$ or if $|\alpha|_1 = |\beta|_1$, $\alpha_j = \beta_j$, $j = 1, \dots, p$, and $\alpha_{p+1} < \beta_{p+1}$.*

DEFINITION 2.2. *Let I_K, where $K \geq 0$, be the polynomially ordered sequence of all multi-indices $\alpha = (\alpha_1, \dots, \alpha_d)$ such that $|\alpha|_1 \leq K$. For $n \in \mathbb{N}$, $I_K(n)$ denotes the nth multi-index in the sequence.*

DEFINITION 2.3. *Let $I_{p,K}$, where $0 \leq p \leq d$ and $K \geq 0$, be the polynomially ordered sequence of all multi-indices $\alpha = (\alpha_1, \dots, \alpha_d)$ with $|\alpha|_1 \leq K$ such that $\alpha_1, \dots, \alpha_p$ are odd numbers, the others are even numbers. For $n \in \mathbb{N}$, $I_{p,K}(n)$ denotes the nth multi-index in the sequence.*

DEFINITION 2.4. *Let I_{2n}^α be the polynomially ordered set of all multi-indices β such that $|\alpha + \beta|_1 = 2n$, and α and β have the same parity.*

Let $X = \{x_j : j = 1, \dots, N\}$ and $(K-1)_d < N \leq K_d$ with K_d in (1.3). Let $p_1, \dots, p_N$ be N independent polynomials in $\Pi_{<K+1}$. We say X is unisolvent with respect to $\{p_j : j = 1, \dots, N\}$ if there is a unique linear combination $\sum \beta_j p_j(x)$

such that it interpolates the data over the point set X. If X is nonunisolvent with respect to any choice of N linearly independent basis functions from $\Pi_{<K+1}$, there is a smallest integer $M > K$ such that we can form a minimal nondegenerate set $\{p_j : j = 1, \ldots, N\}$ chosen from the basis in $\Pi_{<M+1}$, that is,

$$\det(p_j(x_\ell) : j, \ell = 1, \ldots, N) \neq 0.$$

Then the degree of the minimal nondegenerate basis is said to be M. With this definition in hand, we introduce the relation between polynomial interpolation and the distribution of the set X.

PROPOSITION 2.5 (see [9]). *Let $X = \{x_1, \ldots, x_N\}$ and $(K-1)_d < N \leq K_d$ with K_d in (1.3). If X is unisolvent with respect to any set of N linearly independent polynomials in $\Pi_{<K+1}$, then*

(I) *if $N = K_d$, then there is a unique interpolating polynomial of degree K for any given data on X;*

(II) *if $N < K_d$, then there is an interpolating polynomial of degree K for any given data on X, for each choice of N linearly independent basis functions.*

If X is nonunisolvent and the degree of the minimal nondegenerate basis is M, then

(III) *there is an interpolating polynomial of degree M for any given data on X, for each choice of a minimal nondegenerate basis.*

Invoking that the RBFs ϕ considered in this paper are real analytic, the Taylor expansion of $\phi(x - x_j) = \Phi(|x - x_j|)$ with respect to $|x - x_j|$ is given as

$$\phi(x - x_j) = \sum_{n=0}^{\infty} c_n |x - x_j|^{2n}. \tag{2.1}$$

Here, the coefficient of x^α in the expansion of $|x - x_j|^{2n}$ can be written by

$$|x - x_j|^{2n}\big|_{x^\alpha} = \sum_{\beta \in I_{2n}^\alpha} (-1)^{|\alpha|_1} \frac{n!}{(\frac{\alpha+\beta}{2})!} \frac{(\alpha+\beta)!}{\alpha!\beta!} x_j^\beta.$$

For the proof of this identity, the reader is referred to [15]. Combining this with (2.1), we see that the coefficient of x^α in the expansion of $\phi(x - x_j)$ is

$$\phi(x - x_j)|_{x^\alpha} = \sum_{n=|\lfloor \frac{\alpha+\vec{1}}{2} \rfloor|_1}^{\infty} c_n \sum_{\beta \in I_{2n}^\alpha} (-1)^{|\alpha|_1} \frac{n!}{(\frac{\alpha+\beta}{2})!} \frac{(\alpha+\beta)!}{\alpha!\beta!} x_j^\beta, \tag{2.2}$$

where $\vec{1} = (1, \ldots, 1) \in \mathbb{R}^d$ and for any $\gamma = (\gamma_1, \ldots, \gamma_d) \in \mathbb{Z}_+^d$, $\lfloor \gamma \rfloor := (\lfloor \gamma_1 \rfloor, \ldots, \lfloor \gamma_d \rfloor)$ with $\lfloor s \rfloor$ the greatest integer less than or equal to s. Based on this expansion of ϕ, we define the symmetric function $\mathbf{B}_n(\alpha, \beta)$ by

$$\mathbf{B}_n(\alpha, \beta) := c_n (-1)^{|\alpha|_1} \frac{n!}{(\frac{\alpha+\beta}{2})!} \frac{(\alpha+\beta)!}{\alpha!\beta!}, \qquad |\alpha + \beta|_1 = 2n, \tag{2.3}$$

where $\alpha, \beta \in \mathbb{Z}_+^d \times \mathbb{Z}_+^d$ with the same parity. Then the conjectures suggested by Larsson and Fornberg are given as follows.

Larsson–Fornberg's Conjecture I. For $0 \leq p \leq d$ and $0 \leq K$, the matrices $\mathbf{B}_{p,K}$, defined by

$$\mathbf{B}_{p,K} := \big(\mathbf{B}_n(\alpha, \beta) : \alpha, \beta \in I_{p,K}\big), \tag{2.4}$$

are nonsingular for all commonly known RBFs such as (inverse) multiquadrics and Gaussians.

It is necessary to remark that in the above conjecture, the Bessel RBF is not included. The following theorem is the main result of [9].

THEOREM 2.6 (see [9]). *Let $X = \{x_1, \dots, x_N\}$ and $(K-1)_d < N \le K_d$ with K_d in* (1.3). *Assume that Larsson–Fornberg conjecture* I *holds. Then, we have the following properties:*

(a) *If the set X is of type* (I)*, then the limit of the RBF interpolant as the shape parameter $\varepsilon \to 0$ is the unique interpolating polynomial of degree K to the given data.*

(b) *If the set X is of type* (II)*, then the limit of the RBF interpolant as the shape parameter $\varepsilon \to 0$ is a polynomial of degree K that interpolates the given data. The exact polynomial depends on the choice of RBF.*

(c) *If the set X is of type* (III)*, then the limit of the RBF interpolant as the shape parameter $\varepsilon \to 0$, if the limit exists, is a polynomial of degree M that interpolates the given data.*

Remark. It has been proved by Schaback [14] that interpolation by scaled Gaussians always converges to the de Boor–Ron polynomial interpolant when the Gaussian widths increase. However, in the cases of multiquadrics and inverse multiquadrics, there occur some cases where the interpolants diverge in the limit; for examples, see [3] and [9]. Also, for the case of type (II), the reader is referred to the same papers [3] and [9] for an example of different limit polynomials affected by different RBFs.

In [9, Example 2.4], there occurs a case where all commonly used RBFs except the Bessel RBF have the same limit which is the unique interpolating polynomial; see also section 3. Thus, it is conjectured that the expansion coefficients of the Bessel RBF do not fulfill the nonsingularity condition of $\mathbf{B}_{p,J}$ in (2.4).

Larsson–Fornberg's Conjecture II. All matrices $\mathbf{B}_{p,K}$ with $K > 1$ are singular for the expansion coefficients of the Bessel radial basis function.

In the following section, we will prove that the Larsson–Fornberg conjectures are true. Also, the difference between the commonly used RBFs and Bessel RBF will be discussed. Furthermore, we extend Theorem 2.6 up to the case of RBF interpolation from the augmented space $\operatorname{span}\{\phi(x - x_j) : x_j \in X\} + \Pi_{<m}$, using conditionally positive definite ϕ of order $m \ge 0$.

3. The main results.

3.1. Positive definite function. We first prove that Larsson–Fornberg's conjecture I is true for all positive definite functions whose Fourier transforms $\hat{\phi}$ satisfy the condition (1.1).

DEFINITION 3.1. *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a continuous function and $\Omega \subset \mathbb{R}^d$. We say that ϕ is conditionally positive definite of order $m \in \mathbb{Z}_+ := \{0, 1, \dots\}$ if for every finite set of pairwise distinct points $X = \{x_1, \dots, x_N\} \subset \Omega$ and for every $\alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{R}^N \setminus 0$ satisfying*

$$\sum_{j=1}^{N} \alpha_j p(x_j) = 0, \quad p \in \Pi_{<m}, \tag{3.1}$$

the quadratic form

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \phi(x_i - x_j)$$

is positive definite. A conditionally positive definite function of order 0 *is called a positive definite function.*

Indeed, the condition (3.1) may seem a bit technical and hard to verify in practice. However, the positive definiteness of continuous and absolutely integrable functions ϕ is guaranteed when the Fourier transform $\hat{\phi}$ satisfies the condition (1.1) with $m = 0$. We note in passing that this argument is a consequence of the simple identity

$$\sum_{j=1}^{N}\sum_{k=1}^{N} \alpha_j \alpha_k \phi(x_j - x_k) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \hat{\phi}(\theta)\left|\sum_{j=1}^{N} \alpha_j e^{ix_j\cdot\theta}\right|^2 d\theta \tag{3.2}$$

for any $(\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N \setminus 0$ and the fact that the map $\theta \mapsto \sum_{j=1}^{N} \alpha_j e^{ix_j\cdot\theta}$, $\theta \in \mathbb{R}^d$, has zeros at most on a set of measure zero. This identity is also generalized to the case of conditionally positive definite functions ϕ of order $m > 0$.

Suppose that a continuous function $f : \mathbb{R}^d \to \mathbb{R}$ is known only at a set of discrete points $X := \{x_1, \ldots, x_N\}$ in $\Omega \subset \mathbb{R}^d$. A RBF interpolant to the data $(x_j, f(x_j))$, $j = 1, \ldots, N$, with a positive definite function $\phi(\varepsilon\cdot)$ is given by

$$s(x, \varepsilon) := \sum_{j=1}^{N} a_j \phi(\varepsilon(x - x_j)), \tag{3.3}$$

where the coefficients a_j $(j = 1, \ldots, N)$ are obtained by solving the linear system

$$s(x_j, \varepsilon) = f(x_j), \quad j = 1, \ldots, N. \tag{3.4}$$

LEMMA 3.2. *Let* $\mathbf{B}_n(\cdot,\cdot)$ *be the symmetric function defined as in* (2.3)*, and let* $\alpha, \beta \in \mathbb{Z}_+^d$ *with the same parity and* $|\alpha + \beta|_1 = 2n$*. Then, we have*

$$\mathbf{B}_n(\alpha, \beta) = (-1)^{|\alpha|_1} \frac{\phi^{(\alpha+\beta)}(0)}{\alpha!\beta!}.$$

Proof. Since ϕ is radially symmetric, $\phi(\cdot - x_j) = \phi(x_j - \cdot)$. Taking the Taylor expansion of $\phi(x_j - \cdot)$ around x_j yields the expression

$$\begin{aligned}
\phi(x_j - x) &= \sum_{|\alpha|_1=0}^{\infty} \phi^{(\alpha)}(x_j)\frac{(-x)^\alpha}{\alpha!}\\
&= \sum_{|\alpha|_1=0}^{\infty}\left(\sum_{|\beta|_1=0}^{\infty} \phi^{(\alpha+\beta)}(0)\frac{x_j^\beta}{\beta!}\right)\frac{(-x)^\alpha}{\alpha!}\\
&= \sum_{|\alpha|_1=0}^{\infty}\left(\sum_{n=|\lfloor\frac{\alpha+\vec{1}}{2}\rfloor|_1}^{\infty}\sum_{\beta\in I_{2n}^\alpha} \phi^{(\alpha+\beta)}(0)\frac{x_j^\beta}{\beta!}\right)\frac{(-x)^\alpha}{\alpha!};
\end{aligned}$$

where the second identity is a consequence of the Taylor expansion of $\phi^{(\alpha)}(x_j)$ around the origin. Then the coefficient of x^α in this expansion of $\phi(x_j - x)$ is denoted by

$$\phi(x - x_j)|_{x^\alpha} = \sum_{n=|\lfloor\frac{\alpha+\vec{1}}{2}\rfloor|_1}^{\infty}\sum_{\beta\in I_{2n}^\alpha} (-1)^{|\alpha|_1}\frac{\phi^{(\alpha+\beta)}(0)}{\alpha!\beta!}x_j^\beta. \tag{3.5}$$

Comparing (3.5) with (2.2), we obtain that

$$\mathbf{B}_n(\alpha,\beta) = (-1)^{|\alpha|_1} \frac{\phi^{(\alpha+\beta)}(0)}{\alpha!\beta!},$$

where $\alpha, \beta \in \mathbb{Z}_+^d$ with the same parity and $|\alpha+\beta|_1 = 2n$. ∎

The following lemma is useful for the proof of the nonsingularity of $\mathbf{B}_{p,K}$.

LEMMA 3.3. *Let ϕ be a positive definite function, and assume that ϕ satisfies the condition* (1.1) *with $m = 0$, i.e., the Fourier transform $\hat{\phi} \geq 0$ is positive on an open set in $\mathbb{R}^d$. For any integer $K > 0$, define the matrix $\mathbf{T}_\phi$ by*

$$\mathbf{T}_\phi := \mathbf{T}_{\phi,p,K} := \left((-i)^{|\alpha+\beta|_1} \frac{\phi^{(\alpha+\beta)}(0)}{\alpha!\beta!} : \alpha, \beta \in I_{p,K} \right). \tag{3.6}$$

Then $\mathbf{T}_\phi$ is positive definite, and so it is nonsingular.

Proof. Since $\alpha, \beta \in I_{p,K}$, both have the same parity and $|\alpha+\beta|_1$ is always even. It means that each entry of the matrix $\mathbf{T}_\phi$ is real. Now, let $\gamma := (\gamma_\alpha : \alpha \in I_{p,K})$ be an arbitrary nonzero vector. Then, the nonsingularity of $\mathbf{T}_\phi$ is guaranteed by showing that $\gamma \mathbf{T}_\phi \gamma^T > 0$, i.e.,

$$\gamma \mathbf{T}_\phi \gamma^T = \sum_{\beta \in I_{p,K}} \sum_{\alpha \in I_{p,K}} \gamma_\alpha \gamma_\beta (-i)^{|\alpha+\beta|_1} \frac{\phi^{(\alpha+\beta)}(0)}{\alpha!\beta!} > 0. \tag{3.7}$$

It is easy to see the identity

$$\sum_{\beta \in I_{p,K}} \sum_{\alpha \in I_{p,K}} \gamma_\alpha \gamma_\beta (-i)^{|\alpha+\beta|_1} \frac{\phi^{(\alpha+\beta)}(0)}{\alpha!\beta!} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{\phi}(\theta) \left| \sum_{\alpha \in I_{p,K}} \gamma_\alpha \frac{\theta^\alpha}{\alpha!} \right|^2 d\theta. \tag{3.8}$$

Since the Fourier transform $\hat{\phi} \geq 0$ is nonnegative on $\mathbb{R}^d$ and positive at least on an open subset of $\mathbb{R}^d$, the above term in (3.8) is positive, which completes the proof. ∎

Remark. In fact, it is easy to check from the above proof that any symmetric matrix built from $\mathbf{B}_n(\cdot,\cdot)$ is positive definite and so are all symmetrically chosen submatrices.

We now prove Larsson–Fornberg's conjecture I.

THEOREM 3.4. *Let ϕ be a positive definite function and assume that ϕ satisfies the condition* (1.1) *with $m = 0$, i.e., the Fourier transform $\hat{\phi} \geq 0$ is positive on an open set in $\mathbb{R}^d$. Then, for any integers $p, K \geq 0$, the matrix $\mathbf{B}_{p,K}$ in* (2.4) *is nonsingular.*

Proof. Using the form of $\mathbf{B}_{p,K}$ in Lemma 3.2, we see that $\det \mathbf{T}_\phi = c \det \mathbf{B}_{p,K}$ with $|c| = 1$. Thus, by Lemma 3.3, we conclude that $\mathbf{B}_{p,K}$ is nonsingular. ∎

As a conclusion, we can get the following results, which are actually restatements of Theorem 2.6.

THEOREM 3.5. *Let $X = \{x_1, \ldots, x_N\}$ and $(K-1)_d < N \leq K_d$ with K_d in* (1.3)*. Assume that ϕ is a positive definite function and its Fourier transform $\hat{\phi} \geq 0$ is positive on an open set in $\mathbb{R}^d$. Then, we have the following properties:*

(a) *If the set X is of type* (I)*, then the limit of the RBF interpolant $s(\cdot, \varepsilon)$ as the shape parameter $\varepsilon \to 0$ is the unique interpolating polynomial of degree K to the given data.*

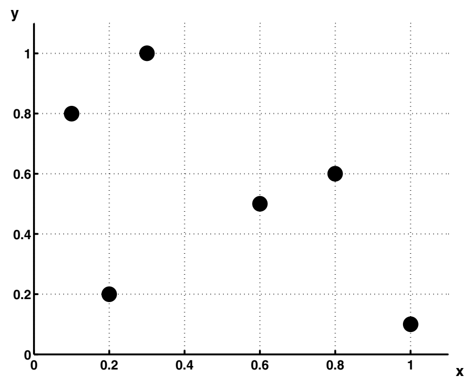


Fig. 1. *Scattered points X.*

(b) *If the set X is of type* (II)*, then the limit of the RBF interpolant $s(\cdot,\varepsilon)$ as the shape parameter $\varepsilon \to 0$ is a polynomial of degree K that interpolates the given data. The exact polynomial depends on the choice of RBFs.*

(c) *If the set X is of type* (III)*, then the limit of the RBF interpolant $s(\cdot,\varepsilon)$ as the shape parameter $\varepsilon \to 0$, if the limit exists, is a polynomial of degree M that interpolates the given data.*

An interesting example was observed in [9, Example 2.4] wherein all the well known RBF interpolants except the Bessel RBF (see (1.2)) interpolant have the same limit (as $\varepsilon \to 0$), which is the unique polynomial interpolant. We revisit this example as follows.

Example 3.6. *Let $X = \{(\frac{1}{10},\frac{4}{5}),(\frac{1}{5},\frac{1}{5}),(\frac{3}{10},1),(\frac{3}{5},\frac{1}{2}),(\frac{4}{5},\frac{3}{5})\}$ be a set of six points as in Figure* 1 *and $f(x_j) = \delta_{0,j}$ with $j = 1,\ldots,6$. These six points do not have any particular pattern. Expanding $\phi(\varepsilon(x_i - x_j))$ in powers of ε^2, the RBF interpolant $s(x,\varepsilon)$ to the given data $(x_j, f(x_j))$ can be written in the form* (see [6])

$$s(x,\varepsilon) = \frac{\varepsilon^{2r}p_{2r}(x) + \varepsilon^{2r+2}p_{2r+2}(x) + \ldots}{\varepsilon^{2q}c_{2q} + \varepsilon^{2q+2}c_{2q+2} + \ldots}, \qquad r,\, q \in \mathbb{N},$$

where c_{2n}, $n = q, q+1, \ldots,$ are some constants and $p_{2\ell}(x)$, $\ell = r, r+1, \ldots,$ are polynomials of degree (at most) 2ℓ. When $r = q$, the limit exists as $\varepsilon \to 0$ and it will be a polynomial of degree at most $2r$. Based on this expansion we programmed the limit of RBF interpolants and obtained the following results. For all the known positive definite RBFs (e.g., inverse multiquadrics and Gaussians), the interpolants converge to the same two-variable polynomial of degree 2*, that is,*

$$\lim_{\varepsilon\to 0} s(x,\varepsilon) = \frac{1}{28274}(-7711 - 81420x + 132915y + 82300x^2 - 55450xy - 91550y^2),$$

which is the unique polynomial interpolant to the given data. However, the Bessel function interpolant converges to a third order polynomial

$$\begin{aligned}\lim_{\varepsilon\to 0} s(x,\varepsilon) = \frac{1}{1017250518}(&-354545067 - 2047021330x + 4593056085y\\ &+ 255438330x^2 - 4166831700xy - 2554383300y^2\\ &- 310763000x^3 + 1319845500x^2y + 932289000xy^2 - 439948500y^3).\end{aligned}$$

It is known that the Bessel radial function ϕ_d, $d > 1$, is positive definite (see [5, Theorem 3.1]). However, its Fourier transform is proportional to a Dirac distribution, i.e., $\hat{\phi}_d \approx \delta(|\cdot| - 1)$ (see [7, p. 364]), which does not satisfy the condition (1.1). As far as we observed, this is the major difference between ϕ_d and the well known RBS. In the next theorem, we prove Larsson–Fornberg's conjecture II, that is, the matrix $\mathbf{B}_{p,K}$ of ϕ_d is singular whenever $p + 2 \leq K$. It explains the deviant behavior of ϕ_d as in Example 3.6.

THEOREM 3.7. *Let $0 \leq p \leq d$ and $p + 2 \leq K$ with $p, K \in \mathbb{N}$. Then the matrix $\mathbf{B}_{p,K}$ of the Bessel RBF ϕ_d is singular.*

Proof. Recall the following Hankel transform on $\mathbb{R}^d$ (see [2, p. 53]):

$$\phi_d(x) = \frac{1}{(2\pi)^{d/2}} \int_{|\theta|=1} e^{ix\cdot\theta} d\theta, \quad x \in \mathbb{R}^d. \tag{3.9}$$

Invoking the definition of $\mathbf{T}_{\phi_d}$ in (3.6), for an arbitrarily given $\gamma = (\gamma_\alpha : \alpha \in I_{p,K})$, we get

$$\begin{aligned}\gamma \mathbf{T}_{\phi_d} \gamma^T &= \sum_{\beta \in I_{p,K}} \sum_{\alpha \in I_{p,K}} \gamma_\alpha \gamma_\beta (-i)^{|\alpha+\beta|_1} \frac{\phi_d^{(\alpha+\beta)}(0)}{\alpha!\beta!} \\ &= \frac{1}{(2\pi)^{d/2}} \int_{|\theta|=1} \left| \sum_{\alpha \in I_{p,K}} \gamma_\alpha \frac{\theta^\alpha}{\alpha!} \right|^2 d\theta \geq 0,\end{aligned}$$

which implies that all the eigenvalues of $\mathbf{T}_{\phi_d}$ are nonnegative real numbers. In fact, since $p + 2 \leq K$, we can choose γ_α with $\alpha \in I_{p,K}$ such that

$$\left| \sum_{\alpha \in I_{p,K}} \gamma_\alpha \frac{\theta^\alpha}{\alpha!} \right|^2 = \left| \theta^{I_{p,K}(1)} [1 - (\theta_1^2 + \cdots + \theta_d^2)] \right|^2, \qquad \theta =: (\theta_1, \ldots, \theta_d),$$

where $I_{p,K}(1)$ is the first index in $I_{p,K}$. Then $\gamma \mathbf{T}_{\phi_d} \gamma^T$ becomes zero, which implies that the matrix $\mathbf{T}_{\phi_d}$ is not positive definite but semipositive definite. It leads to the conclusion that $\mathbf{T}_{\phi_d}$ has an eigenvalue of zero such that $\mathbf{T}_{\phi_d}$ is singular. Since $\det(\mathbf{T}_{\phi_d}) = c \det(\mathbf{B}_{p,K})$ with $|c| = 1$, $\mathbf{B}_{p,K}$ is singular. ∎

3.2. Conditionally positive definite function. The RBF interpolant to the data $(x_j, f(x_j))$, $j = 1, \ldots, N$, with a conditionally positive definite function ϕ of order m is given by

$$s(x, \varepsilon) := \sum_{j=1}^{N} a_j \phi(\varepsilon(x - x_j)) + \sum_{i=1}^{(m-1)_d} b_i p_i(x), \tag{3.10}$$

where $p_1, \ldots, p_{(m-1)_d}$ is a basis for the space $\Pi_{<m}$. The coefficients a_j $(j = 1, \ldots, N)$ and b_i $(i = 1, \ldots, (m-1)_d)$ are obtained by solving the linear system which can be written in a matrix form as

$$\begin{pmatrix} \mathbf{A} & \mathbf{P} \\ \mathbf{P^T} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{0} \end{pmatrix}, \tag{3.11}$$

where $\mathbf{A}$ and $\mathbf{P}$ are the $N \times N$ and $N \times (m-1)_d$ matrices that have the elements $\mathbf{A}_{ij} = \phi(\varepsilon(x_i - x_j))$ and $\mathbf{P}_{ij} = p_j(x_i)$, respectively. Further, $\mathbf{a} \in \mathbb{R}^N$ and $\mathbf{b} \in \mathbb{R}^{(m-1)_d}$

are the vectors of coefficients of $s(\cdot,\varepsilon)$, and the components of $\mathbf{f}$ are the data $f(x_j)$, $j=1,\ldots,N$. Here, for $m>0$, we require X to have the nondegeneracy property for $\Pi_{<m}$, i.e.,

$$q(x_j)=0, \quad 1\le j\le N \quad \text{for} \quad q\in\Pi_{<m} \quad \text{implies} \quad q=0. \tag{3.12}$$

The unique solution of the previous linear system is guaranteed when the function ϕ is conditionally positive definite of order m [12].

DEFINITION 3.8. *Let $I_{p,m,K}$, where $0\le p\le d$ and $m,K\ge 0$, be the polynomially ordered sequence of all multi-indices $\alpha=(\alpha_1,\ldots,\alpha_d)$ with $m\le|\alpha|_1\le K$ such that $\alpha_1,\ldots,\alpha_p$ are odd numbers, and the others are even numbers. For $n\in\mathbb{N}$, $I_{p,m,K}(n)$ denotes the nth multi-index in the sequence.*

Here and in what follows, assume that ϕ is conditionally positive definite of order $m\ge 0$. Recalling the Taylor expansion of ϕ, that is,

$$\phi(x-x_j)=\sum_{n=0}^{\infty}c_n|x-x_j|^{2n}, \tag{3.13}$$

and using the symmetric function $\mathbf{B}_n(\cdot,\cdot)$ in (2.3), let us define the matrix $\mathbf{B}_{p,m,K}$ corresponding to the polynomially ordered sequence $I_{p,m,K}$ by

$$\mathbf{B}_{p,m,K}:=\big(\mathbf{B}_n(\alpha,\beta):\alpha,\beta\in I_{p,m,K}\big). \tag{3.14}$$

Then, we will show that the matrix $\mathbf{B}_{p,m,K}$ is nonsingular.

THEOREM 3.9. *Let ϕ satisfy the condition* (1.1). *Then, the matrix $\mathbf{B}_{p,m,K}$ is nonsingular for any integers $p,m,K>0$.*

Proof. From Lemma 3.2, we find that

$$\mathbf{B}_{p,m,K}=\left((-1)^{|\alpha|_1}\frac{\phi^{(\alpha+\beta)}(0)}{\alpha!\beta!}:\alpha,\beta\in I_{p,m,K}\right).$$

Thus, as in Theorem 3.4, it suffices to prove the nonsingularity of the matrix $\mathbf{T}_\phi^{[m]}$ defined by

$$\mathbf{T}_\phi^{[m]}:=\mathbf{T}_{\phi,p,K}^{[m]}:=\left((-i)^{|\alpha+\beta|_1}\frac{\phi^{(\alpha+\beta)}(0)}{\alpha!\beta!}:\alpha,\beta\in I_{p,m,K}\right).$$

For this, we show $\gamma\mathbf{T}_\phi^{[m]}\gamma^T>0$ for any nonzero vector $\gamma:=(\gamma_\alpha:\alpha\in I_{p,m,K})$. Indeed, we note that this is a consequence of the following relation:

$$\begin{aligned}\gamma\mathbf{T}_\phi^{[m]}\gamma^T&=\sum_{\beta\in I_{p,m,K}}\sum_{\alpha\in I_{p,m,K}}\gamma_\alpha\gamma_\beta(-i)^{|\alpha+\beta|_1}\frac{\phi^{(\alpha+\beta)}(0)}{\alpha!\beta!}\\&=\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\hat\phi(\theta)\left|\sum_{\alpha\in I_{p,m,K}}\gamma_\alpha\frac{\theta^\alpha}{\alpha!}\right|^2d\theta.\end{aligned} \tag{3.15}$$

Since $|\cdot|^{2m}\hat\phi=F\ge 0$ and ϕ is real analytic, we deduce that the function F decays faster than any polynomial degree around ∞. Thus, the last integral in (3.15) makes sense and is positive (see (1.1)). It completes the proof. ∎

The following theorem treats the case of conditionally positive definiteness RBFs. The proof will be given in section 4.

THEOREM 3.10. *Let ϕ be a conditionally positive definite RBF of order $m \geq 0$ with the condition* (1.1). *Let $X = \{x_1, \ldots, x_N\}$ and $(K-1)_d < N \leq K_d$, where $K_d = \dim \Pi_{<K+1}$ and $m < K$. Then, we have the following properties:*

(a) *If the set X is of type* (I)*, then the limit of the RBF interpolant* (3.10) *as the shape parameter $\varepsilon \to 0$ is the unique interpolating polynomial of degree K to the given data.*

(b) *If the set X is of type* (II)*, then the limit of the RBF interpolant* (3.10) *as the shape parameter $\varepsilon \to 0$ is a polynomial of degree K that interpolates the given data. The exact polynomial depends on the choice of RBF.*

(c) *If the set X is of type* (III)*, then the limit of the RBF interpolant* (3.10) *as the shape parameter $\varepsilon \to 0$, if the limit exists, is a polynomial of degree $2K - m$ that interpolates the given data.*

Next, we introduce some examples of RBFs which satisfy the condition (1.1).

EXAMPLE 3.11. *Let the RBF ϕ be chosen to be one of the following:*

(a) $\phi(x) := (|x|^2 + \lambda^2)^{m-d/2}$, *$m > d/2$, $m - d/2 \notin 2\mathbb{Z}$ (multiquadrics);*

(b) $\phi(x) := (|x|^2 + \lambda^2)^{m-d/2} \log(|x|^2 + \lambda^2)^{1/2}$, *$m > d/2$, d even ("shifted" surface splines);*

(c) $\phi(x) := (|x|^2 + \lambda^2)^{m-d/2}$, *$0 < m < d/2$ (inverse multiquadrics);*

(d) $\phi(x) := e^{-|x|^2/\lambda^2}$ *(Gaussians);*

where $d \in \mathbb{N}$ and $\lambda > 0$. In the sense of tempered distributions, the functions ϕ in (a)*,* (b)*, and* (c) *have generalized Fourier transforms of the form* (*see* [7])

$$\hat{\phi}(\theta) = c_{m,\lambda,d} |\theta|^{-2m} \tilde{K}_m(|\lambda\theta|), \tag{3.16}$$

where $c_{m,\lambda,d}$ is a positive constant depending on m, λ and d, and where $\tilde{K}_\nu(|t|) := |t|^\nu K_\nu(|t|)$ with $K_\nu(|t|)$ the modified Bessel function of order ν. From [1]*, we find that*

$$\begin{aligned} &\tilde{K}_\nu(|t|) \in C^{2\nu-1}(\mathbb{R}^d) \cap C^\infty(\mathbb{R}^d \setminus \{0\}), \\ &\tilde{K}_\nu(|t|) > 0, \quad \textit{and} \quad \tilde{K}_\nu(|t|) \approx e^{-|t|}(1 + |t|^{(\nu-1/2)}). \end{aligned} \tag{3.17}$$

In the case of the Gaussian RBF ϕ in (d)*, its Fourier transform is of the form*

$$\hat{\phi}(\theta) := c_0 e^{-|\theta|^2/c_1^2}, \quad c_0, c_1 > 0.$$

4. A proof of Theorem 3.10. The general technique for the proofs is similar to the method given in [9], although the interpolant has an additional polynomial term associated with the order of conditionally positive definiteness of ϕ.

The RBF $\phi(\varepsilon\cdot)$ can be written in even powers of $|x|$ as follows:

$$\phi_\varepsilon(x) := \phi(\varepsilon x) = \sum_{j=0}^{\infty} c_j \varepsilon^{2j} |x|^{2j}. \tag{4.1}$$

Then the entries of the interpolation matrix $\mathbf{A} = (\phi(\varepsilon(x_i - x_j)) : x_i, x_j \in X)$ can be expanded in even powers of ε as above. The coefficients a_j, $j = 1, \ldots, N$, and b_i, $i = 1, \ldots, (m-1)_d$ in (3.10) are obtained by Cramer's rule, and they must be rational functions of ε^2. That is, there exists an integer q such that we can write

$$a_j =: \varepsilon^{-2K} \sum_{n=-q}^{\infty} \varepsilon^{2n} a_{j,n}, \qquad b_i =: \varepsilon^{-2K} \sum_{n=-q}^{\infty} \varepsilon^{2n} b_{i,n}, \tag{4.2}$$

where at least one of $(a_{j,-q} : j = 1, \ldots, N)$ and $(b_{j,-q} : j = 1, \ldots, (m-1)_d)$ is a nonzero vector. Now, for each $n \geq -q$ and $\beta \in \mathbb{Z}_+^d$, the discrete moments of $(a_{1,n}, \ldots, a_{N,n})^T$ are defined by

$$\sigma_n^{[\beta]} := \sum_{j=1}^{N} a_{j,n} x_j^{\beta}, \quad n = -q, -q+1, \ldots. \tag{4.3}$$

Then, for the moments $\sigma_n^{[\beta]}$ with $|\beta|_1 \leq m-1$ and $n \geq -q$, we have the following estimate.

LEMMA 4.1. *Let $\beta \in \mathbb{Z}_+^d$ with $|\beta|_1 \leq m-1$. Then, $\sigma_n^{[\beta]} = 0$ for any $n \geq -q$.*

Proof. Multiplying by ε^{2n-2K} by both sides of (4.3) and summing over $n = -q, -q+1, \ldots$, we obtain from (4.2) that for any $\varepsilon > 0$

$$\begin{aligned} \varepsilon^{-2K} \sum_{n=-q}^{\infty} \varepsilon^{2n} \sigma_n^{[\beta]} &= \varepsilon^{-2K} \sum_{j=1}^{N} \sum_{n=-q}^{\infty} \varepsilon^{2n} a_{j,n} x_j^{\beta} \\ &= \sum_{j=1}^{N} a_j x_j^{\beta} = 0, \end{aligned}$$

where the last identity is a simple consequence of (3.11). Thus, from this relation, we deduce the required result $\sigma_n^{[\beta]} = 0$ with $n \geq -q$. ∎

Let S_N indicate an N-set of polynomially ordered multi-indices, i.e., $\#S_N = N$. Define the matrix $\mathbf{V}$ by

$$\mathbf{V} = (x_j^{\beta} : j = 1, \ldots, N, \ \beta \in S_N).$$

Then from (4.3), we have

$$\mathbf{s}_n = \mathbf{V}\, \mathbf{a}_n, \quad n = -q, -q+1, \ldots, \tag{4.4}$$

where $\mathbf{s}_n := (\sigma_n^{[\beta]} : \beta \in S_N)^T$ and $\mathbf{a}_n := (a_{j,n} : j = 1, \ldots, N)^T$.

Recall that the RBF interpolant with a conditionally positive definite function ϕ of order m is given by

$$s(x, \varepsilon) := \sum_{j=1}^{N} a_j \phi(\varepsilon(x - x_j)) + \sum_{i=1}^{(m-1)_d} b_i p_i(x), \tag{4.5}$$

where $p_1, \ldots, p_{(m-1)_d}$ is a basis for $\Pi_{<m}$. Due to [9, page 123], we find that

$$\sum_{j=1}^{N} a_j \phi(\varepsilon(x - x_j)) = \varepsilon^{-2K} \sum_{s=0}^{\infty} \varepsilon^{2(-q+s)} P_{-q+s}(x) \tag{4.6}$$

with

$$P_{-q+s}(x) = \sum_{\alpha \in I_{2s}} \left(\sum_{n=|\lfloor \frac{\alpha+\vec{1}}{2} \rfloor|_1}^{s} c_n \sum_{\beta \in I_{2n}^{\alpha}} (-1)^{|\alpha|} \frac{n!}{(\frac{\alpha+\beta}{2})!} \frac{(\alpha+\beta)!}{\alpha!\beta!} \sigma_{-q+s-n}^{[\beta]} \right) x^{\alpha}. \tag{4.7}$$

Applying Lemma 4.1, we note that all the coefficients of x^α with $2s-m+1 \le |\alpha|_1 \le 2s$ become zero. It follows that

$$\deg(P_{-q+s}) \le 2s - m. \tag{4.8}$$

Further, inserting (4.2) into (4.5), we obtain that

$$s(x,\varepsilon) = \varepsilon^{-2K} \sum_{s=0}^{\infty} \varepsilon^{2(-q+s)} (P_{-q+s}(x) + Q_{-q+s}(x)), \tag{4.9}$$

where

$$Q_{-q+s}(x) = \sum_{i=1}^{(m-1)_d} b_{i,-q+s} p_i(x). \tag{4.10}$$

Let $C_1 := \{1,\ldots,p\}$ and $C_i = \{i_1,\ldots,i_p : i < \cdots < i_p\}$, $i = 2,\ldots,\binom{d}{p}$, be distinct subsets of $\{1,\ldots,d\}$. Let τ_i be a permutation on $\{1,\ldots,d\}$ such that if $i_j \in C_i$, then $\tau_i(i_j) = j \in C_1$.

DEFINITION 4.2. *We define $I^i_{p,m,K}$ to be the ordered set of all multi-indices $\alpha = (\alpha_{\tau_i(1)},\ldots,\alpha_{\tau_i(d)}) \in \mathbb{Z}^d_+$ with $m \le |\alpha|_1 \le K$ such that $I^i_{p,m,K}(k)$ is a rearrangement of the components of $I^1_{p,m,K}(k)$, where $I^1_{p,m,K} = I_{p,m,K}$ and $I^i_{p,m,K}(k)$ indicates the kth element in $I^i_{p,m,K}$.*

The reader is referred to [9] for an example of such permutation. Now, from this definition, we decompose the set $\tilde{I}_{m,K} := \{\alpha \in \mathbb{Z}^d_+ : m \le |\alpha|_1 \le K\}$ into the disjoint union of $I^i_{p,m,K}$'s as follows:

$$\tilde{I}_{m,K} = \bigcup_{p,i} \left\{ I^i_{p,m,K} : p = 0,\ldots,d \text{ and } i = 1,\ldots,\binom{d}{p} \right\}.$$

For $J \ge m$ and $0 \le p \le d$, let $\alpha \in I^i_{p,m,J}$ be chosen. We investigate the coefficient of x^α in $P_{-q+s}(x)$ (hereafter, denoted by $P_{-q+s}[\alpha]$) with $2s = J + |\alpha|_1$. For this, using the definition of $\mathbf{B}_n(\cdot,\cdot)$ in (2.3), we rewrite the coefficient $P_{-q+s}[\alpha]$ in (4.7) as follows:

$$P_{-q+s}[\alpha] = \sum_{n=|\lfloor \frac{\alpha+\vec{1}}{2} \rfloor|_1}^{s} \sum_{\beta \in I^\alpha_{2n}} \mathbf{B}_n(\alpha,\beta)\sigma^{[\beta]}_{-q+s-n}. \tag{4.11}$$

Note that for any given n with $|\lfloor \frac{\alpha+\vec{1}}{2} \rfloor|_1 \le n \le s$, the index β in the second summation of the right-hand side in (4.11) has the same parity as $\alpha \in I^i_{p,m,J}$ and satisfies $|\beta|_1 \le J$ because $|\alpha+\beta|_1 = 2n \le 2s$. Also, due to Lemma 4.1, we can set $|\beta|_1 \ge m$. As a consequence, since $|\alpha|_1 = 2s - J$ and $|\alpha+\beta|_1 = 2n$, the set of all such indices β is exactly the set $I^i_{p,m,J}$ so that $P_{-q+s}[\alpha]$ can be given as

$$P_{-q+s}[\alpha] = \sum_{\beta \in I^i_{p,m,J}} \mathbf{B}_n(\alpha,\beta)\sigma^{[\beta]}_{-q+(J-|\beta|_1)/2}. \tag{4.12}$$

Next, in order to continue our argument, let us define the following two vectors:

$$\mathbf{s}^i_{p,m,J} := (\sigma^{[\beta]}_{-q+n} : \beta \in I^i_{p,m,J},\ n = (J-|\beta|_1)/2),$$
$$\mathbf{p}^i_{p,m,J} := (P_{-q+s}[\alpha] : \alpha \in I^i_{p,m,J},\ s = (J+|\alpha|_1)/2).$$

Moreover, from the definition of $\mathbf{B}(\cdot,\cdot)$ in (2.3), we see that $\mathbf{B}_n(\alpha,\beta)$ is independent of the permutation τ_i in Definition 4.2, that is,

$$\mathbf{B}_{p,m,J} = \left(\mathbf{B}_n(\alpha,\beta) : \alpha,\beta \in I^i_{p,m,J}\right), \quad i = 2,\dots,\binom{d}{p}.$$

With the matrices from Definition 3.14 and the vectors defined above, we have a sequence of systems of equations for the discrete moments,

$$\mathbf{B}_{p,m,J}\,\mathbf{s}^i_{p,m,J} = \mathbf{p}^i_{p,m,J}, \quad i = 1,2,\dots,\binom{d}{p}, \quad \text{and} \quad J = m, m+1,\dots, \tag{4.13}$$

where p and J have the same parity. By the assumption on ϕ and Theorem 3.9, the systems in (4.13) are nonsingular and we have a complete description of the relation between the discrete moments and the polynomials P_{-q+s}. From the coefficients of the polynomials, the system in (4.13) can be solved directly for determining the moments because $\mathbf{B}_{p,m,J}$ is nonsingular for $0 \le p \le d$ and $J \ge 0$.

There is a range of ε-values for which we get a well defined interpolant $s(x,\varepsilon)$ in (4.5) to the data. If we relate this to the expansion (4.9), we get the following conditions:

(4.14)
- $P_K + Q_K$ interpolates the data on the set $X = \{x_j : j = 1,\dots,N\}$,
- $P_j + Q_j$, $j \ne K$ interpolates 0 on the set X.

Note that if $P_j + Q_j$, $j \ne K$ is of degree $< m$, then the nondegeneracy property (3.12) of X on $\Pi_{<m}$ implies $P_j + Q_j = 0$.

We now introduce two useful lemmas. First, recall that at least one of $(a_{j,-q} : j = 1,\dots,N)$ and $(b_{j,-q} : j = 1,\dots,(m-1)_d)$ is a nonzero vector (see the line below (4.2)). In practice, we will see that $(a_{j,-q} : j = 1,\dots,N)$ should be a nonzero vector.

LEMMA 4.3. *If $a_{j,-q} = 0$ for $j = 1,\dots,N$, then $b_{j,-q} = 0$ for $j = 1,\dots,(m-1)_d$.*

Proof. Assume that $a_{j,-q} = 0$ for $j = 1,\dots,N$. It is clear from Lemma 4.1 that $P_{-q}(x)$ in (4.7) is a zero polynomial. Since $P_{-q} + Q_{-q} = Q_{-q}$ interpolates 0 at X, the nondegeneracy property (3.12) induces that the polynomial $Q_{-q} = 0$. The definition of Q_{-q} in (4.10) leads to the conclusion that $b_{j,-q} = 0$ with $j = 1,\dots,(m-1)_d$. ∎

LEMMA 4.4. *Let $\{p_j(x) : j = 0,\dots,K\}$ be a finite set of polynomials such that $p_0 = 0$ and $p_K \ne 0$, and denote $k := \max\{\ell \ge 0 : p_j = 0,\ j = 0,\dots,\ell\}$. Suppose that*
 (i) *$p_j = 0$ or $\deg(p_j) > K$ for any $j = 0,\dots,K-1$;*
 (ii) *$\deg(p_j) \le 2j - k - 1$ for any $j = k+1,\dots,K$.*

Then $k = K-1$ (i.e., $p_j = 0\ \forall j = 0,\dots,K-1$) and $\deg(p_K) \le K$.

Proof. Suppose, contrary to our claim, that $k < K-1$. From (ii), $\deg(p_{k+1}) \le k+1 < K$. Also, applying (i), we see that $\deg(p_{k+1}) > K$, which leads to a contradiction. Thus, we conclude that $k = K-1$ and by (ii), $\deg(p_K) \le K$. ∎

Now we are ready to prove Theorem 3.10. We consider the case when the set X is of type (I), i.e., $\#X = K_d$, where $K_d = \dim \Pi_{<K+1}$ and $K > m$, and the set X is unisolvent with respect to any basis for $\Pi_{<K+1}$. Then, the relation (4.4) holds for the basis $\{x^\alpha : \alpha \in I_K\}$. Also, due to the unisolvency of X, any polynomial of degree $\le K$ that interpolates zero at N points must be identically zero.

Proof (a) *of Theorem* 3.10. Invoking the fact $\deg(P_{-q+s}) \le 2s - m$ from (4.8), the condition (4.14) shows that at least,

$$P_{-q+j} + Q_{-q+j} = 0 \quad \forall j = 0,\dots,\left\lfloor \frac{K+m}{2} \right\rfloor - 1.$$

Let κ ($\leq K-1$) be a positive integer such that $P_{-q+j}+Q_{-q+j}=0$ for any $j=m,\ldots,\kappa$ and $P_{-q+\kappa+1}+Q_{-q+\kappa+1}\neq 0$. Then

$$P_{-q+j}[\alpha]=0 \quad \text{for} \quad j=m,\ldots,\kappa, \quad |\alpha|_1\geq m,$$

where $P_{-q+j}[\alpha]$ is the coefficient of x^α in $P_{-q+j}(x)$. From the systems (4.13), we derive that for every $J=m,\ldots,\kappa$,

$$\sigma_{-q+n}^{[\beta]}=0 \quad \forall\beta\in I_{m,J} \quad \text{and} \quad 2n=J-|\beta|_1. \tag{4.15}$$

Substituting (4.15) into (4.7), we obtain that for every $s\geq\kappa+1$, $\deg(P_{-q+s})\leq 2s-\kappa-1$, i.e., $\deg(P_{-q+s}+Q_{-q+s})\leq 2s-\kappa-1$. By applying $\{P_{-q+j}+Q_{-q+j}\}_{j=0}^{K}$ to Lemma 4.4, we have that $P_{-q+j}+Q_{-q+j}=0$ for $j=0,\ldots,K-1$ and $\deg(P_{-q+K}+Q_{-q+K})\leq K$. It clearly induces that $\deg(P_{-q+K})\leq K$ and $\deg(P_{-q+j})\leq m-1$ for $j=m,\ldots,K-1$, i.e.,

$$P_{-q+j}[\alpha]=0 \text{ for } \quad j=m,\ldots,K-1, \quad |\alpha|_1\geq m. \tag{4.16}$$

Now, we claim that

$$P_{-q+K}+Q_{-q+K}=P_K+Q_K \quad (\text{i.e., } q=0).$$

Assume, on the contrary, that $q>0$ so that $P_{-q+K}+Q_{-q+K}\neq P_K+Q_K$. Then, $P_{-q+K}+Q_{-q+K}$ interpolates zero on X and $\deg(P_{-q+K}+Q_{-q+K})\leq K$. Hence, $P_{-q+K}+Q_{-q+K}=0$. Since $\deg(Q_{-q+K})<m$ we deduce that

$$P_{-q+K}[\alpha]=0, \quad |\alpha|_1=m,\ldots,K.$$

Accordingly, together with (4.16), we solve the systems (4.13) for $J=m,\ldots,K$ to obtain

$$\sigma_{-q}^{[\beta]}=0, \quad |\beta|_1=m,\ldots,K. \tag{4.17}$$

Combining (4.17) with Lemma 4.1 and (4.4), we obtain that $a_{j,-q}=0$ for $j=1,\ldots,N$. Further, Lemma 4.3 shows that $b_{j,-q}=0$ for $j=1,\ldots,(m-1)_d$, which is impossible since not all of $a_j's$ and $b_i's$ are zeroes. Next, if $q<0$, it follows from (4.16) that $P_K[\alpha]=0$ for any $|\alpha|_1\geq m$, i.e., $\deg P_K<m$. Since $\deg Q_K<m$, $\deg(P_K+Q_K)<m$. Since X is unisolvent with respect to any set of N linearly independent polynomials in $\Pi_{<K+1}$, this is contradictory to (4.14). Thus, we arrive at the conclusion that $q=0$ and the interpolant $s(x,\varepsilon)$ in (4.9) becomes

$$s(x,\varepsilon)=P_K(x)+Q_K(x)+\varepsilon^2\sum_{\ell=1}^{\infty}\varepsilon^{2(\ell-1)}(P_{K+\ell}(x)+Q_{K+\ell}(x)).$$

Unisolvency of X to $\Pi_{<K+1}$ ensures that P_K+Q_K is the unique polynomial interpolant to the data on X. $\square$

In the case of type (II), the set X is unisolvent but the number of points does not coincide with the dimension of the polynomial space, i.e.,

$$(K-1)_d<N<K_d.$$

Thus, the condition that a polynomial $p(x)$ of degree $\leq K$ interpolates zero on X does not ensure $p(x)$ to be identically zero any more, unless its degree is less than K.

The technique for the proof of Theorem 3.10 (b) is almost the same as the method in [9, Theorem 4.2]; hence it is only sketched here.

LEMMA 4.5. *Assume that* $\deg(P_{-q+j}) < m$ *for* $j = 0, \dots, K-1$ *and* $\deg(P_{-q+K}) \le K$. *Then* P_{-q+K} *is a linear combination of* N *linearly independent polynomials of degree* $\le K$.

Proof. This lemma is immediate by applying the same analysis in the proof of [9, Theorem 4.2]. The only difference is that the matrix $\mathbf{B}_{p,K}$ in [9, Theorem 4.2] is replaced by $\mathbf{B}_{p,m,K}$ to consider the conditionally positive definite RBF ϕ of order m. $\square$

Proof (b) *of Theorem* 3.10. By applying $\{P_{-q+j} + Q_{-q+j}\}_{j=0}^{K-1}$ to Lemma 4.4, we obtain that $\deg(P_{-q+K-1}) \le K-1$ and $\deg(P_{-q+j}) \le m-1$ for $j = m, \dots, K-2$, i.e.,

$$P_{-q+j}[\alpha] = 0 \text{ for } \quad j = m, \dots, K-2, \quad |\alpha|_1 \ge m. \tag{4.18}$$

By unisolvency, we have that $P_{-q+K-1} + Q_{-q+K-1} = 0$. The systems (4.13) with $m \le J \le K-1$ yield that

$$\sigma_{-q+n}^{[\beta]} = 0 \quad \forall \beta \in I_{m,J} \quad \text{and} \quad 2n = J - |\beta|_1. \tag{4.19}$$

Then, by substituting (4.19) into (4.7) with $s = K$, we show that $\deg(P_{-q+K}) \le K$.

Now we are in a position to get $P_{-q+K} + Q_{-q+K} = P_K + Q_K$, i.e., $q = 0$. If $q < 0$, by the same method as in the proof of Theorem 3.10 (a), we can induce a contradiction. Suppose that $q > 0$ such that $P_{-q+K} + Q_{-q+K} \ne P_K + Q_K$. Then, $P_{-q+K} + Q_{-q+K}$ is a polynomial of degree $\le K$ which interpolates zero on X. Using Lemma 4.5, we find that $P_{-q+K} + Q_{-q+K} = 0$. It implies that all the coefficients of x^α with $|\alpha|_1 \ge m$ in P_{-q+i} are zero for every $i \le K$ because $\deg(Q_{-q+K}) \le m-1$. Applying (4.13) for $m \le J \le K$, we obtain that $\sigma_{-q}^{[\beta]} = 0$ for $m \le |\beta|_1 \le K$. Combining this with Lemma 4.1 and (4.4), it follows from Lemma 4.3 that $a_{j,-q} = b_{\ell,-q} = 0$, for $j = 1, \dots, N$ and $\ell = 1, \dots, (m-1)_d$, which is a contradiction to the fact that not all of $a_j's$ and $b_i's$ are zeroes. Thus, we must have $q = 0$. Hence, as $\varepsilon \to 0$, the RBF interpolant converges to the polynomial $P_K + Q_K$. $\square$

Proof (c) *of Theorem* 3.10. In this case, in order that the limit exists as $\varepsilon \to \infty$, the interpolant $s(x, \varepsilon)$ in (4.9) must be expressed as

$$s(x, \varepsilon) = P_K(x) + Q_K(x) + \varepsilon^2 (P_{K+1}(x) + Q_{K+1}(x)) + \cdots.$$

Then, the limit is $P_K(x) + Q_K(x)$ and by Lemma 4.1, $\deg(P_K + Q_K) \le 2K - m$. $\square$

Acknowledgment. We are grateful to the anonymous referees for their many valuable corrections and suggestions.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *A Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards Applied Mathematics Series 55, U.S. Government Printing Office, Washington, D.C., 1964.

[2] M. D. BUHMANN, *Radial Basis Functions: Theory and Implementations*, in Cambridge Monographs on Applied and Computational Mathematics 12, Cambridge University Press, Cambridge, UK, 2003.

[3] T. A. DRISCOLL AND B. FORNBERG, *Interpolation in the limit of increasingly flat radial basis functions*, Comput. Math. Appl., 43 (2002), pp. 413–422.

[4] B. Fornberg and G. Wright, *Stable computation of multiquadric interpolation for all values of the shape parameter*, Comput. Math. Appl., 48 (2004), pp. 853–867.

[5] B. Fornberg, E. Larsson, and G. Wright, *A new class of oscillatory radial basis function*, Comput. Math. Appl., 51 (2006), pp. 1209–1222.

[6] B. Fornberg, G. Wright, and E. Larsson, *Some observations regarding interpolants in the limit of flat radial basis functions*, Comput. Math. Appl., 47 (2004), pp. 37–55.

[7] I. M. Gelfand and G. E. Shilov, *Generalized Functions*, Vol. 1, Properties and Operations, Academic Press, NY, 1964.

[8] E. Larsson and B. Fornberg, *A numerical study of some radial basis function based solution methods for elliptic PDEs*, Comput. Math. Appl., 46 (2003), pp. 891–902.

[9] E. Larsson and B. Fornberg, *Theoretical and computational aspects of multivariate interpolation with increasingly flat radial basis functions*, Comput. Math. Appl., 49 (2005), pp. 103–130.

[10] W. R. Madych and S. A. Nelson, *Multivariate interpolation and conditionally positive definite functions,* II, Math. Comp., 54 (1990), pp. 211–230.

[11] W. R. Madych and S. A. Nelson, *Bounds on multivariate polynomials and exponential error estimates for multiquadric interpolation*, J. Approx. Theory, 70 (1992), pp. 94–114.

[12] C. A. Micchelli, *Interpolation of scattered data: Distances matrices and conditionally positive functions*, Constr. Approx., 2 (1986), pp. 11–22.

[13] M. J. D. Powell, *The theory of radial basis function approximation in* 1990, in Advances in Numerical Analysis, Vol. II: Wavelets, Subdivision Algorithms and Radial Basis Functions, W. A. Light, ed., Oxford University Press, New York, 1992, pp. 105–210.

[14] R. Schaback, *Multivariate interpolation by polynomials and radial basis functions*, Constr. Approx., 21 (2005), pp. 293–317.

[15] R. Schaback, *Limit problems for interpolation by analytic radial basis functions*, J. Comp. Appl. Math., (2006), to appear.

[16] Z.-M. Wu and R. Schaback, *Local error estimates for radial basis function interpolation of scattered data*, IMA J. Numer. Anal., 13 (1993), pp. 13–27.

[17] J. Yoon, *Spectral approximation orders of radial basis function interpolation on the Sobolev space*, SIAM J. Math. Anal., 33 (2001), pp. 946–958.

SIAM J. Math. Anal.
Vol. 39, No. 2, pp. 554–564

EXTENDING THE RANGE OF ERROR ESTIMATES FOR RADIAL APPROXIMATION IN EUCLIDEAN SPACE AND ON SPHERES*

R. A. BROWNLEE†, E. H. GEORGOULIS†, and J. LEVESLEY†

Abstract. We adapt Schaback's error doubling trick [R. Schaback, *Math. Comp.*, 68 (1999), pp. 201–216] to give error estimates for radial interpolation of functions with smoothness lying (in some sense) between that of the usual native space and the subspace with double the smoothness. We do this for both bounded subsets of $\mathbb{R}^d$ and spheres. As a step on the way to our ultimate goal we also show convergence of pseudoderivatives of the interpolation error.

Key words. multivariate interpolation, radial basis functions, error estimates, smooth functions

AMS subject classifications. 41A05, 41A15, 41A25, 41A30, 41A63

DOI. 10.1137/060650428

1. Introduction.

In this paper we are interested in extending the range of applicability of error estimates for radial basis function interpolation in Euclidean space and on spheres. Let Ω be a subset of $\mathbb{R}^d$, or the sphere. Let $d(x, y)$ denote the distance between two points in Ω. Let $Y \subset \Omega$ be a finite set of points, and measure the *fill-distance* of Y in Ω with

$$h(Y, \Omega) := \sup_{x \in \overline{\Omega}} \min_{y \in Y} d(x, y).$$

Given a univariate function ϕ defined on either $\mathbb{R}_+$ or $[0, \pi]$, depending on whether we are in Euclidean space or on the sphere, we form an approximation

$$S_\phi^Y(x) = \sum_{y \in Y} \alpha_y \phi(d(x, y)).$$

If the coefficients α_y, $y \in Y$, are determined by the interpolation conditions

$$S_\phi^Y(y) = f(y) \qquad \text{for } y \in Y,$$

we refer to S_ϕ^Y as the ϕ-spline interpolant to f on Y.

We will be approximating functions $f \in \mathcal{H}_\phi$, a Hilbert space of functions which depends on the function ϕ—the so-called *native space*. Later we will be more explicit about this space of functions. With this Hilbert space we have an inner product $\langle \cdot, \cdot \rangle_\phi$, with associated norm $\| \cdot \|_\phi$. We will require the following useful orthogonality and consequent *Pythagorean* property; see, e.g., [11, 13].

Proposition 1.1. *Let S_ϕ^Y be the ϕ-spline interpolant to f on the point set $Y \subset \Omega$. Then, for all $f \in \mathcal{H}_\phi$,*

1. $\langle f - S_\phi^Y, S_\phi^Y \rangle_\phi = 0$;
2. $\|f\|_\phi^2 = \|f - S_\phi^Y\|_\phi^2 + \|S_\phi^Y\|_\phi^2$.

*Received by the editors January 20, 2006; accepted for publication (in revised form) February 7, 2007; published electronically July 18, 2007.
http://www.siam.org/journals/sima/39-2/65042.html

†Department of Mathematics, University of Leicester, Leicester LE1 7RH, England (r.brownlee@mcs.le.ac.uk, e.georgoulis@mcs.le.ac.uk, j.levesley@mcs.le.ac.uk). The first author was partially supported by a studentship from the Engineering and Physical Sciences Research Council.

The usual error estimate for ϕ-spline interpolants is of the form

$$|f(x) - S_\phi^Y(x)| \le P(x, Y, \phi)\|f - S_\phi^Y\|_\phi,$$

where estimation of $P(x, Y, \phi)$—the so-called *power function*—leads to error estimates for interpolation in terms of the fill-distance $h(Y, \Omega)$. For the archetypal function in $\mathcal{H}_\phi$ we can say no more than $\|f - S_\phi^Y\|_\phi \to 0$ as $h(Y, \Omega) \to 0$. However, if f has double the smoothness (in some sense to be made clear later) of the typical function, then Schaback [19] has shown how to double the convergence order of the ϕ-spline interpolant.

We show how to get improved orders of convergence when the target function, f, has less smoothness than Schaback requires, but more smoothness than the typical function. We shall be doing this on the sphere (though this can easily be generalized to other two-point homogeneous spaces) in section 2 and in Sobolev spaces on Euclidean space in section 3. An intermediate result in both cases is to prove approximation orders for pseudoderivatives of the interpolant. We will define this notation at the appropriate place in each of the following sections.

In each case we shall be concerned with the practical scenario in which Y consists of a finite number of points. Forfeiting this assumption is of theoretical interest. In particular, in Euclidean space for (perturbed) gridded data, certain improved error estimates are already known to hold for functions within the native space itself (see, e.g., [4]).

The goal in this paper is quite different from the desire to establish error estimates for functions possessing insufficient smoothness for admittance in the native space. In recent years, contributions in that direction have been provided by several authors, e.g., [16, 17, 12] for the sphere and [23, 3, 18] for the Euclidean case.

2. The sphere. Let $\mathbb{S}^d = \{x \in \mathbb{R}^{d+1} : |x| = 1\}$. Then the geodesic distance between points $x, y \in \mathbb{S}^d$ is $d(x, y) = \cos^{-1} xy$, where xy denotes the usual inner product of vectors in $\mathbb{R}^{d+1}$. We let ν denote the normalized rotationally invariant measure on the sphere and define the inner product

$$\langle f, g\rangle_{L_2(\mathbb{S}^d)} := \int_{\mathbb{S}^d} f(x)g(x)\,\mathrm{d}\nu(x).$$

Let $\|\cdot\|_{L_2(\mathbb{S}^d)} := \langle\,\cdot\,,\,\cdot\,\rangle^{1/2}_{L_2(\mathbb{S}^d)}$, and let $L_2(\mathbb{S}^d)$ denote the set of functions for which $\|\cdot\|_{L_2(\mathbb{S}^d)} < \infty$. Let P_n be the polynomials of degree n in $\mathbb{R}^{d+1}$ restricted to the sphere, and let $H_n = P_n \cap P_n^\perp$ be the space of degree n spherical harmonics. Then, $L_2(\mathbb{S}^d)$ has the decomposition

$$L_2(\mathbb{S}^d) = \bigoplus_{n\ge 0} H_n.$$

Let $Y_1^n, \dots, Y_{d_n}^n$ be an orthonormal basis for H_n.

Related to $\mathbb{S}^d$ (we will see why shortly), we have the Gegenbauer polynomials $C_n^{(\lambda)}(t)$ which are orthogonal on $[-1, 1]$ with respect to the weight $(1-t^2)^{\lambda-1/2}$. It is well known (Müller [15], for instance) that the following addition formula holds:

$$C_n^{(\lambda)}(xy) = \sum_{j=1}^{d_n} Y_j^n(x)Y_j^n(y),$$

with $\lambda = d/2-1$. The normalization of the Gegenbauer polynomials is chosen so that there is no constant in the addition theorem. It is straightforward to see that $C_n^{(\lambda)}$ is the kernel of T_n, the orthogonal projector from $L_2(\mathbb{S}^d)$ onto H_n. Thus,

$$(T_n f)(x) = \int_{\mathbb{S}^d} f(y) C_n^{(\lambda)}(xy)\,\mathrm{d}\nu(y) \qquad \text{for all } f \in L_2(\mathbb{S}^d).$$

The following lemma is a specialization of a result in [11] to the sphere.

LEMMA 2.1. *For* $n \geq 0$,

$$\|T_n f\|_{L_\infty(\mathbb{S}^d)} \leq \sqrt{d_n}\|T_n f\|_{L_2(\mathbb{S}^d)} \qquad \textit{for all } f \in L_2(\mathbb{S}^d).$$

We will be considering interpolation using kernels of the form $\phi(d(x,y))$, where $\phi : [0,\pi] \to \mathbb{R}$. We will assume that the function ϕ has an expansion

$$\phi(d(x,y)) = \sum_{n\geq 0} a_n C_n^{(\lambda)}(xy),$$

where $a_n > 0$, for $n = 0, 1, \ldots$, and

$$\sum_{n\geq 0} d_n a_n < \infty.$$

The first condition ensures that ϕ is positive definite, and the second that it is continuous. Our analysis will take place in the native space for ϕ, $\mathcal{H}_\phi$, defined by

$$\mathcal{H}_\phi := \left\{ f \in L_2(\mathbb{S}^d) : \ \|f\|_\phi := \left(\sum_{n\geq 0} a_n^{-1} \|T_n f\|^2_{L_2(\mathbb{S}^d)} \right)^{\frac{1}{2}} < \infty \right\}.$$

A pseudodifferential operator Λ on $\mathbb{S}^d$ is an operator which acts via multiplication by a constant on each eigenspace H_n:

$$\Lambda p_n = \lambda_n p_n, \qquad p_n \in H_n,\ n = 0, 1, \ldots.$$

For more information on pseudodifferential operators on spheres, see, e.g., [9, 20]. We call the sequence of numbers $\{\lambda_n\}_{n\geq 0}$ the *symbol* of Λ. Let δ_x denote the point evaluation functional at x, and, when it makes sense for the functional μ, let $\Lambda\mu(f) = \mu(\Lambda f)$. Let us denote by Y^* the span of the point evaluation functionals supported on Y. Morton and Neamtu [14] give error estimates for the collocation solution of pseudodifferential equations on spheres. Here we attempt, initially, to find errors in pseudoderivatives of solutions to the interpolation problem.

PROPOSITION 2.2. *Let* S_ϕ^Y *be the* ϕ*-spline interpolant to* $f \in \mathcal{H}_\phi$ *on the point set* $Y \subset \mathbb{S}^d$. *Let* Λ *be a pseudodifferential operator. Then, for each* $x \in \mathbb{S}^d$,

$$|\Lambda(f - S_\phi^Y)(x)| \leq \inf_{\mu\in Y^*} \sup_{\substack{v\in\mathcal{H}_\phi \\ \|v\|_\phi = 1}} |\Lambda v(x) - \mu(v)| \|f - S_\phi^Y\|_\phi.$$

Proof. Since $f(y) - S_\phi(y) = 0$, $y \in Y$, we have, for any coefficients c_y, $y \in Y$,

$$
\begin{aligned}
|\Lambda(f - S_\phi^Y)(x)| &= \left|\Lambda(f - S_\phi^Y)(x) - \sum_{y \in Y} c_y (f(y) - S_\phi^Y(y))\right| \\
&= \left|\Lambda(f - S_\phi^Y)(x) - \sum_{y \in Y} c_y (f(y) - S_\phi^Y(y))\right| \frac{\|f - S_\phi^Y\|_\phi}{\|f - S_\phi^Y\|_\phi} \\
&\le \sup_{\substack{v \in \mathcal{H}_\phi \\ \|v\|_\phi = 1}} \left|\Lambda v(x) - \sum_{y \in Y} c_y v(y)\right| \|f - S_\phi^Y\|_\phi.
\end{aligned}
$$

We now take the infimum over all functionals in Y^* to obtain the result. ∎

In what follows we will need the pseudodifferential operator Λ to satisfy the following assumption.

Assumption 2.3. For all $n \ge 0$, $\lambda_n = (n(d + n - 2))^s$, for some $s > 0$.
From Ditzian [6], if Λ satisfies Assumption 2.3, then for $p \in P_n$,

$$\|\Lambda p\|_{L_\infty(\mathbb{S}^d)} \le E\lambda_n \|p\|_{L_\infty(\mathbb{S}^d)},$$

for some E independent of n.

From [10, Lemma 7] we have the following result.

LEMMA 2.4. *Let Y be a finite set of points with fill-distance $h(Y, \mathbb{S}^d) \le 1/(2N)$, for some fixed $N \in \mathbb{Z}_+$. Then, for any linear functional γ on P_N with*

$$\sup_{\substack{p \in P_N \\ \|p\|_{L_\infty(\mathbb{S}^d)} = 1}} |\gamma p| \le 1,$$

there is a set of real numbers $\{b_y\}_{y \in Y}$, with $\sum_{y \in Y} |b_y| \le 2$, such that

$$\gamma p = \sum_{y \in Y} b_y p(y) \qquad \text{for all } p \in P_N.$$

Now, for a fixed $x \in \mathbb{S}^d$, let

$$\gamma p = \frac{\Lambda p(x)}{E\lambda_N} \qquad \text{for all } p \in P_N.$$

Then,

$$\sup_{0 \ne p \in P_N} \frac{|\gamma p|}{\|p\|_{L_\infty(\mathbb{S}^d)}} \le 1,$$

so that, by the previous lemma, there is a set of coefficients $\{b_y\}_{y \in Y}$ such that

$$\gamma p = \sum_{y \in Y} b_y p(y),$$

with $\sum_{y \in Y} |b_y| \le 2$. Thus, with $c_y = E\lambda_N b_y$, for $y \in Y$, we have

$$\Lambda p(x) = \sum_{y \in Y} c_y p(y) \qquad \text{for all } p \in P_N, \tag{2.1}$$

where

$$\sum_{y\in Y}|c_y| \le 2E\lambda_N. \tag{2.2}$$

We now arrive at the first main result of this section.

THEOREM 2.5. *Let S_ϕ^Y be the ϕ-spline interpolant to $f \in \mathcal{H}_\phi$, on the point set $Y \subset \mathbb{S}^d$, where $h(Y,\mathbb{S}^d) \le 1/(2N)$, for some fixed $N \in \mathbb{Z}_+$. Let Λ be a pseudodifferential operator with symbol $\{\lambda_n\}_{n\ge 0}$ satisfying Assumption* 2.3 *and*

$$\sum_{n\ge 0} d_n\lambda_n^2 a_n < \infty.$$

Then, for $x \in \mathbb{S}^d$,

$$|\Lambda(f - S_\phi^Y)(x)| \le (1+2E)\left(\sum_{n\ge N} d_n\lambda_n^2 a_n\right)^{\frac12} \|f - S_\phi^Y\|_\phi.$$

Proof. Let us choose $\{c_y\}_{y\in Y}$ to be the coefficients described above. Let $v \in \mathcal{H}_\phi$ with $\|v\|_\phi = 1$. Then,

$$\begin{aligned}\inf_{\mu\in Y^*}|\Lambda v(x) - \mu(v)| &\le \left|\sum_{n\ge 0}\left(\Lambda T_n v(x) - \sum_{y\in Y} c_y T_n v(y)\right)\right| \\ &= \left|\sum_{n>N}\left(\lambda_n T_n v(x) - \sum_{y\in Y} c_y T_n v(y)\right)\right|,\end{aligned}$$

by (2.1). Thus,

$$\begin{aligned}\inf_{\mu\in Y^*}|\Lambda v(x) - \mu(v)| &\le \left|\sum_{n>N}\lambda_n T_n v(x)\right| + \left|\sum_{n>N}\sum_{y\in Y} c_y T_n v(y)\right| \\ &\le \sum_{n>N}\left(\lambda_n + \sum_{y\in Y}|c_y|\right)\|T_n v\|_{L_\infty(\mathbb{S}^d)} \\ &\le \sum_{n>N}\left(\lambda_n + \sum_{y\in Y}|c_y|\right)\sqrt{d_n}\|T_n v\|_{L_2(\mathbb{S}^d)},\end{aligned}$$

using Lemma 2.1. Hence, using (2.2) and the Cauchy–Schwarz inequality,

$$\inf_{\mu\in Y^*}|\Lambda v(x) - \mu(v)| \le \left[\left(\sum_{n>N} d_n\lambda_n^2 a_n\right)^{\frac12} + 2E\lambda_N\left(\sum_{n>N} d_n a_n\right)^{\frac12}\right]\|v\|_\phi,$$

and the result follows from Proposition 2.2, since $\|v\|_\phi = 1$ and because $\{\lambda_n\}_{n\ge 0}$ is an increasing sequence. $\square$

Integrating the conclusion of the previous theorem over the sphere, we easily obtain the following result.

COROLLARY 2.6. *Under the hypotheses of Theorem* 2.5,

$$\|\Lambda(f - S_\phi^Y)\|_{L_2(\mathbb{S}^d)} \le (1 + 2E)\left(\sum_{n>N} d_n\lambda_n^2 a_n\right)^{\frac{1}{2}} \|f - S_\phi^Y\|_\phi.$$

Before we give our improved error estimate we need to define a new space $\mathcal{H}_{\Lambda\phi}$ by

$$\mathcal{H}_{\Lambda\phi} := \left\{ f \in \mathcal{H}_\phi : \ \|f\|_{\Lambda\phi} := \left(\sum_{n\ge 0}(\lambda_n a_n)^{-2}\|T_n f\|^2_{L_2(\mathbb{S}^d)}\right)^{\frac{1}{2}} < \infty \right\}.$$

THEOREM 2.7. *Let S_ϕ^Y be the ϕ-spline interpolant to $f \in \mathcal{H}_\phi$ on the point set $Y \subset \mathbb{S}^d$, where $h(Y, \mathbb{S}^d) \le 1/(2N)$, for some fixed $N \in \mathbb{Z}_+$. Let Λ be a pseudodifferential operator with symbol $\{\lambda_n\}_{n\ge 0}$ satisfying Assumption* 2.3 *and*

$$\sum_{n\ge 0} d_n\lambda_n^2 a_n < \infty.$$

Then, for $f \in \mathcal{H}_{\Lambda\phi}$ and for all $x \in \mathbb{S}^d$,

$$|f(x) - S_\phi^Y(x)| \le (1 + 2E)\left(\sum_{n>N} d_n\lambda_n^2 a_n\right)^{\frac{1}{2}}\left(\sum_{n>N} d_n a_n\right)^{\frac{1}{2}} \|f\|_{\Lambda\phi}.$$

Proof. First, using Proposition 1.1 and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\|f - S_\phi^Y\|_\phi^2 &= \langle f - S_\phi^Y, f\rangle_\phi \\
&= \sum_{n\ge 0} a_n^{-1}\langle T_n(f - S_\phi^Y), T_n f\rangle_{L_2(\mathbb{S}^d)} \\
&\le \left(\sum_{n\ge 0}\lambda_n^2\|T_n(f - S_\phi^Y)\|^2_{L_2(\mathbb{S}^d)}\right)^{\frac{1}{2}}\left(\sum_{n\ge 0}(\lambda_n a_n)^{-2}\|T_n f\|^2_{L_2(\mathbb{S}^d)}\right)^{\frac{1}{2}} \\
&= \|\Lambda(f - S_\phi^Y)\|_{L_2(\mathbb{S}^d)}\|f\|_{\Lambda\phi} \\
&\le (1 + 2E)\left(\sum_{n>N} d_n\lambda_n^2 a_n\right)^{\frac{1}{2}}\|f - S_\phi^Y\|_\phi\|f\|_{\Lambda\phi},
\end{aligned}$$

using Corollary 2.6. Cancelling a factor of $\|f - S_\phi^Y\|_\phi$ from both sides yields

$$\|f - S_\phi^Y\|_\phi \le (1 + 2E)\left(\sum_{n>N} d_n\lambda_n^2 a_n\right)^{\frac{1}{2}}\|f\|_{\Lambda\phi}.$$

We can now employ the standard error estimate taken from Jetter, Stöckler, and Ward [10] (our Theorem 2.5 with $\lambda_n = 1$ for all n) to give the required result. ∎

3. The Euclidean case. Our attention now turns to ϕ-spline interpolants of the form

$$S_\phi^Y(x) = \sum_{y \in Y} \alpha_y \phi(|x-y|),$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}$. We will conduct our analysis for positive definite basis functions $\phi \in L_1(\mathbb{R}^d)$ whose Fourier transforms satisfy, for some $s > 0$,

$$C_1(1+|x|)^{-2s} \leq \widehat{\phi}(x) \leq C_2(1+|x|)^{-2s}, \tag{3.1}$$

for some positive constants C_1 and C_2, for example, the Sobolev splines [7] or piecewise polynomial compactly supported radial functions of minimal degree [21]. The exposition contained in this section can be readily adapted to include the polyharmonic splines [8] as well. In that case, the ϕ-spline interpolant must be augmented by a polynomial p with the extra degrees of freedom taken up by the side conditions

$$\sum_{y \in Y} \alpha_y q(y) = 0,$$

where q is polynomial of the same degree (or less) as p.

For a domain $\Omega \subset \mathbb{R}^d$ let $L_2(\Omega)$ denote the usual space of square-integrable functions on Ω with inner product $\langle \cdot\,, \cdot \rangle_{L_2(\Omega)}$ and norm $\|\cdot\|_{L_2(\Omega)}$. For $k \in \mathbb{Z}_+$, the integer-order Sobolev space is defined by

$$\mathcal{H}_k := \left\{ f \in L_2(\mathbb{R}^d) : \; D^\alpha f \in L_2(\mathbb{R}^d) \text{ for all } |\alpha| \leq k \right\},$$

with D^α understood in the distributional sense, which carries the inner product

$$\langle f, g \rangle_k := \langle f, g \rangle_{L_2(\mathbb{R}^d)} + (f, g)_k,$$

where $(f, g)_k$ denotes the Sobolev semi-inner product

$$(f, g)_k := \sum_{|\alpha|=k} c_\alpha^{(k)} \int_{\mathbb{R}^d} (D^\alpha f)(x)(\overline{D^\alpha g})(x)\,\mathrm{d}x,$$

with associated seminorm $|\cdot|_k := (\,\cdot\,, \cdot\,)_k^{1/2}$. The coefficients $c_\alpha^{(k)}$ have been chosen so that

$$\sum_{|\alpha|=k} c_\alpha^{(k)} x^{2\alpha} = |x|^{2k}.$$

We can write the seminorm, using the Fourier transform, in the alternative form

$$|f|_k^2 = \int_{\mathbb{R}^d} |\widehat{f}(x)|^2 |x|^{2k}\,\mathrm{d}x,$$

which facilitates the definition of fractional-order Sobolev space, $\mathcal{H}_s$, for $s > 0$, which has the seminorm

$$|f|_s^2 := \int_{\mathbb{R}^d} |\widehat{f}(x)|^2 |x|^{2s}\,\mathrm{d}x. \tag{3.2}$$

The space $\mathcal{H}_s$ is complete with respect to

$$\|f\|_s := \begin{cases} \left(\|f\|^2_{L_2(\mathbb{R}^d)} + |f|^2_s\right)^{\frac{1}{2}} & \text{if } s \in \mathbb{Z}_+, \\ \left(\|f\|^2_{\lfloor s \rfloor} + |f|^2_s\right)^{\frac{1}{2}} & \text{otherwise,} \end{cases}$$

and, whenever we have $s > d/2$, $\mathcal{H}_s$ is continuously embedded in the continuous functions. The native space for ϕ satisfying (3.1) is equivalent to $\mathcal{H}_s$.

We now wish to make local definitions of our function spaces, which we shall denote by $\mathcal{H}_s(\Omega)$. For $s \in \mathbb{Z}_+$ the definition should be is clear. In what follows we also need the local fractional-order Sobolev spaces:

$$\mathcal{H}_s(\Omega) := \left\{ f \in \mathcal{H}_{\lfloor s \rfloor}(\Omega) : \ \|f\|_{s,\Omega} := \left(\|f\|^2_{\lfloor s \rfloor,\Omega} + |f|^2_{s,\Omega}\right)^{\frac{1}{2}} < \infty \right\},$$

where $|f|_{s,\Omega}$ is the local fractional-order Sobolev seminorm obtained by rewriting (3.2) in an equivalent direct form, i.e., not defined through the Fourier transform of f (see, e.g., Adams [1, p. 214]). For our analysis we find it more useful to exploit an equivalent wavelet representation for the local Sobolev norm [5].

To introduce this equivalent norm we stipulate that the bounded domain, Ω, admits a local multiresolution of closed subspaces $\{V_n(\Omega)\}_{n \geq 0}$ of $L_2(\Omega)$:

$$V_0(\Omega) \subset V_1(\Omega) \subset \cdots \subset L_2(\Omega), \qquad \overline{\bigcup_{n \geq 0} V_n(\Omega)} = L_2(\Omega).$$

Cohen, Dahmen, and DeVore [5] give sufficient conditions on Ω to admit such a local multiresolution. In particular, for $d = 2$, those domains whose boundaries have certain piecewise Lipschitz smoothness are admissible. The following is an incidence of [5, Theorem 4.2].

THEOREM 3.1. *Suppose that $\Omega \subset \mathbb{R}^d$ is a bounded domain that admits a local multiresolution $\{V_n(\Omega)\}_{n \geq 0}$ for $L_2(\Omega)$. For $n \geq 0$, let Q_n^Ω denote the orthogonal projection from $L_2(\Omega)$ onto $W_n(\Omega) = V_n(\Omega) \ominus V_{n-1}(\Omega)$ with the convention that $V_{-1}(\Omega) = \{0\}$. For each $s \geq 0$, let Λ_s be the pseudodifferential operator on Ω defined via*

$$\Lambda_s := \sum_{n \geq 0} 2^{ns} Q_n^\Omega.$$

Then there exist positive constants C_1 and C_2 such that, for all $f \in \mathcal{H}_s(\Omega)$,

$$C_1 \|\Lambda_s f\|_{L_2(\Omega)} \leq \|f\|_{s,\Omega} \leq C_2 \|\Lambda_s f\|_{L_2(\Omega)}.$$

Now, let us return to the task at hand. For ϕ satisfying (3.1), we will denote the ϕ-spline interpolant on the point set Y by S_ϕ^Y. The standard error estimate in this context is

$$|f(x) - S_\phi^Y(x)| \leq C h^{s-d/2} \|f - S_\phi^Y\|_s; \tag{3.3}$$

see [22]. If f is smoother (and satisfies some boundary conditions), we can get a better rate of convergence.

We will exploit the fact that $\mathcal{H}_\mu(\Omega)$, for $0 < \mu < s$, is an interpolation space lying between $L_2(\Omega)$ and $\mathcal{H}_s(\Omega)$ (see Bergh and Löfström [2, p. 131]). We can then use the standard interpolation theorem concerning the norms of operators bounded on the extreme spaces to infer a bound on the norm for the interpolation space. For further information on interpolation spaces the reader can consult, e.g., Bergh and Löfström [2]. We use the following interpolation theorem.

PROPOSITION 3.2. *Let $0 < \mu < s$. Further suppose that $T : \mathcal{H}_s(\Omega) \to L_2(\Omega)$ and that $T : \mathcal{H}_s(\Omega) \to \mathcal{H}_s(\Omega)$ is a bounded operator. Then,*

$$\|T\|_{\mathcal{H}_s(\Omega)\to\mathcal{H}_\mu(\Omega)} \le \|T\|^{1-\mu/s}_{\mathcal{H}_s(\Omega)\to L_2(\Omega)} \|T\|^{\mu/s}_{\mathcal{H}_s(\Omega)\to\mathcal{H}_s(\Omega)}.$$

Since we can write the $\mathcal{H}_s$-norm in entirely direct form, we are at liberty to utilize Duchon's localization technique [8] to enhance the standard error estimate (3.3). Therefore, if Ω is bounded and satisfies an interior cone condition, then, for $f \in \mathcal{H}_s(\Omega)$, $s > d/2$, and sufficiently small $h = h(Y, \Omega)$,

$$\|f - S^Y_\phi\|_{L_2(\Omega)} \le Ch^s \|f - S^Y_\phi\|_{s,\Omega}.$$

Writing $Tf = f - S^Y_\phi$, we see, using the last proposition, that, for $0 < \mu < s$,

$$\begin{aligned} \|f - S^Y_\phi\|_{\mu,\Omega} &\le (Ch^s \|f - S^Y_\phi\|_{s,\Omega})^{1-\mu/s} \|f - S^Y_\phi\|^{\mu/s}_{s,\Omega} \\ &= Ch^{s-\mu} \|f - S^Y_\phi\|_{s,\Omega}. \end{aligned} \tag{3.4}$$

We can now prove our main result of this section, which is a generalization of that of Schaback [19].

THEOREM 3.3. *Suppose that $\Omega \subset \mathbb{R}^d$ is bounded, satisfies an interior cone condition, and admits a local multiresolution. Let $s > d/2$, and let S^Y_ϕ be the ϕ-spline interpolant to $f \in \mathcal{H}_s$ on the point set $Y \subset \Omega$. Suppose further that $f \in \mathcal{H}_\nu$, for $s < \nu \le 2s$, and that f is compactly supported in Ω. Then there exists $C > 0$, independent of f and $h = h(Y, \Omega)$, such that for all $x \in \Omega$ and sufficiently small h,*

$$|f(x) - S^Y_\phi(x)| \le Ch^{\nu-d/2} \|f\|_{\nu,\Omega}.$$

Proof. From Proposition 1.1 we know that

$$\langle f - S^Y_\phi, S^Y_\phi\rangle_s = 0,$$

so that

$$\|f - S^Y_\phi\|^2_s = \langle f - S^Y_\phi, f\rangle_s \le C\langle f - S^Y_\phi, f\rangle_{s,\Omega},$$

where we have used the compact support of f in Ω. Now, the equivalent norm from Theorem 3.1 gives us

$$\begin{aligned} \|f - S^Y_\phi\|^2_s &\le C\langle \Lambda_s(f - S^Y_\phi), \Lambda_s f\rangle_{L_2(\Omega)} \\ &= C\sum_{n\ge 0} 4^{ns} \langle Q^\Omega_n(f - S^Y_\phi), Q^\Omega_n f\rangle_{L_2(\Omega)}, \end{aligned}$$

and successive applications of the continuous and discrete Cauchy–Schwarz inequality yield

$$\begin{aligned}
\|f - S_\phi^Y\|_s^2 &\le C \sum_{n\ge 0} \|2^{n(2s-\nu)} Q_n^\Omega (f - S_\phi^Y)\|_{L_2(\Omega)} \|2^{n\nu} Q_n^\Omega f\|_{L_2(\Omega)} \\
&\le C \left(\sum_{n\ge 0} \|2^{n(2s-\nu)} Q_n^\Omega (f - S_\phi^Y)\|_{L_2(\Omega)}^2 \right)^{\frac{1}{2}} \left(\sum_{n\ge 0} \|2^{n\nu} Q_n^\Omega f\|_{L_2(\Omega)}^2 \right)^{\frac{1}{2}} \\
&= C \|\Lambda_{2s-\nu}(f - S_\phi^Y)\|_{L_2(\Omega)} \|\Lambda_\nu f\|_{L_2(\Omega)}.
\end{aligned}$$

Thus, using the norm equivalence from Theorem 3.1 again together with (3.4), we have

$$\begin{aligned}
\|f - S_\phi^Y\|_s^2 &\le C \|f - S_\phi^Y\|_{2s-\nu,\Omega} \|f\|_{\nu,\Omega} \\
&\le C h^{\nu - s} \|f - S_\phi^Y\|_{s,\Omega} \|f\|_{\nu,\Omega} \\
&\le C h^{\nu - s} \|f - S_\phi^Y\|_s \|f\|_{\nu,\Omega},
\end{aligned}$$

and cancelling one power of $\|f - S_\phi^Y\|_s$ gives

$$\|f - S_\phi^Y\|_s \le C h^{\nu - s} \|f\|_{\nu,\Omega}.$$

The result follows by substitution into the standard error estimate (3.3). □

Acknowledgments. We are grateful to Professor Mikhail Shubin for useful input into questions of pseudodifferential operators and alternative Sobolev space seminorms.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Pure Appl. Math. 65, Academic Press, New York, 1975.

[2] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Grundl. Math. Wiss. 223, Springer-Verlag, Berlin, 1976.

[3] R. A. Brownlee and W. Light, *Approximation orders for interpolation by surface splines to rough functions*, IMA J. Numer. Anal., 24 (2004), pp. 179–192.

[4] M. D. Buhmann, *Radial Basis Functions: Theory and Implementations*, Cambridge Monogr. Appl. Comput. Math. 12, Cambridge University Press, Cambridge, UK, 2003.

[5] A. Cohen, W. Dahmen, and R. DeVore, *Multiscale decompositions on bounded domains*, Trans. Amer. Math. Soc., 352 (2000), pp. 3651–3685.

[6] Z. Ditzian, *Fractional derivatives and best approximation*, Acta Math. Hungar., 81 (1998), pp. 323–348.

[7] J. G. Dix and R. D. Ogden, *An interpolation scheme with radial basis in Sobolev spaces* $H^s(\mathbf{R}^n)$, Rocky Mountain J. Math., 24 (1994), pp. 1319–1337.

[8] J. Duchon, *Sur l'erreur d'interpolation des fonctions de plusieurs variables par les* D^m*-splines*, RAIRO Anal. Numér., 12 (1978), pp. 325–334.

[9] W. Freeden, T. Gervens, and M. Schreiner, *Constructive Approximation on the Sphere (with Applications to Geomathematics)*, Oxford University Press, New York, 1998.

[10] K. Jetter, J. Stöckler, and J. D. Ward, *Error estimates for scattered data interpolation on spheres*, Math. Comp., 68 (1999), pp. 733–747.

[11] J. Levesley, C. Odell, and D. L. Ragozin, *Scattered data on homogeneous manifolds: The norming set approach*, in Curve and Surface Fitting (Saint-Malo, 2002), J.-L. Merrien, A. Cohen, and L. L. Schumaker, eds., Modern Methods in Math., Nashboro Press, Brentwood, TN, 2003, pp. 269–278.

[12] J. Levesley and X. Sun, *Approximation in rough native spaces by shifts of smooth kernels on spheres*, J. Approx. Theory, 133 (2005), pp. 269–283.
[13] W. A. Light and H. Wayne, *On power functions and error estimates for radial basis function interpolation*, J. Approx. Theory, 92 (1998), pp. 245–266.
[14] T. M. Morton and M. Neamtu, *Error bounds for solving pseudodifferential equations on spheres by collocation with zonal kernels*, J. Approx. Theory, 114 (2002), pp. 242–268.
[15] C. Müller, *Spherical Harmonics*, Lecture Notes in Math. 17, Springer-Verlag, Berlin, 1966.
[16] F. J. Narcowich, R. Schaback, and J. D. Ward, *Approximations in Sobolev spaces by kernel expansions*, J. Approx. Theory, 114 (2002), pp. 70–83.
[17] F. J. Narcowich and J. D. Ward, *Scattered data interpolation on spheres: Error estimates and locally supported basis functions*, SIAM J. Math. Anal., 33 (2002), pp. 1393–1410.
[18] F. J. Narcowich and J. D. Ward, *Scattered-data interpolation on* $\mathbb{R}^n$*: Error estimates for radial basis and band-limited functions*, SIAM J. Math. Anal., 36 (2004), pp. 284–300.
[19] R. Schaback, *Improved error bounds for scattered data interpolation by radial basis functions*, Math. Comp., 68 (1999), pp. 201–216.
[20] S. L. Svensson, *Pseudodifferential operators—A new approach to the boundary problems of physical geodesy*, Manuscripta Geod., 8 (1983), pp. 1–40.
[21] H. Wendland, *Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree*, Adv. Comput. Math., 4 (1995), pp. 389–396.
[22] Z. M. Wu and R. Schaback, *Local error estimates for radial basis function interpolation of scattered data*, IMA J. Numer. Anal., 13 (1993), pp. 13–27.
[23] J. Yoon, *Interpolation by radial basis functions on Sobolev space*, J. Approx. Theory, 112 (2001), pp. 1–15.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 565–592

STABLE SOLUTIONS FOR THE BILAPLACIAN WITH EXPONENTIAL NONLINEARITY*

JUAN DÁVILA†, LOUIS DUPAIGNE‡, IGNACIO GUERRA§, AND MARCELO MONTENEGRO¶

Abstract. Let $\lambda^* > 0$ denote the largest possible value of λ such that $\{\Delta^2 u = \lambda e^u$ in B, $u = \frac{\partial u}{\partial n} = 0$ on $\partial B\}$ has a solution, where B is the unit ball in $\mathbb{R}^N$ and n is the exterior unit normal vector. We show that for $\lambda = \lambda^*$ this problem possesses a unique *weak* solution u^*. We prove that u^* is smooth if $N \leq 12$ and singular when $N \geq 13$, in which case $u^*(r) = -4\log r + \log(8(N-2)(N-4)/\lambda^*) + o(1)$ as $r \to 0$. We also consider the problem with general constant Dirichlet boundary conditions.

Key words. biharmonic, singular solutions, stability

AMS subject classifications. Primary, 35J65; Secondary, 35J40

DOI. 10.1137/060665579

1. Introduction.

We study the fourth order problem

$$\begin{cases} \Delta^2 u = \lambda e^u & \text{in } B, \\ u = a & \text{on } \partial B, \\ \dfrac{\partial u}{\partial n} = b & \text{on } \partial B, \end{cases} \tag{1}$$

where a, $b \in \mathbb{R}$, B is the unit ball in $\mathbb{R}^N$, $N \geq 1$, n is the exterior unit normal vector, and $\lambda \geq 0$ is a parameter.

Recently higher order equations have attracted the interest of many researchers. In particular, fourth order equations with an exponential nonlinearity have been studied in four dimensions in a setting analogous to Liouville's equation in [3, 12, 24] and in higher dimensions by [1, 2, 4, 5, 13].

We shall pay special attention to (1) in the case $a = b = 0$, as it is the natural fourth order analogue of the classical Gelfand problem

$$\begin{cases} -\Delta u = \lambda e^u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega \end{cases} \tag{2}$$

*Received by the editors July 20, 2006; accepted for publication (in revised form) January 9, 2007; published electronically July 27, 2007.

http://www.siam.org/journals/sima/39-2/66557.html

†Departamento de Ingeniería Matemática and CMM, Universidad de Chile, Casilla 170 Correo 3, Santiago, Chile (jdavila@dim.uchile.cl). This author has been partially supported by grants FONDECYT 1050725 and FONDAP Chile. Part of this research was done when this author was visiting Laboratoire Amiénois de Mathématique Fondamentale et Appliquée at Faculté de Mathématique et d'Informatique, Amiens, France.

‡LAMFA CNRS UMR 6140, Université de Picardie Jules Verne, 33 rue Saint-Leu 80039, Amiens Cedex 1, France (louis.dupaigne@u-picardie.fr). Part of this research was done when this author was visiting the Departamento de Ingeniería Matemática at Universidad de Chile, Santiago, Chile.

§Departamento de Matemática y C.C., Facultad de Ciencia, Universidad de Santiago de Chile, Casilla 307, Correo 2, Santiago, Chile (iguerra@usach.cl). This author was supported by FONDECYT 1061110.

¶Universidade Estadual de Campinas, IMECC, Departamento de Matemática, Caixa Postal 6065, CEP 13083-970, Campinas, SP, Brazil (msm@ime.unicamp.br). This author was partially supported by project CNPq-CNRS 490093/2004-3.

(Ω is a smooth bounded domain in $\mathbb{R}^N$) for which a vast literature exists [7, 8, 9, 10, 18, 19, 20, 21].

From the technical point of view, one of the basic tools in the analysis of (2) is the maximum principle. As pointed out in [2], in general domains the maximum principle for Δ^2 with Dirichlet boundary condition is not valid anymore. One of the reasons to study (1) in a ball is that a maximum principle holds in this situation; see [6]. In this simpler setting, though there are some similarities between the two problems, several tools that are well suited for (2) no longer seem to work for (1).

As a start, let us introduce the class of weak solutions we shall be working with: we say that $u \in H^2(B)$ is a weak solution to (1) if $e^u \in L^1(B)$, $u = a$ on ∂B, $\frac{\partial u}{\partial n} = b$ on ∂B, and

$$\int_B \Delta u \Delta \varphi = \lambda \int_B e^u \varphi \quad \forall \varphi \in C_0^\infty(B).$$

The following basic result is a straightforward adaptation of Theorem 3 in [2].

THEOREM 1.1 (see [2]). *There exists λ^* such that if $0 \le \lambda < \lambda^*$ then* (1) *has a minimal smooth solution u_λ and if $\lambda > \lambda^*$ then* (1) *has no weak solution.*

The limit $u^* = \lim_{\lambda \nearrow \lambda^*} u_\lambda$ exists pointwise, belongs to $H^2(B)$, and is a weak solution to (1). *It is called the extremal solution.*

The functions u_λ, $0 \le \lambda < \lambda^*$, and u^* are radially symmetric and radially decreasing.

The branch of minimal solutions of (1) has an important property; namely, u_λ is stable in the sense that

$$\int_B (\Delta \varphi)^2 \ge \lambda \int_B e^{u_\lambda} \varphi^2 \quad \forall \varphi \in C_0^\infty(B); \tag{3}$$

see [2, Proposition 37].

The authors in [2] pose several questions, some of which we address in this work. First we show that the extremal solution u^* is the unique solution to (1) in the class of weak solutions. Actually the statement is stronger, asserting that for $\lambda = \lambda^*$ there are no strict supersolutions.

THEOREM 1.2. *If*

$$v \in H^2(B),\ e^v \in L^1(B),\ v|_{\partial B} = a,\ \tfrac{\partial v}{\partial n}|_{\partial B} \le b, \tag{4}$$

and

$$\int_B \Delta v \Delta \varphi \ge \lambda^* \int_B e^v \varphi \quad \forall\, \varphi \in C_0^\infty(B),\ \varphi \ge 0, \tag{5}$$

then $v = u^*$. In particular, for $\lambda = \lambda^*$ problem (1) *has a unique weak solution.*

This result is analogous to the work of Martel [19] for more general versions of (2), where the exponential function is replaced by a positive, increasing, convex, and superlinear function.

Next, we discuss the regularity of the extremal solution u^*. In dimensions $N = 5, \ldots, 16$ the authors of [2] find, with a computer assisted proof, a radial singular solution U_σ to (1) with $a = b = 0$ associated to a parameter $\lambda_\sigma > 8(N-2)(N-4)$. They show that $\lambda_\sigma < \lambda^*$ if $N \le 10$ and claim to have numerical evidence that this holds for $N \le 12$. They leave open the question of whether u^* is singular in dimension $N \le 12$. We prove the following theorem.

THEOREM 1.3. *If $N \le 12$ then the extremal solution u^* of* (1) *is smooth.*

The method introduced in [10, 20] to prove the boundedness of u^* in low dimensions for (2) seems not useful for (1), thus requiring a new strategy. A first indication that the borderline dimension for the boundedness of u^* is 12 is Rellich's inequality [23], which states that if $N \ge 5$ then

$$\int_{\mathbb{R}^N} (\Delta\varphi)^2 \ge \frac{N^2(N-4)^2}{16} \int_{\mathbb{R}^N} \frac{\varphi^2}{|x|^4} \qquad \forall \varphi \in C_0^\infty(\mathbb{R}^N), \tag{6}$$

where the constant $N^2(N-4)^2/16$ is known to be optimal. The proof of Theorem 1.3 is based on the observation that if u^* is singular then $\lambda^* e^{u^*} \sim 8(N-2)(N-4)|x|^{-4}$ near the origin. But $8(N-2)(N-4) > N^2(N-4)^2/16$ if $N \le 12$, which would contradict the stability condition (3).

In view of Theorem 1.3, it is natural to ask whether u^* is singular in dimension $N \ge 13$. If $a = b = 0$, we prove the following theorem.

THEOREM 1.4. *Let $N \ge 13$ and $a = b = 0$. Then the extremal solution u^* to* (1) *is unbounded.*

For general boundary values, it seems more difficult to determine the dimensions for which the extremal solution is singular. We observe first that given any $a, b \in \mathbb{R}$, u^* is the extremal solution of (1) if and only if $u^* - a$ is the extremal solution of the same equation with boundary condition $u = 0$ on ∂B. In particular, if $\lambda^*(a, b)$ denotes the extremal parameter for problem (1), one has that $\lambda^*(a, b) = e^{-a}\lambda^*(0, b)$. So the value of a is irrelevant. But one may ask if Theorem 1.4 still holds for any $N \ge 13$ and any $b \in \mathbb{R}$. The situation turns out to be somewhat more complicated.

PROPOSITION 1.5.

(a) *Fix $N \ge 13$ and take any $a \in \mathbb{R}$. Assume $b \ge -4$. There exists a critical parameter $b^{max} > 0$, depending only on N, such that the extremal solution u^* is singular if and only if $b \le b^{max}$.*

(b) *Fix $b \ge -4$ and take any $a \in \mathbb{R}$. There exists a critical dimension $N^{min} \ge 13$, depending only on b, such that the extremal solution u^* to* (1) *is singular if $N \ge N^{min}$.*

Remark 1.6.

- We have not investigated the case $b < -4$.
- If follows from item (a) that for $b \in [-4, 0]$, the extremal solution is singular if and only if $N \ge 13$.
- It also follows from item (a) that there exist values of b for which $N^{min} > 13$. We do not know whether u^* remains bounded for $13 \le N < N^{min}$.

Our proof of Theorem 1.4 is related to an idea that Brezis and Vázquez applied to the Gelfand problem and is based on a characterization of singular *energy* solutions through linearized stability (see Theorem 3.1 in [8]). In our context we show the following.

PROPOSITION 1.7. *Assume that $u \in H^2(B)$ is an unbounded weak solution of* (1) *satisfying the stability condition*

$$\lambda \int_B e^u \varphi^2 \le \int_B (\Delta\varphi)^2 \quad \forall \varphi \in C_0^\infty(B). \tag{7}$$

Then $\lambda = \lambda^*$ and $u = u^*$.

In the proof of Theorem 1.4 we do not use Proposition 1.7 directly but some variants of it—see Lemma 2.6 and Remark 2.7—because we do not have at our disposal an explicit solution to (1). Instead, we show that it is enough to find a sufficiently good approximation to u^*. When $N \geq 32$ we are able to construct such an approximation by hand. However, for $13 \leq N \leq 31$ we resort to a computer assisted generation and verification.

Only in very few situations may one take advantage of Proposition 1.7 directly. For instance, for problem (1) with $a = 0$ and $b = -4$ we have an explicit solution

$$\bar{u}(x) = -4 \log |x|$$

associated to $\bar{\lambda} = 8(N-2)(N-4)$. Thanks to Rellich's inequality (6) the solution $\bar{u}$ satisfies condition (7) when $N \geq 13$. Therefore, by Theorem 1.3 and a direct application of Proposition 1.7 we obtain Theorem 1.4 in the case $b = -4$.

In [2] the authors say that a radial weak solution u to (1) is weakly singular if

$$\lim_{r \to 0} r u'(r) \text{ exists.}$$

For example, the singular solutions U_σ of [2] verify this condition.

As a corollary of Theorem 1.2 we show the following.

PROPOSITION 1.8. *The extremal solution u^* to* (1) *with $b \geq -4$ is always weakly singular.*

A weakly singular solution either is smooth or exhibits a log-type singularity at the origin. More precisely, if u is a nonsmooth weakly singular solution of (1) with parameter λ, then (see [2])

$$\begin{aligned} &\lim_{r \to 0} u(r) + 4 \log r = \log \frac{8(N-2)(N-4)}{\lambda}, \\ &\lim_{r \to 0} r u'(r) = -4. \end{aligned}$$

In section 2 we describe the comparison principles we use later. Section 3 is devoted to the proof of the uniqueness of u^* and Propositions 1.7 and 1.8. We prove Theorem 1.3, the boundedness of u^* in low dimensions, in section 4. The argument for Theorem 1.4 is contained in section 5 for the case $N \geq 32$ and section 6 for $13 \leq N \leq 31$. In section 7 we give the proof of Proposition 1.5.

Notation.

- B_R is the ball of radius R in $\mathbb{R}^N$ centered at the origin. $B = B_1$.
- n is the exterior unit normal vector to B_R.
- All inequalities or equalities for functions in L^p spaces are understood to be a.e.

2. Comparison principles.

LEMMA 2.1 (Boggio's principle [6]). *If $u \in C^4(\overline{B}_R)$ satisfies*

$$\begin{cases} \Delta^2 u \geq 0 & \text{in } B_R, \\ u = \dfrac{\partial u}{\partial n} = 0 & \text{on } \partial B_R, \end{cases}$$

then $u \geq 0$ in B_R.

LEMMA 2.2. *Let $u \in L^1(B_R)$ and suppose that*

$$\int_{B_R} u \Delta^2 \varphi \geq 0$$

for all $\varphi \in C^4(\overline{B}_R)$ such that $\varphi \geq 0$ in B_R, $\varphi|_{\partial B_R} = 0 = \frac{\partial \varphi}{\partial n}|_{\partial B_R}$. Then $u \geq 0$ in B_R. Moreover, $u \equiv 0$ or $u > 0$ a.e. in B_R.

For a proof see Lemma 17 in [2].

LEMMA 2.3. *If $u \in H^2(B_R)$ is radial, $\Delta^2 u \geq 0$ in B_R in the weak sense, that is,*

$$\int_{B_R} \Delta u \Delta \varphi \geq 0 \quad \forall \varphi \in C_0^\infty(B_R),\ \varphi \geq 0,$$

and $u|_{\partial B_R} \geq 0$, $\frac{\partial u}{\partial n}|_{\partial B_R} \leq 0$, then $u \geq 0$ in B_R.

Proof. We deal only with the case $R = 1$ for simplicity. Solve

$$\begin{cases} \Delta^2 u_1 = \Delta^2 u & \text{in } B_1, \\ u_1 = \dfrac{\partial u_1}{\partial n} = 0 & \text{on } \partial B_1 \end{cases}$$

in the sense $u_1 \in H_0^2(B_1)$ and $\int_{B_1} \Delta u_1 \Delta \varphi = \int_{B_1} \Delta u \Delta \varphi$ for all $\varphi \in C_0^\infty(B_1)$. Then $u_1 \geq 0$ in B_1 by Lemma 2.2.

Let $u_2 = u - u_1$ so that $\Delta^2 u_2 = 0$ in B_1. Define $f = \Delta u_2$. Then $\Delta f = 0$ in B_1 and since f is radial we find that f is constant. It follows that $u_2 = ar^2 + b$. Using the boundary conditions we deduce $a + b \geq 0$ and $a \leq 0$, which imply $u_2 \geq 0$. ☐

Similarly, we have the following lemma.

LEMMA 2.4. *If $u \in H^2(B_R)$ and $\Delta^2 u \geq 0$ in B_R in the weak sense, that is,*

$$\int_{B_R} \Delta u \Delta \varphi \geq 0 \quad \forall \varphi \in C_0^\infty(B_R),\ \varphi \geq 0,$$

and $u|_{\partial B_R} = 0$, $\frac{\partial u}{\partial n}|_{\partial B_R} \leq 0$, then $u \geq 0$ in B_R.

The next lemma is a consequence of a decomposition lemma of Moreau [22]. For a proof see [14, 15].

LEMMA 2.5. *Let $u \in H_0^2(B_R)$. Then there exist unique $w, v \in H_0^2(B_R)$ such that $u = w + v$, $w \geq 0$, $\Delta^2 v \leq 0$ in B_R and $\int_{B_R} \Delta w \Delta v = 0$.*

We need the following comparison principle.

LEMMA 2.6. *Let u_1, $u_2 \in H^2(B_R)$ with e^{u_1}, $e^{u_2} \in L^1(B_R)$. Assume that*

$$\Delta^2 u_1 \leq \lambda e^{u_1} \quad \text{in } B_R$$

in the sense that

$$\int_{B_R} \Delta u_1 \Delta \varphi \leq \lambda \int_{B_R} e^{u_1} \varphi \quad \forall \varphi \in C_0^\infty(B_R),\ \varphi \geq 0, \tag{8}$$

and $\Delta^2 u_2 \geq \lambda e^{u_2}$ in B_R in the similar weak sense. Suppose also

$$u_1|_{\partial B_R} = u_2|_{\partial B_R} \quad \text{and} \quad \frac{\partial u_1}{\partial n}|_{\partial B_R} = \frac{\partial u_2}{\partial n}|_{\partial B_R}.$$

Assume, furthermore, that u_1 is stable in the sense that

$$\lambda \int_{B_R} e^{u_1} \varphi^2 \leq \int_{B_R} (\Delta \varphi)^2 \quad \forall \varphi \in C_0^\infty(B_R). \tag{9}$$

Then

$$u_1 \leq u_2 \quad \text{in } B_R.$$

Proof. Let $u = u_1 - u_2$. By Lemma 2.5 there exist $w, v \in H_0^2(B_R)$ such that $u = w + v$, $w \geq 0$ and $\Delta^2 v \leq 0$. Observe that $v \leq 0$ so $w \geq u_1 - u_2$.

By hypothesis we have for all $\varphi \in C_0^\infty(B_R)$, $\varphi \geq 0$,

$$\int_{B_R} \Delta(u_1 - u_2)\Delta\varphi \leq \lambda \int_{B_R} (e^{u_1} - e^{u_2})\varphi \leq \lambda \int_{B_R \cap [u_1 \geq u_2]} (e^{u_1} - e^{u_2})\varphi$$

and by density this holds also for w:

$$\begin{aligned} \int_{B_R} (\Delta w)^2 &= \int_{B_R} \Delta(u_1 - u_2)\Delta w \\ &\leq \lambda \int_{B_R \cap [u_1 \geq u_2]} (e^{u_1} - e^{u_2}) w = \lambda \int_{B_R} (e^{u_1} - e^{u_2}) w, \end{aligned} \tag{10}$$

where the first equality holds because $\int_{B_R} \Delta w \Delta v = 0$. By density we deduce from (9)

$$\lambda \int_{B_R} e^{u_1} w^2 \leq \int_{B_R} (\Delta w)^2. \tag{11}$$

Combining (10) and (11), we obtain

$$\int_{B_R} e^{u_1} w^2 \leq \int_{B_R} (e^{u_1} - e^{u_2}) w.$$

Since $u_1 - u_2 \leq w$ the previous inequality implies

$$0 \leq \int_{B_R} (e^{u_1} - e^{u_2} - e^{u_1}(u_1 - u_2)) w. \tag{12}$$

But by convexity of the exponential function $e^{u_1} - e^{u_2} - e^{u_1}(u_1 - u_2) \leq 0$, and we deduce from (12) that $(e^{u_1} - e^{u_2} - e^{u_1}(u_1 - u_2))w = 0$. Recalling that $u_1 - u_2 \leq w$ we deduce that $u_1 \leq u_2$. ◻

Remark 2.7. The following variant of Lemma 2.6 also holds.

Let u_1, $u_2 \in H^2(B_R)$ be *radial* with e^{u_1}, $e^{u_2} \in L^1(B_R)$. Assume $\Delta^2 u_1 \leq \lambda e^{u_1}$ in B_R in the sense of (8) and $\Delta^2 u_2 \geq \lambda e^{u_2}$ in B_R. Suppose $u_1|_{\partial B_R} \leq u_2|_{\partial B_R}$ and $\frac{\partial u_1}{\partial n}|_{\partial B_R} \geq \frac{\partial u_2}{\partial n}|_{\partial B_R}$ and that the stability condition (9) holds. Then $u_1 \leq u_2$ in B_R.

Proof. We solve for $\tilde{u} \in H_0^2(B_R)$ such that

$$\int_{B_R} \Delta\tilde{u}\Delta\varphi = \int_{B_R} \Delta(u_1 - u_2)\Delta\varphi \quad \forall \varphi \in C_0^\infty(B_R).$$

By Lemma 2.3 it follows that $\tilde{u} \geq u_1 - u_2$. Next we apply the decomposition of Lemma 2.5 to $\tilde{u}$, that is, $\tilde{u} = w + v$ with $w, v \in H_0^2(B_R)$, $w \geq 0$, $\Delta^2 v \leq 0$ in B_R, and $\int_{B_R} \Delta w \Delta v = 0$. Then the argument follows that of Lemma 2.6. ◻

Finally, in several places we will need the method of sub- and supersolutions in the context of weak solutions.

LEMMA 2.8. *Let $\lambda > 0$ and assume that there exists $\bar{u} \in H^2(B_R)$ such that $e^{\bar{u}} \in L^1(B_R)$,*

$$\int_{B_R} \Delta\bar{u}\Delta\varphi \geq \lambda \int_{B_R} e^{\bar{u}}\varphi \quad \forall \varphi \in C_0^\infty(B_R), \varphi \geq 0,$$

and

$$\bar{u} = a, \quad \frac{\partial \bar{u}}{\partial n} \le b \quad \text{on } \partial B_1.$$

Then there exists a weak solution to (1) *such that* $u \le \bar{u}$.

The proof is similar to that of Lemma 19 in [2].

3. Uniqueness of the extremal solution: Proof of Theorem 1.2.

Proof of Theorem 1.2. Suppose that $v \in H^2(B)$ satisfies (4), (5), and $v \not\equiv u^*$. Notice that we do not need v to be radial. The idea of the proof is as follows.

Step 1. The function

$$u_0 = \frac{1}{2}(u^* + v)$$

is a supersolution to the problem

$$\begin{cases} \Delta^2 u = \lambda^* e^u + \mu \eta e^u & \text{in } B, \\ u = a & \text{on } \partial B, \\ \dfrac{\partial u}{\partial n} = b & \text{on } \partial B \end{cases} \tag{13}$$

for some $\mu = \mu_0 > 0$, where $\eta \in C_0^\infty(B)$, $0 \le \eta \le 1$, is a fixed radial cut-off function such that

$$\eta(x) = 1 \quad \text{for } |x| \le \tfrac{1}{2}, \qquad \eta(x) = 0 \quad \text{for } |x| \ge \tfrac{3}{4}.$$

Step 2. Using a solution to (13) we construct, for some $\lambda > \lambda^*$, a supersolution to (1). This provides a solution u_λ for some $\lambda > \lambda^*$, which is a contradiction.

Proof of Step 1. Observe that given $0 < R < 1$ we must have for some $c_0 = c_0(R) > 0$

$$v(x) \ge u^*(x) + c_0, \quad |x| \le R. \tag{14}$$

To prove this we recall the Green's function for Δ^2 with Dirichlet boundary conditions

$$\begin{cases} \Delta_x^2 G(x,y) = \delta_y, & x \in B, \\ G(x,y) = 0, & x \in \partial B, \\ \dfrac{\partial G}{\partial n}(x,y) = 0, & x \in \partial B, \end{cases}$$

where δ_y is the Dirac mass at $y \in B$. Boggio gave an explicit formula for $G(x,y)$ which was used in [16] to prove that in dimension $N \ge 5$ (the case $1 \le N \le 4$ can be treated similarly)

$$G(x,y) \sim |x-y|^{4-N} \min\left(1, \frac{d(x)^2 d(y)^2}{|x-y|^4}\right), \tag{15}$$

where

$$d(x) = \operatorname{dist}(x, \partial B) = 1 - |x|$$

and $a \sim b$ means that for some constant $C > 0$ we have $C^{-1}a \le b \le Ca$ (uniformly for $x, y \in B$). Formula (15) yields

$$G(x,y) \ge cd(x)^2 d(y)^2 \tag{16}$$

for some $c > 0$ and this in turn implies that for smooth functions $\tilde v$ and $\tilde u$ such that $\tilde v - \tilde u \in H_0^2(B)$ and $\Delta^2(\tilde v - \tilde u) \ge 0$,

$$\begin{aligned}\tilde v(y) - \tilde u(y) &= \int_{\partial B} \Big(\frac{\partial \Delta_x G}{\partial n_x}(x,y)(\tilde v - \tilde u) - \Delta_x G(x,y) \frac{\partial(\tilde v - \tilde u)}{\partial n} \Big)\, dx \\ &\quad + \int_B G(x,y) \Delta^2(\tilde v - \tilde u)\, dx \\ &\ge cd(y)^2 \int_B (\Delta^2 \tilde v - \Delta^2 \tilde u) d(x)^2\, dx.\end{aligned}$$

Using a standard approximation procedure, we conclude that

$$v(y) - u^*(y) \ge cd(y)^2 \lambda^* \int_B (e^v - e^{u^*}) d(x)^2\, dx.$$

Since $v \ge u^*$, $v \not\equiv u^*$ we deduce (14).

Let $u_0 = (u^* + v)/2$. Then by Taylor's theorem

$$e^v = e^{u_0} + (v - u_0)e^{u_0} + \frac{1}{2}(v-u_0)^2 e^{u_0} + \frac{1}{6}(v - u_0)^3 e^{u_0} + \frac{1}{24}(v-u_0)^4 e^{\xi_2} \tag{17}$$

for some $u_0 \le \xi_2 \le v$ and

$$e^{u^*} = e^{u_0} + (u^* - u_0)e^{u_0} + \frac{1}{2}(u^*-u_0)^2 e^{u_0} + \frac{1}{6}(u^* - u_0)^3 e^{u_0} + \frac{1}{24}(u^*-u_0)^4 e^{\xi_1} \tag{18}$$

for some $u^* \le \xi_1 \le u_0$. Adding (17) and (18) yields

$$\frac{1}{2}(e^v + e^{u^*}) \ge e^{u_0} + \frac{1}{8}(v - u^*)^2 e^{u_0}. \tag{19}$$

From (14) with $R = 3/4$ and (19) we see that $u_0 = (u^* + v)/2$ is a supersolution of (13) with $\mu_0 := c_0/8$.

Proof of Step 2. Let us show now how to obtain a weak supersolution of (1) for some $\lambda > \lambda^*$. Given $\mu > 0$, let u denote the minimal solution to (13). Define φ_1 as the solution to

$$\begin{cases} \Delta^2 \varphi_1 = \mu \eta e^u & \text{in } B, \\ \varphi_1 = 0 & \text{on } \partial B, \\ \dfrac{\partial \varphi_1}{\partial n} = 0 & \text{on } \partial B, \end{cases}$$

and φ_2 as the solution to

$$\begin{cases} \Delta^2 \varphi_2 = 0 & \text{in } B, \\ \varphi_2 = a & \text{on } \partial B, \\ \dfrac{\partial \varphi_2}{\partial n} = b & \text{on } \partial B. \end{cases}$$

If $N \geq 5$ (the case $1 \leq N \leq 4$ can be treated similarly), relation (16) yields

$$\varphi_1(x) \geq c_1 d(x)^2 \quad \forall x \in B, \tag{20}$$

for some $c_1 > 0$. But u is a radial solution of (13) and therefore it is smooth in $B \setminus B_{1/4}$. Thus

$$u(x) \leq M\varphi_1 + \varphi_2 \quad \forall x \in B_{1/2}, \tag{21}$$

for some $M > 0$. Therefore, from (20) and (21), for $\lambda > \lambda^*$ with $\lambda - \lambda^*$ sufficiently small we have

$$\left(\frac{\lambda}{\lambda^*} - 1\right) u \leq \varphi_1 + \left(\frac{\lambda}{\lambda^*} - 1\right) \varphi_2 \quad \text{in } B.$$

Let $w = \frac{\lambda}{\lambda^*} u - \varphi_1 - (\frac{\lambda}{\lambda^*} - 1)\varphi_2$. The inequality just stated guarantees that $w \leq u$. Moreover,

$$\Delta^2 w = \lambda e^u + \frac{\lambda\mu}{\lambda^*}\eta e^u - \mu\eta e^u \geq \lambda e^u \geq \lambda e^w \quad \text{in } B$$

and

$$w = a, \quad \frac{\partial w}{\partial n} = b \quad \text{on } \partial B.$$

Therefore, w is a supersolution to (1) for λ. By the method of sub- and supersolutions a solution to (1) exists for some $\lambda > \lambda^*$, which is a contradiction. ∎

Proof of Proposition 1.7. Let $\lambda > 0$ and $u \in H^2(B)$ be a weak unbounded solution of (1). If $\lambda < \lambda^*$ from Lemma 2.6 we find that $u \leq u_\lambda$, where u_λ is the minimal solution. This is impossible because u_λ is smooth and u is unbounded. If $\lambda = \lambda^*$ then necessarily $u = u^*$ by Theorem 1.2. ∎

Proof of Proposition 1.8. Let u denote the extremal solution of (1) with $b \geq -4$. If u is smooth, then the result is trivial. So we restrict our attention to the case where u is singular. By Theorem 1.3 we have, in particular, that $N \geq 13$. We may also assume that $a = 0$. If $b = -4$ by Theorem 1.2 we know that if $N \geq 13$, then $u = -4\log|x|$ so that the desired conclusion holds. Henceforth we assume $b > -4$ in this section.

For $\rho > 0$ define

$$u_\rho(r) = u(\rho r) + 4\log\rho$$

so that

$$\Delta^2 u_\rho = \lambda^* e^{u_\rho} \qquad \text{in } B_{1/\rho}.$$

Then

$$\frac{du_\rho}{d\rho}\bigg|_{\rho=1,r=1} = u'(1) + 4 > 0.$$

Hence, there is $\delta > 0$ such that

$$u_\rho(r) < u(r) \quad \forall 1 - \delta < r \leq 1,\ 1 - \delta < \rho \leq 1.$$

This implies

$$u_\rho(r) < u(r) \quad \forall 0 < r \le 1,\, 1-\delta < \rho \le 1. \tag{22}$$

Otherwise set

$$r_0 = \sup\{\, 0 < r < 1 \,|\, u_\rho(r) \ge u(r)\,\}.$$

This definition yields

$$u_\rho(r_0) = u(r_0) \quad \text{ and } \quad u_\rho'(r_0) \le u'(r_0). \tag{23}$$

Write $\alpha = u(r_0)$, $\beta = u'(r_0)$. Then u satisfies

$$\begin{cases} \Delta^2 u = \lambda e^u & \text{on } B_{r_0}, \\ u(r_0) = \alpha, \\ u'(r_0) = \beta. \end{cases} \tag{24}$$

Observe that u is an unbounded $H^2(B_{r_0})$ solution to (24), which is also stable. Thus Proposition 1.7 shows that u is the extremal solution to this problem. On the other hand, u_ρ is a supersolution to (24), since $u_\rho'(r_0) \le \beta$ by (23). We may now use Theorem 1.2 and we deduce that

$$u(r) = u_\rho(r) \quad \forall 0 < r \le r_0,$$

which in turn implies by standard ODE theory that

$$u(r) = u_\rho(r) \quad \forall 0 < r \le 1,$$

which is a contradiction to (22). This proves estimate (22).

From (22) we see that

$$\frac{du_\rho}{d\rho}\Big|_{\rho=1}(r) \ge 0 \quad \forall 0 < r \le 1. \tag{25}$$

But

$$\frac{du_\rho}{d\rho}\Big|_{\rho=1}(r) = u'(r)r + 4 \quad \forall 0 < r \le 1,$$

and this together with (25) implies

$$\frac{du_\rho}{d\rho}(r) = \frac{1}{\rho}(u'(\rho r)\rho r + 4) \ge 0 \quad \forall 0 < r \le \frac{1}{\rho},\, 0 < \rho \le 1, \tag{26}$$

which means that $u_\rho(r)$ is nondecreasing in ρ. We wish to show that $\lim_{\rho\to 0} u_\rho(r)$ exists for all $0 < r \le 1$. For this we shall show

$$u_\rho(r) \ge -4\log(r) + \log\left(\frac{8(N-2)(N-4)}{\lambda^*}\right) \quad \forall 0 < r \le \frac{1}{\rho},\, 0 < \rho \le 1. \tag{27}$$

Set

$$u_0(r) = -4\log(r) + \log\left(\frac{8(N-2)(N-4)}{\lambda^*}\right),$$

and suppose that (27) is not true for some $0 < \rho < 1$. Let

$$r_1 = \sup \{ 0 < r < 1/\rho \,|\, u_\rho(r) < u_0(r) \}.$$

Observe that

$$\lambda^* > 8(N-2)(N-4). \tag{28}$$

Otherwise $w = -4 \ln r$ would be a strict supersolution of the equation satisfied by u, which is not possible by Theorem 1.2. In particular, $r_1 < 1/\rho$ and

$$u_\rho(r_1) = u_0(r_1) \quad \text{and} \quad u_\rho'(r_1) \geq u_0'(r_1).$$

It follows that u_0 is a supersolution of

$$\begin{cases} \Delta^2 u = \lambda^* e^u & \text{in } B_{r_1}, \\ u = A & \text{on } \partial B_{r_1}, \\ \dfrac{\partial u}{\partial n} = B & \text{on } \partial B_{r_1}, \end{cases} \tag{29}$$

with $A = u_\rho(r_1)$ and $B = u_\rho'(r_1)$. Since u_ρ is a singular stable solution of (29), it is the extremal solution of the problem by Proposition 1.7. By Theorem 1.2, there is no strict supersolution of (29), and we conclude that $u_\rho \equiv u_0$ first for $0 < r < r_1$ and then for $0 < r \leq 1/\rho$. This is impossible for $\rho > 0$ because $u_\rho(1/\rho) = 4 \log \rho$ and $u_0(1/\rho) < 4 \log \rho + \log(\frac{8(N-2)(N-4)}{\lambda^*}) < u_\rho(1/\rho)$ by (28). This proves (27).

By (26) and (27) we see that

$$v(r) = \lim_{\rho \to 0} u_\rho(r) \quad \text{exists } \forall 0 < r < +\infty,$$

where the convergence is uniform (even in C^k for any k) on compact sets of $\mathbb{R}^N \setminus \{0\}$. Moreover, v satisfies

$$\Delta^2 v = \lambda^* e^v \quad \text{in } \mathbb{R}^N \setminus \{0\}. \tag{30}$$

Then for any $r > 0$

$$v(r) = \lim_{\rho \to 0} u_\rho(r) = \lim_{\rho \to 0} u(\rho r) + 4 \log(\rho r) - 4 \log(r) = v(1) - 4 \log(r).$$

Hence, using (30) we obtain

$$v(r) = -4 \log r + \log \left(\frac{8(N-2)(N-4)}{\lambda^*} \right) = u_0(r).$$

But then

$$u_\rho'(r) = u'(\rho r)\rho \to -4 \quad \text{as } \rho \to 0,$$

and therefore, with $r = 1$

$$\rho u'(\rho) \to -4 \quad \text{as } \rho \to 0. \qquad \square \tag{31}$$

4. Proof of Theorem 1.3. First we will show the following lemma.

LEMMA 4.1. *Suppose that the extremal solution u^* to* (1) *is singular. Then for any $\sigma > 0$ there exists $0 < R < 1$ such that*

$$u^*(x) \geq (1-\sigma)\log\left(\frac{1}{|x|^4}\right) \quad \forall\, |x| < R. \tag{32}$$

Proof. Assume by contradiction that (32) is false. Then there exist $\sigma > 0$ and a sequence $x_k \in B$ with $x_k \to 0$ such that

$$u^*(x_k) < (1-\sigma)\log\left(\frac{1}{|x_k|^4}\right). \tag{33}$$

Let $s_k = |x_k|$ and choose $0 < \lambda_k < \lambda^*$ such that

$$\max_{\overline{B}} u_{\lambda_k} = u_{\lambda_k}(0) = \log\left(\frac{1}{s_k^4}\right). \tag{34}$$

Note that $\lambda_k \to \lambda^*$; otherwise u_{λ_k} would remain bounded. Let

$$v_k(x) = \frac{u_{\lambda_k}(s_k x)}{\log(\frac{1}{s_k^4})}, \qquad x \in B_k \equiv \frac{1}{s_k}B.$$

Then $0 \leq v_k \leq 1$, $v_k(0) = 1$,

$$\begin{aligned}\Delta^2 v_k(x) &= \lambda_k \frac{s_k^4}{\log(\frac{1}{s_k^4})} e^{u_{\lambda_k}(s_k x)} \\ &\leq \frac{\lambda_k}{\log(\frac{1}{s_k^4})} \to 0 \quad \text{in } B_k\end{aligned}$$

by (34). By elliptic regularity $v_k \to v$ uniformly on compact sets of $\mathbb{R}^N$ to a function v satisfying $0 \leq v \leq 1$, $v(0) = 1$, $\Delta^2 v = 0$ in $\mathbb{R}^N$. By Liouville's theorem for biharmonic functions [17] we conclude that v is constant and therefore $v \equiv 1$.

Since $|x_k| = s_k$ we deduce that

$$\frac{u_{\lambda_k}(x_k)}{\log(\frac{1}{s_k^4})} \to 1,$$

which contradicts (33). ◻

Proof of Theorem 1.3. We write for simplicity $u = u^*$, $\lambda = \lambda^*$. Assume by contradiction that u^* is unbounded and $5 \leq N \leq 12$. If $N \leq 4$ the problem is subcritical, and the boundedness of u^* can be proved by other means: no singular solutions exist for positive λ (see [2]), though in dimension $N = 4$, a family of solutions (u_λ) can blow up as $\lambda \to 0$ (see [24]).

For $\varepsilon > 0$ let $\psi = |x|^{\frac{4-N}{2}+\varepsilon}$ and let $\eta \in C_0^\infty(\mathbb{R}^N)$ with $\eta \equiv 1$ in $B_{1/2}$ and $\operatorname{supp}(\eta) \subseteq$ B. Observe that

$$(\Delta\psi)^2 = (H_N + O(\varepsilon))|x|^{-N+2\varepsilon}, \quad \text{where } H_N = \frac{N^2(N-4)^2}{16}.$$

Using a standard approximation argument as in the proof of Lemma 2.6, we can use $\psi\eta$ as a test function in (9) and we obtain

$$\int_B (\Delta\psi)^2 + O(1) \geq \lambda \int_B e^u \psi^2,$$

since the contribution of the integrals outside a fixed ball around the origin remains bounded as $\varepsilon \to 0$ (here $O(1)$ denotes a bounded function as $\varepsilon \to 0$).

This implies

$$\lambda \int_B e^u |x|^{4-N+2\varepsilon} \leq (H_N + O(\varepsilon)) \int_B |x|^{-N+2\varepsilon} = \omega_N \frac{H_N}{2\varepsilon} + O(1), \tag{35}$$

where ω_N is the surface area of the unit $N-1$ dimensional sphere S^{N-1}. In particular, $\int_B e^u|x|^{4-N+2\varepsilon} < +\infty$.

For $\varepsilon > 0$ we define $\varphi = |x|^{4-N+2\varepsilon}$. Note that away from the origin

$$\Delta^2\varphi = \varepsilon k_N |x|^{-N+2\varepsilon}, \quad \text{where } k_N = 4(N-2)(N-4) + O(\varepsilon). \tag{36}$$

Let φ_j solve

$$\begin{cases} \Delta^2 \varphi_j = \varepsilon k_N \min(|x|^{-N+2\varepsilon}, j) & \text{in } B, \\ \varphi_j = \dfrac{\partial \varphi_j}{\partial n} = 0 & \text{on } \partial B. \end{cases} \tag{37}$$

Then $\varphi_j \uparrow \varphi$ as $j \to +\infty$. Using (35) and (37)

$$\begin{aligned} \varepsilon k_N \int_B u \ \min(|x|^{-N+2\varepsilon}, j) = \int_B u\Delta^2\varphi_j &= \lambda \int_B e^u \varphi_j \\ &\leq \lambda \int_B e^u \varphi \\ &\leq \omega_N \frac{H_N}{2\varepsilon} + O(1), \end{aligned}$$

where $O(1)$ is bounded as $\varepsilon \to 0$ independently of j. Letting $j \to +\infty$ yields

$$\varepsilon k_N \int_B u \ |x|^{-N+2\varepsilon} \leq \omega_N \frac{H_N}{2\varepsilon} + O(1), \tag{38}$$

showing that the integral on the left-hand side is finite. On the other hand, by (32)

$$\varepsilon k_N \int_B u \ |x|^{-N+2\varepsilon} \geq \varepsilon k_N \omega_N (1-\sigma) \int_0^1 \log\left(\frac{1}{r^4}\right) r^{-1+2\varepsilon}\, dr = k_N \omega_N (1-\sigma) \frac{1}{\varepsilon}. \tag{39}$$

Combining (38) and (39), we obtain

$$(1-\sigma) k_N \leq \frac{H_N}{2} + O(\varepsilon).$$

Letting $\varepsilon \to 0$ and then $\sigma \to 0$, we have

$$8(N-2)(N-4) \leq H_N = \frac{N^2(N-4)^2}{16}.$$

This is valid only if $N \geq 13$, which is a contradiction. □

Remark 4.2. The conclusion of Theorem 1.3 can be obtained also from Proposition 1.8. However, that proposition depends crucially on the radial symmetry of the solutions, while the argument in this section can be generalized to other domains.

5. The extremal solution is singular in large dimensions. In this section we take $a = b = 0$ and prove Theorem 1.4 for $N \geq 32$.

The idea for the proof of Theorem 1.4 is to estimate accurately from above the function $\lambda^* e^{u^*}$, and to deduce that the operator $\Delta^2 - \lambda^* e^{u^*}$ has a strictly positive first eigenvalue (in the $H_0^2(B)$ sense). Then, necessarily, u^* is singular.

Upper bounds for both λ^* and u^* are obtained by finding suitable sub- and supersolutions. For example, if for some λ_1 there exists a supersolution, then $\lambda^* \geq \lambda_1$. If for some λ_2 one can exhibit a stable singular subsolution u, then $\lambda^* \leq \lambda_2$. Otherwise, $\lambda_2 < \lambda^*$, and one can then prove that the minimal solution u_{λ_2} is above u, which is impossible. The bound for u^* also requires a stable singular subsolution.

It turns out that in dimension $N \geq 32$ we can construct the necessary subsolutions and verify their stability by hand. For dimensions $13 \leq N \leq 31$ it seems difficult to find these subsolutions explicitly. We adopt then an approach that involves a computer assisted construction of subsolutions and verification of the desired inequalities. We present this part in the next section.

LEMMA 5.1. *Assume $N \geq 13$. Then $u^* \leq \bar{u} = -4 \log |x|$ in B_1.*

Proof. Define $\bar{u}(x) = -4 \log |x|$. Then $\bar{u}$ satisfies

$$\begin{cases} \Delta^2 \bar{u} = 8(N-2)(N-4)e^{\bar{u}} & \text{in } \mathbb{R}^N, \\ \bar{u} = 0 & \text{on } \partial B_1, \\ \dfrac{\partial \bar{u}}{\partial n} = -4 & \text{on } \partial B_1. \end{cases}$$

Observe that since $\bar{u}$ is a supersolution to (1) with $a = b = 0$ we deduce immediately that $\lambda^* \geq 8(N-2)(N-4)$.

In the case $\lambda^* = 8(N-2)(N-4)$ we have $u_\lambda \leq \bar{u}$ for all $0 \leq \lambda < \lambda^*$ because $\bar{u}$ is a supersolution, and therefore $u^* \leq \bar{u}$ holds. Alternatively, one can invoke Theorem 3 in [2] to conclude that we always have $\lambda^* > 8(N-2)(N-4)$.

Suppose now that $\lambda^* > 8(N-2)(N-4)$. We prove that $u_\lambda \leq \bar{u}$ for all $8(N-2)(N-4) < \lambda < \lambda^*$. Fix such λ and assume by contradiction that $u_\lambda \leq \bar{u}$ is not true. Note that for $r < 1$ and sufficiently close to 1 we have $u_\lambda(r) < \bar{u}(r)$ because $u_\lambda'(1) = 0$ while $\bar{u}'(1) = -4$. Let

$$R_1 = \inf\{\, 0 \leq R \leq 1 \mid u_\lambda < \bar{u} \text{ in } (R, 1)\,\}.$$

Then $0 < R_1 < 1$, $u_\lambda(R_1) = \bar{u}(R_1)$, and $u_\lambda'(R_1) \leq \bar{u}'(R_1)$. So u_λ is a supersolution to the problem

$$\begin{cases} \Delta^2 u = 8(N-2)(N-4)e^{u} & \text{in } B_{R_1}, \\ u = u_\lambda(R_1) & \text{on } \partial B_{R_1}, \\ \dfrac{\partial u}{\partial n} = u_\lambda'(R_1) & \text{on } \partial B_{R_1}, \end{cases} \tag{40}$$

while $\bar{u}$ is a subsolution to (40). Moreover it is stable for this problem, since from Rellich's inequality (6) and $8(N-2)(N-4) \leq N^2(N-4)^2/16$ for $N \geq 13$, we have

$$8(N-2)(N-4)\int_{B_{R_1}} e^{\bar{u}}\varphi^2 \leq \frac{N^2(N-4)^2}{16}\int_{\mathbb{R}^N} \frac{\varphi^2}{|x|^4} \leq \int_{\mathbb{R}^N} (\Delta\varphi)^2 \quad \forall \varphi \in C_0^\infty(B_{R_1}).$$

By Remark 2.7 we deduce that $\bar{u} \leq u_\lambda$ in B_{R_1}, which is impossible. $\square$

An upper bound for λ^* is obtained by considering again a stable, singular subsolution to the problem (with another parameter, though).

LEMMA 5.2. *For $N \geq 32$ we have*

$$\lambda^* \leq 8(N-2)(N-4)e^2. \tag{41}$$

Proof. Consider $w = 2(1-r^2)$ and define

$$u = \bar{u} - w,$$

where $\bar{u}(x) = -4\log|x|$. Then

$$\begin{aligned}\Delta^2 u = 8(N-2)(N-4)\frac{1}{r^4} = 8(N-2)(N-4)e^{\bar{u}} &= 8(N-2)(N-4)e^{u+w} \\ &\leq 8(N-2)(N-4)e^2 e^u.\end{aligned}$$

Also $u(1) = u'(1) = 0$, so u is a subsolution to (1) with parameter $\lambda_0 = 8(N-2)(N-4)e^2$.

For $N \geq 32$ we have $\lambda_0 \leq N^2(N-4)^2/16$. Then by (6) u is a stable subsolution of (1) with $\lambda = \lambda_0$. If $\lambda^* > \lambda_0 = 8(N-2)(N-4)e^2$ the minimal solution u_{λ_0} to (1) with parameter λ_0 exists and is smooth. From Lemma 2.6 we find $u \leq u_{\lambda_0}$ which is impossible because u is singular and u_{λ_0} is bounded. Thus we have proved (41) for $N \geq 32$. ∎

Proof of Theorem 1.4 *in the case $N \geq 32$.* Combining Lemmas 5.1 and 5.2, we have that if $N \geq 32$ then $\lambda^* e^{u^*} \leq r^{-4}\, 8(N-2)(N-4)e^2 \leq r^{-4} N^2(N-4)^2/16$. This and (6) show that

$$\inf_{\varphi \in C_0^\infty(B)} \frac{\int_B (\Delta\varphi)^2 - \lambda^* \int_B e^{u^*}\varphi^2}{\int_B \varphi^2} > 0,$$

which is not possible if u^* is bounded. ∎

6. A computer assisted proof for dimensions $13 \leq N \leq 31$. Throughout this section we assume $a = b = 0$. As was mentioned in the previous section, the proof of Theorem 1.4 relies on precise estimates for u^* and λ^*. We present first some conditions under which it is possible to find these estimates. Later we show how to meet such conditions with a computer assisted verification.

The first lemma is analogous to Lemma 5.2.

LEMMA 6.1. *Suppose there exist $\varepsilon > 0$, $\lambda > 0$, and a radial function $u \in H^2(B) \cap W^{4,\infty}_{loc}(B \setminus \{0\})$ such that*

$$\Delta^2 u \leq \lambda e^u \quad \forall 0 < r < 1,$$

$$|u(1)| \leq \varepsilon, \quad \left|\frac{\partial u}{\partial n}(1)\right| \leq \varepsilon,$$

$$u \notin L^\infty(B),$$

$$\lambda e^\varepsilon \int_B e^u \varphi^2 \leq \int_B (\Delta\varphi)^2 \quad \forall \varphi \in C_0^\infty(B). \tag{42}$$

Then

$$\lambda^* \leq \lambda e^{2\varepsilon}.$$

Proof. Let

$$\psi(r) = \varepsilon r^2 - 2\varepsilon \tag{43}$$

so that

$$\Delta^2 \psi \equiv 0, \quad \psi(1) = -\varepsilon, \quad \psi'(1) = 2\varepsilon,$$

and

$$-2\varepsilon \le \psi(r) \le -\varepsilon \quad \forall 0 \le r \le 1.$$

It follows that

$$\Delta^2(u+\psi) \le \lambda e^u = \lambda e^{-\psi} e^{u+\psi} \le \lambda e^{2\varepsilon} e^{u+\psi}.$$

On the boundary we have $u(1) + \psi(1) \le 0$, $u'(1) + \psi'(1) \ge 0$. Thus $u + \psi$ is a singular subsolution to the equation with parameter $\lambda e^{2\varepsilon}$. Moreover, since $\psi \le -\varepsilon$ we have $\lambda e^{2\varepsilon} e^{u+\psi} \le \lambda e^{\varepsilon} e^u$, and hence, from (42) we see that $u + \psi$ is stable for the problem with parameter $\lambda e^{2\varepsilon}$. If $\lambda e^{2\varepsilon} < \lambda^*$ then the minimal solution associated to the parameter $\lambda e^{2\varepsilon}$ would be above $u + \psi$, which is impossible because u is singular. ∎

LEMMA 6.2. *Suppose we can find $\varepsilon > 0$, $\lambda > 0$, and $u \in H^2(B) \cap W^{4,\infty}_{loc}(B \setminus \{0\})$ such that*

$$\Delta^2 u \ge \lambda e^u \quad \forall 0 < r < 1,$$
$$|u(1)| \le \varepsilon, \quad \left| \frac{\partial u}{\partial n}(1) \right| \le \varepsilon.$$

Then

$$\lambda e^{-2\varepsilon} \le \lambda^*.$$

Proof. Let ψ be given by (43). Then $u - \psi$ is a supersolution to the problem with parameter $\lambda e^{-2\varepsilon}$. ∎

The next result is the main tool to guarantee that u^* is singular. The proof, as in Lemma 5.1, is based on an upper estimate of u^* by a stable singular subsolution.

LEMMA 6.3. *Suppose there exist ε_0, $\varepsilon > 0$, $\lambda_a > 0$, and a radial function $u \in H^2(B) \cap W^{4,\infty}_{loc}(B \setminus \{0\})$ such that*

$$\Delta^2 u \le (\lambda_a + \varepsilon_0) e^u \quad \forall 0 < r < 1, \tag{44}$$

$$\Delta^2 u \ge (\lambda_a - \varepsilon_0) e^u \quad \forall 0 < r < 1, \tag{45}$$

$$|u(1)| \le \varepsilon, \quad \left| \frac{\partial u}{\partial n}(1) \right| \le \varepsilon, \tag{46}$$

$$u \notin L^\infty(B), \tag{47}$$

$$\beta_0 \int_B e^u \varphi^2 \le \int_B (\Delta \varphi)^2 \quad \forall \varphi \in C_0^\infty(B), \tag{48}$$

where

$$\beta_0 = \frac{(\lambda_a + \varepsilon_0)^3}{(\lambda_a - \varepsilon_0)^2} e^{9\varepsilon}. \tag{49}$$

Then u^* is singular and

$$(\lambda_a - \varepsilon_0)e^{-2\varepsilon} \le \lambda^* \le (\lambda_a + \varepsilon_0)e^{2\varepsilon}. \tag{50}$$

Proof. By Lemmas 6.1 and 6.2 we have (50). Let

$$\delta = \log\left(\frac{\lambda_a + \varepsilon_0}{\lambda_a - \varepsilon_0}\right) + 3\varepsilon$$

and define

$$\varphi(r) = -\frac{\delta}{4}r^4 + 2\delta.$$

We claim that

$$u^* \le u + \varphi \quad \text{ in } B_1. \tag{51}$$

To prove this, we shall show that for $\lambda < \lambda^*$

$$u_\lambda \le u + \varphi \quad \text{ in } B_1. \tag{52}$$

Indeed, we have

$$\begin{aligned} &\Delta^2 \varphi = -\delta 2N(N+2),\\ &\varphi(r) \ge \delta \quad \forall 0 \le r \le 1,\\ &\varphi(1) \ge \delta \ge \varepsilon, \quad \varphi'(1) = -\delta \le -\varepsilon, \end{aligned}$$

and therefore

$$\begin{aligned} \Delta^2(u+\varphi) &\le (\lambda_a + \varepsilon_0)e^u + \Delta^2\varphi \le (\lambda_a+\varepsilon_0)e^u = (\lambda_a+\varepsilon_0)e^{-\varphi}e^{u+\varphi}\\ &\le (\lambda_a+\varepsilon_0)e^{-\delta}e^{u+\varphi}. \end{aligned} \tag{53}$$

By (50) and the choice of δ

$$(\lambda_a + \varepsilon_0)e^{-\delta} = (\lambda_a - \varepsilon_0)e^{-3\varepsilon} < \lambda^*. \tag{54}$$

To prove (52) it suffices to consider λ in the interval $(\lambda_a - \varepsilon_0)e^{-3\varepsilon} < \lambda < \lambda^*$. Fix such λ and assume that (52) is not true. Write

$$\bar{u} = u + \varphi$$

and let

$$R_1 = \sup\{\, 0 \le R \le 1 \,|\, u_\lambda(R) = \bar{u}(R)\,\}.$$

Then $0 < R_1 < 1$ and $u_\lambda(R_1) = \bar{u}(R_1)$. Since $u'_\lambda(1) = 0$ and $\bar{u}'(1) < 0$ we must have $u'_\lambda(R_1) \le \bar{u}'(R_1)$. Then u_λ is a solution to the problem

$$\begin{cases} \Delta^2 u = \lambda e^u & \text{in } B_{R_1},\\ u = u_\lambda(R_1) & \text{on } \partial B_{R_1},\\ \dfrac{\partial u}{\partial n} = u'_\lambda(R_1) & \text{on } \partial B_{R_1}, \end{cases}$$

while, thanks to (53) and (54), $\bar{u}$ is a subsolution to the same problem. Moreover, $\bar{u}$ is stable thanks to (48) since, by Lemma 6.1,

$$\lambda < \lambda^* \le (\lambda_a + \varepsilon_0)e^{2\varepsilon} \tag{55}$$

and hence

$$\lambda e^{\bar{u}} \le (\lambda_a + \varepsilon_0)e^{2\varepsilon}e^{2\delta}e^{u} \le \beta_0 e^{u}.$$

We deduce $\bar{u} \le u_\lambda$ in B_{R_1} which is impossible, since $\bar{u}$ is singular while u_λ is smooth. This establishes (51).

From (51) and (55) we have

$$\lambda^* e^{u^*} \le \beta_0 e^{-\varepsilon} e^{u}$$

and therefore

$$\inf_{\varphi \in C_0^\infty(B)} \frac{\int_B (\Delta\varphi)^2 - \lambda^* e^{u^*}\varphi^2}{\int_B \varphi^2} > 0.$$

This is not possible if u^* is a smooth solution. $\square$

For each dimension $13 \le N \le 31$ we construct u satisfying (44)–(48) of the form

$$u(r) = \begin{cases} -4\log r + \log\left(\frac{8(N-2)(N-4)}{\lambda}\right) & \text{for } 0 < r < r_0, \\ \tilde{u}(r) & \text{for } r_0 \le r \le 1, \end{cases} \tag{56}$$

where $\tilde{u}$ is explicitly given. Thus u satisfies (47) automatically.

Numerically it is better to work with the change of variables

$$w(s) = u(e^s) + 4s, \quad -\infty < s < 0,$$

which transforms the equation $\Delta^2 u = \lambda e^u$ into

$$Lw + 8(N-2)(N-4) = \lambda e^{w}, \quad -\infty < s < 0,$$

where

$$Lw = \frac{d^4 w}{ds^4} + 2(N-4)\frac{d^3 w}{ds^3} + (N^2 - 10N + 20)\frac{d^2 w}{ds^2} - 2(N-2)(N-4)\frac{dw}{ds}.$$

The boundary conditions $u(1) = 0$, $u'(1) = 0$ then yield

$$w(0) = 0, \quad w'(0) = 4.$$

Regarding the behavior of w as $s \to -\infty$, observe that

$$u(r) = -4\log r + \log\left(\frac{8(N-2)(N-4)}{\lambda}\right) \quad \text{for } r < r_0$$

if and only if

$$w(s) = \log\frac{8(N-2)(N-4)}{\lambda} \quad \forall s < \log r_0.$$

The steps we perform are the following.

(1) We fix $x_0 < 0$ and using numerical software we follow a branch of solutions to

$$\begin{cases} L\hat{w} + 8(N-2)(N-4) = \lambda e^{\hat{w}}, \quad x_0 < s < 0, \\ \hat{w}(0) = 0, \quad \hat{w}'(0) = t, \\ \hat{w}(x_0) = \log \dfrac{8(N-2)(N-4)}{\lambda}, \quad \dfrac{d^2\hat{w}}{ds^2}(x_0) = 0, \quad \dfrac{d^3\hat{w}}{ds^3}(x_0) = 0 \end{cases}$$

as t increases from 0 to 4. The numerical solution $(\hat{w}, \hat{\lambda})$ we are interested in corresponds to the case $t = 4$. The five boundary conditions are due to the fact that we are solving a fourth order equation with an unknown parameter λ.

(2) Based on $\hat{w}$, $\hat{\lambda}$ we construct a C^3 function w which is constant for $s \le x_0$ and piecewise polynomial for $x_0 \le s \le 0$. More precisely, we first divide the interval $[x_0, 0]$ into smaller intervals of length h. Then we generate a cubic spline approximation g_{fl} with floating point coefficients of $\frac{d^4\hat{w}}{ds^4}$. From g_{fl} we generate a piecewise cubic polynomial g_{ra} which uses rational coefficients and we integrate it four times to obtain w, where the constants of integration are such that $\frac{d^j w}{ds^j}(x_0) = 0$, $1 \le j \le 3$, and $w(x_0)$ is a rational approximation of $\log(8(N-2)(N-4)/\lambda)$. Thus w is a piecewise polynomial function that in each interval is of degree 7 with rational coefficients, and which is globally C^3. We also let λ be a rational approximation of $\hat{\lambda}$. With these choices note that $Lw + 8(N-2)(N-4) - \lambda e^w$ is a small constant (not necessarily 0) for $s \le x_0$.

(3) The conditions (44) and (45) we need to check for u are equivalent to the following inequalities for w:

$$Lw + 8(N-2)(N-4) - (\lambda + \varepsilon_0)e^w \le 0, \quad -\infty < s < 0, \tag{57}$$

$$Lw + 8(N-2)(N-4) - (\lambda - \varepsilon_0)e^w \ge 0, \quad -\infty < s < 0. \tag{58}$$

Using a program in Maple we verify that w satisfies (57) and (58). This is done by evaluating a second order Taylor approximation of $Lw + 8(N-2)(N-4) - (\lambda + \varepsilon_0)e^w$ at sufficiently close mesh points. All arithmetic computations are done with rational numbers and thus obtain exact results. The exponential function is approximated by a Taylor polynomial of degree 14, and the difference with the real value is controlled.

More precisely, we write

$$f(s) = Lw + 8(N-2)(N-4) - (\lambda + \varepsilon_0)e^w,$$
$$\tilde{f}(s) = Lw + 8(N-2)(N-4) - (\lambda + \varepsilon_0)T(w),$$

where T is the Taylor polynomial of order 14 of the exponential function around 0. Applying Taylor's formula to f at y_j, we have for $s \in [y_j, y_{j+h}]$

$$\begin{aligned} f(s) &\le f(y_j) + |f'(y_j)|h + \frac{1}{2}Mh^2 \\ &\le \tilde{f}(y_j) + |\tilde{f}'(y_j)|h + \frac{1}{2}Mh^2 + |f(y_j) - \tilde{f}(y_j)| + |f'(y_j) - \tilde{f}'(y_j)|h \\ &\le \tilde{f}(y_j) + |\tilde{f}'(y_j)|h + \frac{1}{2}Mh^2 + E_1 + E_2 h, \end{aligned}$$

where

M is a bound for $|f''|$ in $[y_j, y_j + h]$,
E_1 is such that $(\lambda + \varepsilon_0)|e^w - T(w)| \le E_1$ in $[y_j, y_j + h]$,
E_2 is such that $(\lambda + \varepsilon_0)|(e^w - T'(w))w'| \le E_2$ in $[y_j, y_j + h]$.

So, inequality (57) will be verified on each interval $[y_j, y_j+h]$ where w is a polynomial as soon as

$$\tilde{f}(y_j) + |\tilde{f}'(y_j)|h + \frac{1}{2}Mh^2 + E_1 + E_2 h \le 0. \tag{59}$$

When more accuracy is desired, instead of (59) one can verify that

$$\tilde{f}(x_i) + |\tilde{f}'(x_i)|\frac{h}{m} + \frac{1}{2}M\left(\frac{h}{m}\right)^2 + E_1 + E_2\frac{h}{m} \le 0,$$

where $(x_i)_{i=1\ldots m+1}$ are $m+1$ equally spaced points in $[y_j, y_j+h]$.

We obtain exact values for the upper bounds M, E_1, E_2 as follows. First note that $f'' = Lw'' - (\lambda+\varepsilon_0)e^w((w')^2 + w'')$. On $[y_j, y_j+h]$, we have $w(s) = \sum_{i=0}^7 a_i(s-y_j)^i$ and we estimate $|w(s)| \le \sum_{i=0}^7 |a_i|h^i$ for $s \in [y_j, y_j+h]$. Similarly,

$$\left|\frac{d^\ell w}{ds^\ell}(s)\right| \le \sum_{i=\ell}^{7} i(i-1)\ldots(i-\ell+1)|a_i|h^{i-\ell} \quad \forall s \in [y_j, y_j+h]. \tag{60}$$

The exponential is estimated by $e^w \le e^1 \le 3$, since our numerical data satisfies the rough bounds $-3/2 \le w \le 1$. Using this information and (60) yields a rational upper bound M. E_1 is estimated using Taylor's formula:

$$E_1 = (\lambda+\varepsilon_0)\frac{(3/2)^{15}}{15!}.$$

Similarly, $E_2 = (\lambda+\varepsilon_0)\frac{(3/2)^{14}}{14!}B_1$, where B_1 is the right-hand side of (60) when $\ell = 1$.

(4) We show that the operator $\Delta^2 - \beta e^u$ where $u(r) = w(\log r) - 4\log r$, satisfies condition (48) for some $\beta \ge \beta_0$ where β_0 is given by (49). In dimension $N \ge 13$ the operator $\Delta^2 - \beta e^u$ has indeed a positive eigenfunction in $H_0^2(B)$ with finite eigenvalue if β is not too large. The reason is that near the origin

$$\beta e^u = \frac{c}{|x|^4},$$

where c is a number close to $8(N-2)(N-4)\beta/\lambda$. If β is not too large compared to λ, then $c < N^2(N-4)^2/16$, and hence, using (6), $\Delta^2 - \beta e^u$ is coercive in $H_0^2(B_{r_0})$ (this holds under even weaker conditions; see [11]). It follows that there exists a first eigenfunction $\varphi_1 \in H_0^2(B)$ for the operator $\Delta^2 - \beta e^u$ with a finite first eigenvalue μ_1; that is,

$$\begin{aligned}
&\Delta^2\varphi_1 - \beta e^u\varphi_1 = \mu_1\varphi_1 \quad \text{in } B,\\
&\varphi_1 > 0 \quad \text{in } B,\\
&\varphi_1 \in H_0^2(B).
\end{aligned}$$

Moreover, μ_1 can be characterized as

$$\mu_1 = \inf_{\varphi\in C_0^\infty(B)} \frac{\int_B (\Delta\varphi)^2 - \beta e^u\varphi^2}{\int_B \varphi^2}$$

and is the smallest number for which a positive eigenfunction in $H_0^2(\Omega)$ exists.

Thus to prove that (48) holds it suffices to verify that $\mu_1 \geq 0$ and for this it is enough to show the existence of a nonnegative $\varphi \in H_0^2(B)$, $\varphi \not\equiv 0$, such that

$$\left\{\begin{aligned} \Delta^2\varphi - \beta e^u \varphi \geq 0 &\qquad \text{in } B,\\ \varphi = 0 &\qquad \text{on } \partial B,\\ \frac{\partial \varphi}{\partial n} \leq 0 &\qquad \text{on } \partial B. \end{aligned}\right. \tag{61}$$

Indeed, multiplication of (61) by φ_1 and integration by parts yield

$$\mu_1 \int_B \varphi\varphi_1 + \int_{\partial B} \frac{\partial\varphi}{\partial n}\Delta\varphi_1 \geq 0.$$

But $\Delta\varphi_1 \geq 0$ on ∂B and thus $\mu_1 \geq 0$. To achieve (61) we again change variables and define

$$\phi(s) = \varphi(e^s), \quad -\infty < s \leq 0.$$

Then we have to find $\phi \geq 0$, $\phi \not\equiv 0$, satisfying

$$\left\{\begin{aligned} L\phi - \beta e^w \phi \geq 0 &\quad \text{in } -\infty < s \leq 0,\\ \phi(0) = 0&,\\ \phi'(0) \leq 0&. \end{aligned}\right. \tag{62}$$

Regarding the behavior as $s \to -\infty$, we note that w is constant for $-\infty < s < x_0$, and therefore, if

$$L\phi - \beta e^w \phi \equiv 0, \quad -\infty < s \leq x_0,$$

then ϕ is a linear combination of exponential functions $e^{-\alpha s}$, where α must be a solution to

$$\alpha^4 - 2(N-4)\alpha^3 + (N^2 - 10N + 20)\alpha^2 + 2(N-2)(N-4)\alpha = \beta e^{w(x_0)},$$

where $\beta e^{w(x_0)}$ is close to $8(N-2)(N-4)\beta/\lambda$. If $N \geq 13$ the polynomial

$$\alpha^4 - 2(N-4)\alpha^3 + (N^2 - 10N + 20)\alpha^2 + 2(N-2)(N-4)\alpha - 8(N-2)(N-4)$$

has four distinct real roots, while if $N \leq 12$ there are two real roots and two complex conjugates. If $N \geq 13$ there is exactly one root in the interval $(0, (N-4)/2)$, two roots greater than $(N-4)/2$, and one negative. We know that $\varphi(r) = \phi(\log r) \sim r^{-\alpha}$ is in H^2, which forces $\alpha < (N-4)/2$. It follows that for $s < x_0$, ϕ is a combination of $e^{-\alpha_0 s}$, $e^{-\alpha_1 s}$ where $\alpha_0 > 0$, $\alpha_1 < 0$ are the two roots smaller than $\alpha < (N-4)/2$. For simplicity, however, we will look for ϕ such that $\phi(s) = Ce^{-\alpha_0 s}$ for $s < x_0$, where $C > 0$ is a constant. This restriction will mean that we will not be able to impose $\phi'(0) = 0$ at the end. This is not a problem because $\phi'(0) \leq 0$.

Notice that we need only the inequality in (62), and hence we need to choose $\alpha \in (0, N-4/2)$ such that

$$\alpha^4 - 2(N-4)\alpha^3 + (N^2 - 10N + 20)\alpha^2 + 2(N-2)(N-4)\alpha \geq \beta e^{w(x_0)}.$$

The precise choice we employed in each dimension is in a summary table at the end of this section.

To find a suitable function ϕ with the behavior $\phi(s) = Ce^{-\alpha s}$ for $s < x_0$ we set $\phi = \psi e^{-\alpha s}$ and solve the equation

$$T_\alpha \psi - \beta e^w \psi = f,$$

where the operator T_α is given by

$$\begin{aligned} T_\alpha \psi = \frac{d^4\psi}{ds^4} &+ (-4\alpha + 2(N-4))\frac{d^3\psi}{ds^3} + (6\alpha^2 - 6\alpha(N-4) + N^2 - 10N + 20)\frac{d^2 w}{ds^2} \\ &+ (-4\alpha^3 + 6\alpha^2(N-4) - 2\alpha(N^2 - 10N + 20) - 2(N-2)(N-4))\frac{d\psi}{ds} \\ &+ (\alpha^4 - 2\alpha^3(N-4) + \alpha^2(N^2 - 10N + 20) + 2\alpha(N-2)(N-4))\psi \end{aligned}$$

and f is some smooth function such that $f \geq 0$, $f \not\equiv 0$. Actually we choose $\bar\beta > \beta_0$ (where β_0 is given in (49)) and find $\bar\alpha$ satisfying approximately

$$\bar\alpha^4 - 2(N-4)\bar\alpha^3 + (N^2 - 10N + 20)\bar\alpha^2 + 2(N-2)(N-4)\bar\alpha = \bar\beta e^{w(x_0)}.$$

We solve numerically

$$\begin{aligned} T_{\bar\alpha}\hat\psi - \bar\beta e^w \hat\psi = f, &\quad x_0 < s < 0, \\ \hat\psi(x_0) = 1, &\quad \hat\psi''(x_0) = 0, \quad \hat\psi'''(x_0) = 0, \\ \hat\psi(0) = 0. & \end{aligned}$$

Using the same strategy as in (2) from the numerical approximation of $\frac{d^4\hat\psi}{ds^4}$ we compute a piecewise polynomial ψ of degree 7, which is globally C^3 and constant for $s \leq x_0$. The constant $\psi(x_0)$ is chosen so that $\psi(0) = 0$. We then use Maple to verify the inequalities

$$\begin{aligned} \psi \geq 0, &\quad x_0 \leq s \leq 0, \\ T_\alpha \psi - \beta e^w \psi \geq 0, &\quad x_0 \leq s \leq 0, \\ \psi'(0) \leq 0, & \end{aligned}$$

where $\beta_0 < \beta < \bar\beta$ and $0 < \alpha < (N-4)/2$ are suitably chosen.

At the URLs http://www.lamfa.u-picardie.fr/dupaigne/ and http://www.ime.unicamp.br/~msm/ we provide the data of the functions w and ψ defined as piecewise polynomials of degree 7 in $[x_0, 0]$ with rational coefficients for each dimension in $13 \leq N \leq 31$. We also give a rational approximation of the constants involved in the corresponding problems.

We use Maple to verify that w and ψ (with suitable extensions) are C^3 global functions and satisfy the corresponding inequalities, using only its capability to operate on arbitrary rational numbers. These operations are exact and are limited only by the memory of the computer and clearly slower than floating point operations. We chose Maple since it is a widely used software, but the reader can check the validity of our results with any other software (see, e.g., the open-source solution pari/gp).

The tests were conducted using Maple 9. See Table 1 for a summary of parameters and results.

Remark 6.4. (1) Although we work with λ rational, in Table 1 we prefer to display a decimal approximation of λ.

(2) In Table 1 we selected a "large" value of ε_0 in order to have a fast verification with Maple. By requiring more accuracy in the numerical calculations, using a

TABLE 1

N	λ	ε_0	ε	β	β	α
13	2438.6	1	$5 \cdot 10^{-7}$	2550	2500	3.9
14	2911.2	1	$3 \cdot 10^{-6}$	3100	3000	3.4
15	3423.8	1	$3 \cdot 10^{-6}$	3600	3500	3.1
16	3976.4	1	$1 \cdot 10^{-5}$	4100	4000	3.0
17	4568.8	1	$2 \cdot 10^{-4}$	4800	4600	3.0
18	5201.1	2	$2 \cdot 10^{-4}$	5400	5300	2.7
19	5873.2	2	$2 \cdot 10^{-4}$	6100	6000	2.7
20	6585.1	3	$7 \cdot 10^{-4}$	7000	6800	2.7
21	7336.7	3	$7 \cdot 10^{-4}$	7700	7500	2.6
22	8128.1	4	$1 \cdot 10^{-3}$	8600	8400	2.6
23	8959.1	4	$1 \cdot 10^{-3}$	9400	9200	2.5
24	9829.8	4	$1 \cdot 10^{-3}$	10400	10200	2.5
25	10740.1	4	$1 \cdot 10^{-3}$	11400	11200	2.5
26	11690.1	6	$2 \cdot 10^{-3}$	12400	12200	2.5
27	12679.7	7	$2 \cdot 10^{-3}$	13400	13200	2.4
28	13709.0	7	$2 \cdot 10^{-3}$	14500	14300	2.4
29	14777.8	7	$2 \cdot 10^{-3}$	15400	15200	2.4
30	15886.2	8	$2 \cdot 10^{-3}$	16600	16400	2.4
31	17034.3	10	$2 \cdot 10^{-3}$	17600	17500	2.3

TABLE 2

N	λ	ε_0	ε	λ^*_{min}	λ^*_{max}	β	β	α
13	2438.589	0.003	$5 \cdot 10^{-7}$	2438.583	2438.595	2550	2510	3.9
14	2911.194	0.003	$5 \cdot 10^{-7}$	2911.188	2911.200	3100	3000	3.4

smaller value of ε_0, and using more subintervals to verify the inequalities in the Maple program, it is possible to obtain better estimates of λ^*. For instance, using formulas (50), we obtained the results in Table 2.

The verification above, however, is required to check 1500 subintervals of each of the 4500 intervals of length 0.002, which amounts to substantial computer time.

7. Proof of Proposition 1.5. Throughout this section, we restrict our attention, as permitted, to the case $a = 0$.

(a) Let u denote the extremal solution of (1) with homogeneous Dirichlet boundary condition $a = b = 0$. We extend u on its maximal interval of existence $(0, \bar{R})$.

LEMMA 7.1. *$\bar{R} < \infty$ and $u(r) \sim \log(\bar{R} - r)^{-4}$ for $r \sim \bar{R}$.*

Proof. The fact that $\bar{R} < \infty$ can be readily deduced from section 2 of [1]. We present an alternative (and more quantitative) argument. We first observe that

$$u'' - \frac{1}{r}u' > 0 \qquad \forall r \in [1, \bar{R}). \tag{63}$$

Integrate indeed (1) over a ball of radius r to conclude that

$$0 < \lambda \int_{B_r} e^u = \int_{\partial B_r} \frac{\partial}{\partial r}\Delta u = \omega_N r^{N-1}\left(u''' + \frac{N-1}{r}\left(u'' - \frac{1}{r}u'\right)\right). \tag{64}$$

If $r = 1$, since u is nonnegative in $(0, 1)$ and $u(1) = u'(1) = 0$, we must have $u''(1) \geq 0$. In fact, $u''(1) > 0$. Otherwise, we would have $u''(1) = 0$ and $u'''(1) > 0$ by (64), contradicting $u > 0$ in $(0, 1)$. So, we may define

$$R = \sup\left\{r > 1 : u''(t) - \frac{1}{t}u'(t) > 0 \;\; \forall t \in [1, r)\right\},$$

and we just need to prove that $R = \bar{R}$. Assume this is not the case; then $u''(R) - \frac{1}{R}u'(R) = 0$ and $u'''(R) = \left(u'' - \frac{1}{R}u'\right)'(R) \leq 0$. This contradicts (64) and we have just proved (63). In particular, we see that u is convex increasing on $(1, \bar{R})$.

Since u is radial, (1) reduces to

$$u^{(4)} + \frac{2(N-1)}{r}u''' + \frac{(N-1)(N-3)}{r^2}u'' - \frac{(N-1)(N-3)}{r^3}u' = \lambda e^u. \tag{65}$$

Multiply by u':

$$u^{(4)}u' + \frac{2(N-1)}{r}u'''u' + \frac{(N-1)(N-3)}{r^2}u''u' - \frac{(N-1)(N-3)}{r^3}(u')^2 = \lambda(e^u)',$$

which we rewrite as

$$\begin{aligned}[(u'''u')' - u'''u''] + 2(N-1)\left[\left(\frac{1}{r}u''u'\right)' - u''\left(\frac{1}{r}u'\right)'\right]& \\ +(N-1)(N-3)\left(\frac{(u')^2}{2r^2}\right)' &= \lambda(e^u)'.\end{aligned}$$

By (63), it follows that for $r \in [1, \bar{R})$,

$$[(u'''u')' - u'''u''] + 2(N-1)\left(\frac{1}{r}u''u'\right)' + (N-1)(N-3)\left(\frac{(u')^2}{2r^2}\right)' \geq \lambda(e^u)'.$$

Integrating, we obtain for some constant A

$$u'''u' - \frac{(u'')^2}{2} + 2(N-1)\frac{1}{r}u''u' + \frac{(N-1)(N-3)}{2}\frac{(u')^2}{r^2} \geq \lambda e^u - A.$$

We multiply again by u':

$$\begin{aligned}\left[(u''(u')^2)' - u''\left((u')^2\right)'\right] - \frac{1}{2}(u'')^2u' + 2(N-1)\frac{1}{r}u''(u')^2& \\ +\frac{(N-1)(N-3)}{2}\frac{1}{r^2}(u')^3 &\geq (\lambda e^u - Au)'.\end{aligned} \tag{66}$$

We deduce from (63) that

$$\frac{1}{r}u''(u')^2 = \frac{1}{2}\left(\frac{1}{r}(u')^3\right)' - \frac{1}{2}(u')^2\left(\frac{1}{r}u'\right)' \leq \frac{1}{2}\left(\frac{1}{r}(u')^3\right)' \quad \text{and}$$

$$\frac{1}{r^2}(u')^3 \leq \frac{1}{r}(u')^2u'' \leq \frac{1}{2}\left(\frac{1}{r}(u')^3\right)'.$$

Using this information in (66), dropping nonpositive terms, and integrating, we obtain for some constant B

$$u''(u')^2 + \frac{(N^2-1)}{4}\frac{1}{r}(u')^3 \geq \lambda e^u - Au - B.$$

Applying (63) again, it follows that for $C = \frac{N^2-1}{4} + 1$

$$Cu''(u')^2 \geq \lambda e^u - Au - B,$$

which after multiplication by u' and integration provides positive constants c, C such that

$$(u')^4 \geq c(e^u - Au^2 - Bu - C).$$

At this point, we observe that since u is convex and increasing, u converges to $+\infty$ as r approaches $\bar{R}$. Hence, for r close enough to $\bar{R}$ and for $c > 0$ perhaps smaller,

$$u' \geq c\, e^{u/4}.$$

By Gronwall's lemma, $\bar{R}$ is finite and

$$u \leq -4\log(\bar{R} - r) + C \qquad \text{for } r \text{ close to } \bar{R}.$$

It remains to prove that $u \geq -4\log(\bar{R} - r) - C$. This time, we rewrite (1) as

$$\left[r^{N-1}(\Delta u)'\right]' = \lambda r^{N-1} e^u.$$

We multiply by $r^{N-1}(\Delta u)'$ and obtain

$$\frac{1}{2}\left[r^{2N-2}((\Delta u)')^2\right]' = \lambda r^{2N-2} e^u (\Delta u)' \leq C e^u (\Delta u)' \leq C(e^u \Delta u)'.$$

Hence, for r close to $\bar{R}$ and C perhaps larger,

$$((\Delta u)')^2 \leq C e^u \Delta u,$$

and so

$$\sqrt{\Delta u}(\Delta u)' \leq C e^{u/2} \Delta u \leq C' e^{u/2} u'' \leq C'(e^{u/2} u')',$$

where we have used (63). Integrate to conclude that

$$(\Delta u)^{3/2} \leq C e^{u/2} u'.$$

Solving for Δu and multiplying by $(u')^{1/3}$, we obtain in particular that

$$(u')^{1/3} u'' \leq C e^{u/3} u'.$$

Integrating again, it follows that $(u')^{4/3} \leq C e^{u/3}$, i.e.,

$$u' \leq C e^{u/4}.$$

It then follows easily that (for r close to $\bar{R}$)

$$u \geq -4\log(\bar{R} - r) - C. \qquad \square$$

Proof of Proposition 1.5(a). Given $N \geq 13$, let b^{max} denote the supremum of all parameters $b \geq -4$ such that the corresponding extremal solution is singular. We first observe that

$$b^{max} > 0.$$

In fact, it follows from sections 5 and 6 that the extremal solution u associated to parameters $a = b = 0$ is strictly stable:

$$\inf_{\varphi \in C_0^\infty(B)} \frac{\int_B (\Delta\varphi)^2 - \lambda^* \int_B e^u \varphi^2}{\int_B \varphi^2} > 0. \tag{67}$$

Extend u as before on its maximal interval of existence $(0,\bar{R})$. Choosing $R \in (1,\bar{R})$ close to 1, we deduce that (67) still holds on the ball B_R. In particular, letting $v(x) = u(Rx) - u(R)$ for $x \in B$, we conclude that v is a singular stable solution of (1) with $a = 0$ and $b = Ru'(R) > 0$. By Proposition 1.7, we conclude that $b^{max} > 0$. We now prove that

$$b^{max} < \infty.$$

Assume this is not the case and let u_n denote the (singular) extremal solution associated to b_n, where $b_n \nearrow \infty$. We first observe that there exists $\rho_n \in (0,1)$ such that $u_n'(\rho_n) = 0$. Otherwise, u_n would remain monotone increasing on $(0,1)$ and hence bounded above by $u_n(1) = 0$. It would then follow from (1) and elliptic regularity that u_n is bounded. Let $v_n(x) = u_n(\rho_n x) - u_n(\rho_n)$ for $x \in B$ and observe that v_n solves (1) with $a = b = 0$ and some $\lambda = \lambda_n$. Clearly v_n is stable and singular. By Proposition 1.7, v_n coincides with u, the extremal solution of (1) with $a = b = 0$. By standard ODE theory, $v_n = u$ on $(0,\bar{R})$. In addition,

$$b_n = u_n'(1) = \frac{1}{\rho_n} v_n'\left(\frac{1}{\rho_n}\right) = \frac{1}{\rho_n} u'\left(\frac{1}{\rho_n}\right) \to +\infty,$$

which can happen only if $1/\rho_n \to \bar{R}$.

Now, since u_n is stable on B, $u = v_n$ is stable on B_{1/ρ_n}. Letting $n \to \infty$, we conclude that u is stable on $B_{\bar{R}}$. This clearly contradicts Lemma 7.1.

We have just proved that b^{max} is finite. It remains to prove that u^* is singular when $-4 \le b \le b^{max}$. We begin with the case $b = b^{max}$. Choose a sequence (b_n) converging to b^{max} and such that the corresponding extremal solution u_n is singular. Using the same notation as above, we find a sequence $\rho_n \in (0,1)$ such that

$$\frac{1}{\rho_n} u'\left(\frac{1}{\rho_n}\right) = b_n \to b^{max}.$$

Taking subsequences if necessary and passing to the limit as $n \to \infty$, we obtain for some $\rho \in (0,1)$

$$\frac{1}{\rho} u'\left(\frac{1}{\rho}\right) = b^{max}.$$

Furthermore, by construction of ρ_n, u is stable in B_{1/ρ_n} and hence in $B_{1/\rho}$. This implies that v defined for $x \in B$ by $v(x) = u(\frac{x}{\rho}) - u(\frac{1}{\rho})$ is a stable singular solution of (1) with $b = b^{max}$. By Proposition 1.7, we conclude that the extremal solution is singular when $b = b^{max}$.

When $b = -4$, as we have already mentioned in the introduction, u^* is singular for $N \ge 13$ as a direct consequence of Proposition 1.7 and Rellich's inequality.

So we are left with the case $-4 < b < b^{max}$. Let u_m^* denote the extremal solution when $b = b^{max}$, which is singular, and λ_m^* the corresponding parameter. For $0 < R < 1$ set

$$u_R(x) = u_m^*(Rx) - u_m^*(R).$$

Then

$$\Delta^2 u_R = \lambda_R e^{u_R}, \quad \text{where} \quad \lambda_R = \lambda_0^* R^4 e^{u_m^*(R)},$$

and $u_R = 0$ on ∂B, while

$$\frac{du_R}{dr}(1) = R\frac{du_m^*}{dr}(R).$$

By (31), note that

$$R\frac{du_m^*}{dr}(R) \to b^{max} \quad \text{as } R \to 1, \quad \text{and} \quad R\frac{du_m^*}{dr}(R) \to -4 \quad \text{as } R \to 0.$$

Thus, for any $-4 < b < b^{max}$ we have found a singular stable solution to (1) (with $a = 0$). By Proposition 1.7 the extremal solution to this problem is singular. $\square$

Proof of Proposition 1.5(b). Let $b \geq -4$. Lemma 5.1 applies also for $b \geq -4$ and yields $u^* \leq \bar{u}$, where $\bar{u}(x) = -4\log|x|$. We now modify slightly the proof of Lemma 5.2. Indeed, consider $w = (4+b)(1-r^2)/2$ and define $u = \bar{u} - w$. Then

$$\begin{aligned}\Delta^2 u &= 8(N-2)(N-4)\frac{1}{r^4} = 8(N-2)(N-4)e^{\bar{u}} = 8(N-2)(N-4)e^{u+w}\\ &\leq 8(N-2)(N-4)e^{(4+b)/2}e^u.\end{aligned}$$

Also $u(1) = 0$ and $u'(1) = b$, so u is a subsolution to (1) with parameter $\lambda_0 = 8(N-2)(N-4)e^{(4+b)/2}$.

If N is sufficiently large, depending on b, we have $\lambda_0 < N^2(N-4)^2/16$. Then by (6) u is a stable subsolution of (1) with $\lambda = \lambda_0$. As in Lemma 5.2 this implies $\lambda^* \leq \lambda_0$.

Thus for large enough N we have $\lambda^* e^{u^*} \leq r^{-4}\, 8(N-2)(N-4)e^{(4+b)/2} < r^{-4}N^2(N-4)^2/16$. This and (6) show that

$$\inf_{\varphi \in C_0^\infty(B)} \frac{\int_B (\Delta\varphi)^2 - \lambda^* \int_B e^{u^*}\varphi^2}{\int_B \varphi^2} > 0,$$

which is not possible if u^* is bounded.

Acknowledgments. Juan Dávila and Louis Dupaigne are indebted to J. Coville for useful conversations on some of the topics in the paper. The authors thank the referee for his careful reading of the manuscript.

REFERENCES

[1] G. Arioli, F. Gazzola, and H.-C. Grunau, *Entire solutions for a semilinear fourth order elliptic problem with exponential nonlinearity*, J. Differential Equations, 230 (2006), pp. 743–770.

[2] G. Arioli, F. Gazzola, H.-C. Grunau, and E. Mitidieri, *A semilinear fourth order elliptic problem with exponential nonlinearity*, SIAM J. Math. Anal., 36 (2005), pp. 1226–1258.

[3] S. Baraket, M. Dammak, T. Ouni, and F. Pacard, *Singular limits for 4-dimensional semilinear elliptic problems with exponential nonlinearity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.

[4] E. Berchio and F. Gazzola, *Some remarks on biharmonic elliptic problems with positive, increasing, and convex nonlinearities*, Electron. J. Differential Equations, 34 (2005), pp. 1–20.

[5] E. Berchio, F. Gazzola, and E. Mitidieri, *Positivity preserving property for a class of biharmonic problems*, J. Differential Equations, 229 (2006), pp. 1–23.

[6] T. Boggio, *Sulle funzioni di Freen d'ordine m*, Rend. Circ. Mat. Palermo, 20 (1905), pp. 97–135.

[7] H. Brezis, T. Cazenave, Y. Martel, and A. Ramiandrisoa, *Blow-up for $u_t - \Delta u = g(u)$ revisited*, Adv. Differential Equations, 1 (1996), pp. 73–90.

[8] H. Brezis and J. L. Vázquez, *Blow-up solutions of some nonlinear elliptic problems*, Rev. Mat. Univ. Comput. Madrid, 10 (1997), pp. 443–469.

[9] D. S. Cohen and H. B. Keller, *Some positone problems suggested by nonlinear heat generation*, J. Math. Mech., 16 (1967), pp. 1361–1376.

[10] M. G. Crandall and P. H. Rabinowitz, *Some continuation and variational methods for positive solutions of nonlinear elliptic eigenvalue problems*, Arch. Ration. Mech. Anal., 58 (1975), pp. 207–218.

[11] J. Dávila and L. Dupaigne, *Comparison results for PDEs with a singular potential*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 61–83.

[12] Z. Djadli and A. Malchiodi, *Existence of conformal metrics with constant Q-curvature*, Ann. of Math (2), to appear.

[13] A. Ferrero and H.-C. Grunau, *The Dirichlet problem for supercritical biharmonic equations with power-type nonlinearity*, J. Differential Equations, 234 (2007), pp. 582–606.

[14] F. Gazzola and H.-C. Grunau, *Critical dimensions and higher order Sobolev inequalities with remainder terms*, NoDEA Nonlinear Differential Equations Appl., 8 (2001), pp. 35–44.

[15] F. Gazzola, H.-C. Grunau, and E. Mitidieri, *Hardy inequalities with optimal constants and remainder terms*, Trans. Amer. Math. Soc., 356 (2004), pp. 2149–2168.

[16] H.-C. Grunau and G. Sweers, *Positivity for equations involving polyharmonic operators with Dirichlet boundary conditions*, Math. Ann., 307 (1997), pp. 589–626.

[17] R. R. Huilgol, *On Liouville's theorem for biharmonic functions*. SIAM J. Appl. Math., 20 (1971), pp. 37–39.

[18] D. D. Joseph and T. S. Lundgren, *Quasilinear Dirichlet problems driven by positive sources*, Arch. Ration. Mech. Anal., 49 (1972/73), pp. 241–269.

[19] Y. Martel, *Uniqueness of weak extremal solutions of nonlinear elliptic problems*, Houston J. Math., 23 (1997), pp. 161–168.

[20] F. Mignot and J. P. Puel, *Sur une classe de problèmes non linéaires avec non linéairité positive, croissante, convexe*, Comm. Partial Differential Equations, 5 (1980), pp. 791–836.

[21] F. Mignot and J. P. Puel, *Solution radiale singulière de $-\Delta u = \lambda e^u$*, C. R. Acad. Sci. Paris Sér. I Math., 307 (1988), pp. 379–382.

[22] J. J. Moreau, *Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires*, C. R. Acad. Sci. Paris, 255 (1962), pp. 238–240.

[23] F. Rellich, *Halbbeschränkte Differentialoperatoren höherer Ordnung*, in Proceedings of the International Congress of Mathematicians (Amsterdam 1954), Vol. III, J. C. H. Gerretsen et al., eds., Nordhoff, Groningen, The Netherlands, 1956, pp. 243–250.

[24] J. Wei, *Asymptotic behavior of a nonlinear fourth order eigenvalue problem*, Comm. Partial Differential Equations, 21 (1996), pp. 1451–1467.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 593–607

GENERIC VALIDITY OF THE MULTIFRACTAL FORMALISM*

A. FRAYSSE†

Abstract. The multifractal formalism is a conjecture which gives the spectrum of singularities of a signal using numerically computable quantities. We prove its generic validity by showing that almost every function in a given function space is multifractal and satisfies the multifractal formalism.

Key words. Hölder regularity, multifractal analysis, prevalence, wavelet expansion

AMS subject classifications. 28C20, 26A15, 28A80, 42C40

DOI. 10.1137/060669760

1. Introduction.

One motivation of multifractal analysis was the study of fully developed turbulent flows. Indeed, some experimental results obtained in wind tunnels showed that the regularity of the velocity of a turbulent fluid changes wildly from point to point. This quantity is therefore hard to compute. Hence, rather than measure the exponent at some point, one rather estimates the fractal dimension of sets where it takes a given value H.

The *spectrum of singularities* $d(H)$ is the function which gives the Hausdorff dimension of those sets. From its definition, it is also almost impossible to obtain numerically the spectrum of singularities.

In [9], physicists Frisch and Parisi proposed an algorithm in order to derive the spectrum of singularities from quantities that are effectively computable on a signal. They proposed using the L^p modulus of continuity of the velocity, used in the theory of turbulent flows; see [17]. This average quantity is called the *scaling* function, or scaling exponent, and is denoted by ξ_f. It is defined by $\int |f(x+l)-f(x)|^p dx \sim |l|^{\xi_f(p)}$, where $\sim$ means that $\int |f(x+l)-f(x)|^p dx$ is of the order of magnitude of $|l|^{\xi_f(p)}$ when l tends to 0 (assuming that the limit exists). Numerical estimations and further results about the scaling function and its wavelet decomposition can be found in [1, 2].

Frisch and Parisi proposed that the spectrum of singularities of a function can be obtained as follows:

$$d(H) = \inf_{p\in\mathbb{R}} (pH - \xi_f(p) + d); \tag{1}$$

see [9] for a heuristic derivation of this formula.

First, we state the mathematical framework of multifractal analysis. The main notion we need to define is the Hölder exponent.

DEFINITION 1. *Let $\alpha \geq 0$; a function $f : \mathbb{R}^d \to \mathbb{R}$ is $C^\alpha(x_0)$ if for all $x \in \mathbb{R}^d$ such that $\|x - x_0\| \leq 1$ there exist a polynomial P of degree less than $[\alpha]$ and a constant $C > 0$ such that*

$$|f(x) - P(x - x_0)| \leq C\|x - x_0\|^\alpha. \tag{2}$$

*Received by the editors September 13, 2006; accepted for publication (in revised form) March 9, 2007; published electronically August 1, 2007.
http://www.siam.org/journals/sima/39-2/66976.html

†Laboratoire d'analyse et de mathématiques appliquées, Université Paris XII, France. Current address: TSI, ENST, 46, rue Barrault, 75634 Paris Cédex 13, France (fraysse@tsi.enst.fr).

The Hölder exponent of f at x_0 is

$$h_f(x_0) = \sup\{\alpha : f \in C^\alpha(x_0)\}.$$

It was proved in [14] that for $p \geq 1$, the scaling function $\xi_f(p)$ is closely related to Sobolev or Besov smoothness. It is thus natural to replace the scaling function $\xi_f(p)$ as follows:

$$\text{If } p > 0, \quad \eta_f(p) = \sup\{s : \ f \in B_p^{s/p,\infty}\}. \tag{3}$$

So (1) applied to η_f can at most give the increasing part of the spectrum.

Defining, as in [16], an auxiliary function $s(1/p) = \eta(p)/p$, the Besov domain of a function f is the set of (q,t) such that $f \in B_{1/q}^{t,1/q}$. The boundary of the Besov domain of f is then given by the graph of $s(q)$. And by Sobolev embeddings, the Besov domain of a function is a convex set. Thus, the functions η satisfying (3) are increasing and concave. Furthermore the auxiliary function s is such that $0 \leq s'(q) \leq d$. These facts lead us to the following definition.

Definition 2. *A function η is admissible if $s(q) = q\eta(1/q)$ is concave and satisfies $0 \leq s'(q) \leq d$. Furthermore it is strongly admissible if $s(0) > 0$.*

The following important result from [16] allows us to define a metric space using admissible functions.

Proposition 3. *Any concave function s satisfying $0 \leq s'(q) \leq d$ defines the Besov domain of a distribution f.*

Thanks to Proposition 3, a metric space V can be associated to each admissible function η by taking

$$V = \bigcap_{\varepsilon>0,0<p<\infty} B_{p,loc}^{(\eta(p)-\varepsilon)/p,p}.$$

For the sake of completeness, we also recall the definition of Legendre transforms.

Definition 4. *Let f be a lower semicontinuous function defined in a normed vector space E. Then the Legendre transform of f is*

$$f^*(x) = \sup_{y\in E}(f(y) - xy). \tag{4}$$

This function is convex and lower semicontinuous.

In the present paper, we propose to study the validity of (1) for $\eta_f(p)$. An equivalent form of this heuristic formula is satisfied by a large class of invariant measures; see [4, 6, 19]. In the context of signal analysis, this conjecture is often satisfied if we add particular assumptions on f, such as self-similarity. On the other hand, there exist counterexamples to the general validity of this formula. And then, if it does not hold for every function, what is its range of validity? Our purpose here is to show that the validity of formula (1) is not an exceptional phenomenon, but it is satisfied for a large class of functions, without any additional assumption. More precisely, we study the validity of this formula for "almost every" function, i.e., in a measure-theoretic sense.

In a finite-dimensional space, the notion of "almost every" means "for the Lebesgue measure." The particular role played by this measure is justified by the fact that this is the only one which is σ-finite and invariant under translation. In a metric infinite-dimensional space no measure enjoys these properties. The following definition (see [5, 7, 11]) can thus replace the notion of the vanishing Haar measure.

Definition 5. *Let V be a complete metric vector space. A Borel set B in V is called Haar-null if there exists a probability measure μ with compact support such that*

$$\mu(B+v)=0 \quad \forall v\in V. \tag{5}$$

In this case the measure μ is said to be transverse to B.

A subset of V is called Haar-null if it is contained in a Borel Haar-null set.

The complement of a Haar-null set is called a prevalent set.

In a slight abuse of language we will say that a property is satisfied almost everywhere when it holds on a prevalent set.

Let us recall some properties of Haar-null sets; see [7, 11].

Proposition 6.

1. *If S is Haar-null, then for all $x\in V$, $x+S$ is Haar-null.*
2. *If $\dim(V)<\infty$, S is Haar-null if and only if $\mathrm{meas}(S)=0$ (where* meas *denotes the Lebesgue measure).*
3. *Prevalent sets are dense.*
4. *If S is Haar-null and $S'\subset S$, then S' is Haar-null.*
5. *The union of a countable collection of Haar-null sets is Haar-null.*
6. *If $\dim(V)=\infty$, compact subsets of V are Haar-null.*

Several kinds of measures can be used as transverse measures of a Borel set. Here, we will use only the following notion.

Definition 7. *A finite-dimensional space P is called a probe for a set $T\subset V$ if the Lebesgue measure on P is transverse to the complement of T.*

Those measures are not compactly supported probability measures. However, one immediately checks that Definition 7 is equivalent to the same one stated with the Lebesgue measure defined on the unit ball of P. Note that in this case, the support of the measure is included in the unit ball of a finite-dimensional subspace. The compactness assumption is therefore fulfilled.

The study of generic regularity for a "large" set of functions goes back to Banach [3], who gave differentiability properties of continuous functions for quasi-all functions in the Baire's categories sense. Later Hunt [12] proved the same result in the measure-theoretic sense of prevalence.

In [16], Jaffard studied properties of generic functions, in the Baire's categories sense, in Sobolev spaces. He also proved that in the sense of Baire's categories quasi-all functions in V satisfy

$$d(H)=\inf_{p\geq p_c}(pH-\eta(p)+d), \tag{6}$$

where p_c is the only critical point such that $\eta(p)=d$.

In this paper we will study the validity of the Frisch–Parisi conjecture for almost every function in the prevalence setting. The aim of this paper is to prove the following theorem.

Theorem 8. *Let η be a strongly admissible function, and let V be the space defined by*

$$V=\bigcap_{\varepsilon>0,0<p<\infty} B^{(\eta(p)-\varepsilon)/p,p}_{p,loc}; \tag{7}$$

then, in the sense of prevalence, almost every function f in V satisfies the following two conditions:

1. *For all* $p > 0$,

$$\eta_f(p) = \eta(p).$$

2. *The spectrum of singularities is defined on the interval* $[s(0), \frac{d}{p_c}]$ *and is given by*

$$d_f(H) = \inf_{p \geq p_c} (pH - \eta_f(p) + d), \tag{8}$$

where p_c *is the only critical point such that* $\eta(p_c) = d$.

REMARK 1. *We require that* η *is strongly admissible. Otherwise, according to* [8], *almost every function in* V *is nowhere locally bounded.*

In section 2 we will solve a simpler problem. We will prove that almost every function in a given intersection of a Sobolev or a Besov space and a Hölder space satisfies a slight modification of the Frisch–Parisi conjecture. We will first establish their spectrum of singularities.

THEOREM 9. *If* $\gamma > 0$ *and* $s - \frac{d}{p} < 0$, *the spectrum of singularities of almost every function in* $B_p^{s,q} \bigcap C^\gamma$ *or in* $L^{p,s} \bigcap C^\gamma$ *is given by*

$$d(H) = \begin{cases} \frac{d+(\gamma-s)p}{\gamma} H & \text{if } H \in \left[\gamma, \frac{d\gamma}{d+(\gamma-s)p}\right], \\ -\infty & \text{otherwise.} \end{cases}$$

REMARK 2. *Using the Sobolev embeddings* $B_p^{s,1} \hookrightarrow L^{p,s} \hookrightarrow B_p^{s,\infty}$, *the same result holds in Sobolev and in Besov spaces. As Besov spaces have a very simple wavelet characterization, we will prove the result in these spaces only. To obtain the Sobolev case, we only need to set* $q = \infty$ *in the following.*

In Theorem 9 we state only the spectrum of singularities of functions in the case $B_p^{s,q} \cap C^\gamma$ where $s - \frac{d}{p} < 0$. Other cases are proven in [8]. Moreover we recall the following result from [8].

PROPOSITION 10.

(i) *If* $s - d/p < 0$, *then almost every function in* $L^{p,s}$ *or in* $B_p^{s,q}$ *is nowhere locally bounded, and therefore its spectrum of singularities is not defined.*

(ii) *If* $s - d/p > 0$, *then the Hölder exponent of almost every function* f *of* $L^{p,s}$, *or of* $B_p^{s,q}$, *takes values in* $[s - d/p, s]$ *and*

$$\forall H \in [s - d/p, s], \qquad d_f(H) = Hp - sp + d; \tag{9}$$

furthermore, for almost every x, $h_f(x) = s$.

Our purpose here is to expand the result of [8] by taking an intersection of Besov spaces.

The main tool that we will use in the following is the wavelet expansion of functions. First, it yields a simple characterization of functional spaces and offers a simple condition for pointwise regularity. Let us recall some properties of the wavelet expansion.

There exist $2^d - 1$ oscillating functions $(\psi^{(i)})_{i \in \{1, \ldots, 2^d - 1\}}$ in the Schwartz class such that the functions

$$2^{dj} \psi^{(i)}(2^j x - k), \quad j \in \mathbb{Z}, \ k \in \mathbb{Z}^d,$$

form an orthonormal basis of $L^2(\mathbb{R}^d)$; see [18]. Wavelets are indexed by dyadic cubes $\lambda = (\frac{k}{2^j}; \frac{k+1}{2^j})^d$. Thus, any function $f \in L^2$ can be written as

$$f(x) = \sum c_{j,k}^{(i)} \psi^{(i)}(2^j x - k),$$

where

$$c_{j,k}^{(i)} = 2^{dj} \int f(x) \psi^{(i)}(2^j x - k) dx.$$

(Note that we use an L^∞ normalization instead of an L^2 one, which simplifies the formulas.) If $p > 1$ and $s > 0$, the Sobolev space has thus the following characterization (see [18]):

$$f \in L^{p,s} \Leftrightarrow \left(\sum_{\lambda \in \Lambda} |c_\lambda|^2 (1 + 4^{js}) \chi_\lambda(x) \right)^{1/2} \in L^p(\mathbb{R}^d), \tag{10}$$

where $\chi_\lambda(x)$ denotes the characteristic function of the cube λ and Λ is the set of all dyadics cubes. Homogeneous Besov spaces, which will also be considered, are characterized (for $p, q > 0$ and $s \in \mathbb{R}$) by

$$f \in B_p^{s,q} \iff \sum_j \left(\sum_{\lambda \in \Lambda_j} |c_\lambda|^p 2^{(sp-d)j} \right)^{q/p} \leq C, \tag{11}$$

where Λ_j denotes the set of dyadics cubes at scale j; see [18]. Note that if $p \in (0, 1)$, Besov spaces are not Banach spaces since they are not locally convex but nonetheless are separable complete metric vector spaces.

Pointwise regularity can also be expressed in terms of a condition on wavelet coefficients; see [14].

Proposition 11. *Let x be in $\mathbb{R}^d$. If f is in $C^\alpha(x)$, then there exists $c > 0$ such that for all λ,*

$$|c_\lambda| \leq c 2^{-\alpha j} (1 + |2^j x - k|)^\alpha. \tag{12}$$

In the following we will also call the "cone of influence above x_0 of width L" the set of couples (j, k) (or of cubes λ) such that

$$\left| k - 2^j x_0 \right| \leq L$$

(we use the norm on $\mathbb{R}^d$: $|x| = \sup_{i=1,\dots,d} |x_i|$).

2. Multifractal formalism in a given Besov space. The Frisch–Parisi conjecture gives the spectrum of singularities as the Legendre transform of the scaling function. We will determine the validity of this formula for measure-theoretic generic functions in a given Besov space in two steps. First we will prove Theorem 9, which gives the spectrum of singularities of almost every function. Afterwards, we will give the prevalent scaling function. This allows us to compare the spectrum obtained with the one given by formula (1) applied to the scaling function.

2.1. Proof of Theorem 9. Proposition 10 states that if $s-\frac{d}{p}<0$, almost every function in $B_p^{s,q}$ is nowhere locally bounded and the spectrum of singularities is not defined for any H. To define this spectrum, we need to assume a minimum uniform regularity. That is why, in the following, we choose $s-\frac{d}{p}<0$ and $0<\gamma<s$ and we study almost every function in $B_p^{s,q}\cap C^\gamma$.

Theorem 2.1 from [13] yields an upper bound of the spectrum of singularities.

LEMMA 12. *Let $s-\frac{d}{p}<0$. For all functions $f\in B_p^{s,q}\bigcap C^\gamma$, the Hausdorff dimension of the set $\{x : f\notin C^\alpha(x)\}$ is bounded by $\frac{d+(\gamma-s)p}{\gamma}\alpha$.*

We also need the following definition.

DEFINITION 13. *Let $\alpha\in[1,\frac{d}{d+(\gamma-s)p}]$. A point x_0 belongs to J_α if there exists an infinite sequence $(j,k)\in\mathbb{N}\times\{0,\dots,2^j-1\}^d$, $k=(k_1,\dots,k_d)$, such that for each $i=1,\dots,d$, k_i can be written as l_i2^{j-L} and*

$$\frac{1}{2^j}+\left|x_0-\frac{k}{2^j}\right|<\frac{1}{2^{\alpha L}}, \tag{13}$$

where $L:=[\frac{(d+(\gamma-s)p)j}{d}]$. We define the exponent of approximation of x as $\alpha'(x)=\sup\{\alpha : x\in J_\alpha\}$.

In [15], it is proved that the Hausdorff dimension of J_α is $\frac{d}{\alpha}$.

Let $\alpha\in[1,\frac{d}{d+(\gamma-s)p}]$, $\varepsilon>0$, and $n\in\mathbb{N}$ such that $N=2^{dn}>\frac{d}{\varepsilon\alpha}+1$ is fixed. We denote $H(\alpha)=\frac{d\gamma}{\alpha(d+(\gamma-s)p)}$ and $\beta(\alpha)=H(\alpha)+\varepsilon$. Each dyadic cube of size 2^{-dj} can be split into 2^{dn} subcubes $i(\lambda)$ with side $2^{-(j+n)}$. We define the probe P spanned by N functions g^r with the following wavelet coefficients d_λ^r:

$$d_\lambda^r=\begin{cases} j^{-2/q}2^{-\gamma j} & \text{if each } k_i \text{ is a multiple of } 2^{j-L} \text{ and } r=i(\lambda),\\ 0 & \text{elsewhere,}\end{cases} \tag{14}$$

where for each j we denote $L=[\frac{(d+(\gamma-s)p)j}{d}]$.

One can check that these functions g^r belong to $B_p^{s,q}\cap C^\gamma$; see [16].

Let $J_\alpha(i,l)=\frac{l}{2^i}+[-\frac{1}{2^{\alpha L}},\frac{1}{2^{\alpha L}}]^d$.

Let us first check that the set of points $S_c(\alpha)$ defined by

$$S_c(\alpha)=\left\{f=\sum c_\lambda\psi_\lambda\in B_p^{s,q}\cap C^\gamma :\ \exists x\in J_\alpha\ \ \forall j,k\ \ |c_\lambda|\le c2^{-\beta(\alpha)j}(1+|2^jx-k|)^{\beta(\alpha)}\right\}$$

is a Borel Haar-null set. Indeed this set can be included in the lim sup on i of the countable union over l of sets:

$$\begin{aligned} S_c(\alpha)^{i,l}=\Big\{f=\sum c_\lambda\psi_\lambda\in B_p^{s,q}\cap C^\gamma :\ \exists x\in J_\alpha(i,l)\ \ \forall j,k\ \ |c_\lambda|\\ \le c2^{-\beta(\alpha)j}(1+|2^jx-k|)^{\beta(\alpha)}\Big\}. \end{aligned} \tag{15}$$

To prove that the sets $S_c(\alpha)^{i,l}$ are closed, we pick a sequence of functions f_n in $S_c(\alpha)^{i,l}$. Suppose that f_n converges to f in $B_p^{s,q}\cap C^\gamma$. For each n, there exists x_n in $J_\alpha(i,l)$ such that f_n satisfies condition (12) at x_n. But $J_\alpha(i,l)$ is a compact set, so there exists x and a subsequence $(x_{n(i)})_{i\in\mathbb{N}}$ such that $x_{n(i)}$ converges to x. As the mapping that gives wavelet coefficients of a function is continuous, f satisfies also (12) at x. This means that $S_c(\alpha)^{i,l}$ are closed.

Let $f\in B_p^{s,q}\cap C^\gamma$ be fixed. Consider the affine subset $M=\{\delta\in\mathbb{R}^N;\ \ f+\sum\delta^ig^i\in S_c(\alpha)\}$. Let δ_1 and δ_2 be in M. There exists $x_1\in J_\alpha$ and $x_2\in J_\alpha$ such that for $l=1,2$,

$$\left|c_\lambda+\sum\delta_l^id_\lambda^i\right|\le c2^{-\beta(\alpha)j}(1+|2^jx_l-k|)^{\beta(\alpha)}\le c2^{-\alpha\beta(\alpha)L}. \tag{16}$$

Furthermore if λ is such that each k is a multiple of 2^{j-L},

$$|d^i_\lambda| > \frac{1}{j^{2/q}} 2^{-\alpha H(\alpha) L}. \tag{17}$$

So, by taking (16) and (17) we obtain

$$\|\delta_1 - \delta_2\|_{\mathbb{R}^N} \leq 2c2^{-\alpha\beta(\alpha)L} 2^{\alpha H(\alpha)L} j^{2/q} = 2cj^{2/q} 2^{-\alpha\varepsilon L}.$$

When j tends to infinity, the Lebesgue measure of $S_c(\alpha)$ tends to zero.

Now, we take the countable union over c and $\varepsilon_n \to 0$. As Haar-null sets are stable under inclusion, we obtain

$$\forall \alpha \in \left[1, \frac{d}{d+(\gamma-s)p}\right] \ \text{ a.e. in } B^{s,q}_p \cap C^\gamma \ \ \forall x \in J_\alpha \ h_f(x) \leq H(\alpha).$$

Let (α_n) be a dense sequence in $[1, \frac{d}{d+(\gamma-s)p}]$. As a countable union of Haar-null sets is still a Haar-null set, for almost every function in $B^{s,q}_p \cap C^\gamma$,

$$h_f(x) \leq H(\alpha_n) \quad \forall n \, \forall x \in J_{\alpha_n}. \tag{18}$$

Let f be a function satisfying (18) and let α be fixed. There exists a nondecreasing subsequence (α_{φ_n}) which converges to α, and the intersection of the subsets $J_{\alpha_{\varphi_n}}$ $(:= \tilde{J}_\alpha)$ contains J_α. Furthermore there exists a measure such that any set of dimension less than d/α is of measure zero. And the measure of J_α is positive. If $G_H = \{x : \ h_f(x) \leq H\}$, with Lemma 12 we have that the Hausdorff dimension of G_H is $\frac{d+(\gamma-s)p}{\gamma}H$. And the $\frac{d}{\alpha}$ Hausdorff measure of the set $\{x : \ h_f(x) < H\}$ equals zero. Thus for almost every function in $B^{s,q}_p \cap C^\gamma$,

$$d(H) = \frac{d+(\gamma-s)p}{\gamma} H \quad \text{for } H \in \left[\gamma, \frac{d\gamma}{d+(\gamma-s)p}\right].$$

2.2. The scaling function. Let us now determine the scaling function of almost every function in a given Besov space. We will now show the following result.

Proposition 14. *Let s_0 and p_0 be fixed such that $s_0 - \frac{d}{p_0} > 0$. Outside a Haar-null set in $B^{s_0,\infty}_{p_0}$, we have*

$$\eta_f(p) = \begin{cases} ps_0 & \text{if } p \leq p_0, \\ d + p(s_0 - \frac{d}{p_0}) & \text{if } p \geq p_0. \end{cases} \tag{19}$$

Let $0 < \gamma < s_0$ be fixed. If $s_0 - \frac{d}{p_0} < 0$, then outside a Haar-null set in $B^{s_0,p_0}_{p_0} \bigcap C^\gamma$,

$$\eta_f(p) = \begin{cases} ps_0 & \text{if } p \leq p_0, \\ \gamma p + p_0(s_0 - \gamma) & \text{if } p \geq p_0. \end{cases} \tag{20}$$

Proof. In each case, we can find in [20] the lower bound. Indeed, this bound is given by the Sobolev embedding.

To prove the upper bound, we will first consider the case $s_0 - \frac{d}{p_0} > 0$. Let $\varepsilon > 0$ be fixed and denote

$$\tilde{s}(p) = \begin{cases} s_0 + \varepsilon, & p \leq p_0, \\ \frac{d}{p} + (s_0 - \frac{d}{p_0}) + \varepsilon, & p \geq p_0. \end{cases}$$

Let $0 < p_0 < \infty$ be fixed. We want to show that the set of functions belonging to $B_p^{\tilde{s}(p),\infty}$ for all $0 < p < \infty$ is Haar-null. This set is clearly closed and Borel in $B_{p_0}^{s_0,\infty}$. Let $j \geq 1$ and $k \in \{0, \dots, 2^j - 1\}^d$. We define $J \leq j$ and $K \in \mathbb{Z}^d$ such that

$$\frac{K}{2^J} = \frac{k}{2^j}$$

is an irreducible fraction. Let $a > \frac{3}{p_0}$. We define a probe spanned by the function F with the following wavelet coefficients:

$$d_\lambda = j^{-a} 2^{(\frac{d}{p_0} - s_0)j} 2^{-\frac{d}{p_0}J}.$$

This function belongs to $B_{p_0}^{s_0,p_0}$.

Let f be in $B_{p_0}^{s_0,p_0}$ and consider the affine subset

$$M = \{\alpha \in \mathbb{R};\ f + \alpha F \in B_p^{\tilde{s}(p),\infty}\}.$$

Suppose that there exist α_1 and α_2 in M. We then have three cases, following the position of p.

(i) If $p = p_0$, then $\tilde{s}(p) = p_0 + \varepsilon$ and

$$\begin{aligned}
\|f + \alpha_1 F - (f + \alpha_2 F)\|_{B_p^{\tilde{s}(p),\infty}} &= \sup_j \sum_{k \in \{0,\dots,2^j-1\}^d} \left| \frac{\alpha_1 - \alpha_2}{j^a} 2^{(\tilde{s} - \frac{d}{p_0})j} 2^{(\frac{d}{p_0} - s_0)j} 2^{-\frac{d}{p_0}J} \right|^{p_0} \\
&= \sup_j \left| \frac{\alpha_1 - \alpha_2}{j^a} \right|^{p_0} 2^{p_0 \varepsilon j} \sum_{J=0}^{j} \sum_{K \in \{0,\dots,2^J-1\}^d} 2^{-dJ} \\
&= \sup_j \frac{j+1}{2} \left| \frac{\alpha_1 - \alpha_2}{j^a} \right|^{p_0} 2^{p_0 \varepsilon j}.
\end{aligned}$$

But if α_1 and α_2 belong to M, this implies that $f + \alpha_1 F - (f + \alpha_2 F)$ belong to $B_p^{\tilde{s}(p),\infty}$. This is possible only if $\alpha_1 = \alpha_2$.

(ii) If $p > p_0$, then $\tilde{s}(p) = \frac{d}{p} + (s_0 - \frac{d}{p_0}) + \varepsilon$. In this case, $f + \alpha_1 F - (f + \alpha_2 F)$ belong to $B_p^{\tilde{s}(p),\infty}$, implying that there exist $c > 0$ such that

$$\|f + \alpha_1 F - (f + \alpha_2 F)\|_{B_p^{\tilde{s}(p),\infty}} \leq \|f + \alpha_1 F\|_{B_p^{\tilde{s}(p),\infty}} + \|f + \alpha_2 F\|_{B_p^{\tilde{s}(p),\infty}} \leq c.$$

We then have the following inequalities:

$$\begin{aligned}
&\forall j > 0 \quad \sum_{k \in \{0,\dots,2^j-1\}^d} \left| \frac{\alpha_1 - \alpha_2}{j^a} 2^{(\tilde{s} - \frac{d}{p})j} 2^{(\frac{d}{p_0} - s_0)j} 2^{-\frac{d}{p_0}J} \right|^p \leq c, \\
&\forall j > 0 \quad \left| \frac{\alpha_1 - \alpha_2}{j^a} \right|^p 2^{(\tilde{s} - \frac{d}{p})pj} 2^{(\frac{d}{p_0} - s_0)pj} \sum_{J=0}^{j} \sum_{K \in \{0,\dots,2^J-1\}^d} 2^{-\frac{dp}{p_0}J} \leq c, \\
&\forall j > 0 \quad \left| \frac{\alpha_1 - \alpha_2}{j^a} \right|^p \sum_{J=0}^{j} 2^{(d - \frac{dp}{p_0})J} \leq c 2^{(-\tilde{s} + \frac{d}{p} - \frac{d}{p_0} + s_0)pj}, \\
&\forall j > 0 \quad \left| \frac{\alpha_1 - \alpha_2}{j^a} \right| \leq c 2^{(-\tilde{s} + \frac{d}{p} - \frac{d}{p_0} + s_0)j} \left| \frac{1}{1 - 2^{j(d - d\frac{p}{p_0})}} \right|^{\frac{1}{p}}.
\end{aligned} \tag{21}$$

As $p > p_0$, $1 - 2^{j(d-d\frac{p}{p_0})}$ is equivalent to 1 and (21) implies

$$|\alpha_1 - \alpha_2| \le cj^a 2^{-\varepsilon j},$$

which tends to zero when j tends to infinity.

(iii) If $p < p_0$, then $\tilde{s}(p) = s_0 + \varepsilon$ and $|1 - 2^{j(d-d\frac{p}{p_0})}|$ is equivalent to $2^{j(d-d\frac{p}{p_0})}$. Thus in (21), we obtain again

$$|\alpha_1 - \alpha_2| \le cj^a 2^{-\varepsilon j}.$$

In each case we have obtained that M is of Lebesgue measure zero. Taking a countable union over $\varepsilon \to 0$, and over p, we obtain the desired scaling exponent.

The second case, for $s_0 - \frac{d}{p_0} < 0$, can be treated the same way as for $p \le p_0$. For the case $p > p_0$ a modification is made by taking, instead of F, the function G with wavelet coefficients $d_{j,k}$ given by

$$d_{j,k} = j^{-2/q} 2^{-\gamma j} \text{ if each } k_i \text{ is a multiple of } 2^{j-L},$$

where for each j we take $L = [\frac{(d+(\gamma-s)p)j}{d}]$. □

From Theorem 9 and Proposition 14, we obtain the following Legendre transform of the scaling function of almost every function in a given Besov space.

PROPOSITION 15. *Let $s_0 > 0$ and $0 < p_0 < \infty$.*

(i) *If $s_0 - \frac{d}{p_0} > 0$, then for almost every function in $B_p^{s,q}$,*

$$\forall H \in \left[s_0 - \frac{d}{p_0}, s_0\right] \qquad \inf_{p>0}(d - \eta(p) + Hp) = d - p_0 s_0 + Hp_0. \tag{22}$$

(ii) *If $s_0 - \frac{d}{p_0} < 0$, then for almost every function in $B_p^{s,q} \bigcap C^\gamma$, we have*

$$\forall H \in [\gamma, s_0] \qquad \inf_{p>0}(d - \eta(p) + Hp) = d - p_0 s_0 + Hp_0. \tag{23}$$

This proposition shows that for $s_0 - \frac{d}{p_0} > 0$, the increasing part of the spectrum given by Frisch–Parisi conjecture is valid for almost every function. But for $s_0 - \frac{d}{p_0} < 0$, this Legendre transform does not correspond to the spectrum of singularities given by Theorem 9.

3. The Frisch–Parisi conjecture. We will now prove Theorem 8. Instead of $B_{p_0}^{s_0,q_0}$ we will now work with

$$V = \bigcap_{\varepsilon>0, 0<p<\infty} B_{p,loc}^{(\eta(p)-\varepsilon)/p,p}.$$

This set V can also be written as a countable intersection over $B_{p_n,loc}^{(\eta(p_n)-\varepsilon_n)/p_n,p_n}$.

Note that V is a topological vector space. For $p < 1$ Besov spaces are only quasi-Banach spaces; as the triangle inequality is satisfied only up to a constant, V is not a Banach space but a complete metric space. Indeed, if $p \ge 1$, we take for distance between two functions f and g in $B_p^{s,q}$

$$d(f,g) = \sum_{j\ge0} \left(\sum_{k\in\{0,\dots,2^j-1\}^d} \left|(c_{j,k} - d_{j,k}) 2^{(s-\frac{d}{p})j}\right|^p \right)^{\frac{q}{p}},$$

where $c_{j,k}$ are the wavelet coefficients of f, and $d_{j,k}$ are those of g.

If $p < 1$, Besov spaces are not Banach spaces, but complete metric spaces with the following distance:

$$d(f,g) = \left(\sum_{j\geq 0} \left(\sum_{k\in\{0,\dots,2^j-1\}^d} |(c_{j,k} - d_{j,k})2^{(s-\frac{d}{p})j}|^p \right)^{\frac{q}{p}} \right)^{\frac{\min(p,q)}{q}}.$$

Thus, we obtain a distance in V by taking

$$\forall f,g \in V \quad d(f,g) = \sum_n 2^{-n} \frac{d_n(f,g)}{1+d_n(f,g)},$$

where d_n denotes the distance in $B^{(\eta(p_n)-\varepsilon_n)/p_n,p_n}_{p_n,loc}$. With this distance, V is clearly a complete space. Note that the measure used is the Lebesgue measure on the unit ball of a probe, so this is a probability measure with compact support.

In the following subsection we prove that the spectrum of singularities of almost every function in V satisfies

$$d(H) = \inf_{p\geq p_c} (pH - \eta(p) + d).$$

3.1. Proof of Theorem 8. Let us now study the spectrum of singularities on a prevalent set of functions in V.

PROPOSITION 16. *For almost every function $f \in V$, the spectrum of singularities satisfies*

$$\forall H \in \left[s(0), \frac{d}{p_c} \right] \quad d(H) = \inf_{p\geq p_c} (Hp - \eta(p) + d). \tag{24}$$

Proof. We will first construct the probe. Denote

$$a(j,k) = \inf_p \left(\frac{d(j-J) - \eta(p)j}{p} \right)$$

and define g via its wavelet coefficients,

$$d_\lambda = \frac{1}{j^a} 2^{a(j,k)}, \tag{25}$$

where we define $a = a_j = \log j$, and $J \leq j$ is such that there exists $K \in \mathbb{Z}^d$ and $\frac{k}{2^j} = \frac{K}{2^J}$ is an irreducible form.

First, we check that g belongs to V. Let $p > 0$ be fixed. Thus we have to show that $g \in B^{\eta(p)/p,\infty}_p$. Let $s = \frac{\eta(p)}{p}$. Since $a(j,k) \leq \frac{d(j-J)}{p} - sj$, $pa(j,k) + (\eta(p) - p)j = -Jd$ and $g \in B^{\eta(p)/p,\infty}_p$. For further details on this function g, the reader is referred to [16].

DEFINITION 17. *Let α be fixed. We denote by F_α the set of α-approximable points defined by*

$$F_\alpha = \left\{ x : \exists \text{ a sequence } ((k_n,j_n))_{n\in\mathbb{N}} \ \left| x - \frac{k_n}{2^{j_n}} \right| \leq \frac{1}{2^{\alpha j_n}} \right\}. \tag{26}$$

The dyadic exponent of x is defined by $\alpha(x_0) = \sup\{\alpha : x_0$ is α-approximable by dyadics$\}$.

As stated in [16], the Hausdorff dimension of the set F_α is at least $\frac{d}{\alpha}$.

First, let $\alpha \in (1, \infty)$ be fixed, and let F_α be the set given by Definition 17. Let $\varepsilon > 0$ be fixed, and let

$$H(\alpha) = \frac{1}{\alpha} \sup_{\omega \geq \alpha} \left(\omega \sup_{q>0} \left(s(q) - d\left(1 - \frac{1}{\omega}\right) q \right)\right)$$

and $\gamma = \gamma(\alpha) = H(\alpha) + \varepsilon$.

Let $n \in \mathbb{N}$ be such that $N = 2^{dn} > \frac{d}{\varepsilon} + 1$ is fixed. The probe P is spanned by N functions g_i which are deduced from g by taking its wavelet coefficients only over some subcubes $i(\lambda)$ with size $2^{-d(j+n)}$. The aim of this part is to prove that the set of functions f such that there exists a point in F_α where f is C^γ is a Haar-null set. This set is included in the countable union over $c > 0$ of

$$S_c(\alpha) = \left\{ f = \sum c_\lambda \psi_\lambda : \ \exists x \in F_\alpha \ \forall j, k \ |c_\lambda| \leq c 2^{-\gamma(\alpha) j} (1 + |2^j x - k|)^{\gamma(\alpha)} \right\}.$$

And we have already seen that

$$H(\alpha) = \frac{1}{\alpha} \sup_{\omega \geq \alpha} \left(\omega \sup_{q>0} \left(s(q) - d\left(1 - \frac{1}{\omega}\right) q \right)\right) \geq -a(j,k)$$

for some j and k.

Let $x \in F_\alpha$ be fixed and let λ be such that $|x - \lambda| \leq A$. If $A > 2N$, for all functions g_i, the wavelet coefficients indexed by those λ satisfy

$$|d_\lambda^i| \geq \frac{c(A)}{j^a} 2^{-H(\alpha) j}. \tag{27}$$

We will now prove that the set $S_c(\alpha)$ is a Borel Haar-null set. First, this set is included in the countable union over (i, l) of

$$S_c(\alpha)^{i,l} = \left\{ f = \sum c_\lambda \psi_\lambda : \ \exists x \in F_\alpha^{i,l} \ \forall j, k \ |c_\lambda| \leq c 2^{-\gamma(\alpha) j} (1 + |2^j x - k|)^{\gamma(\alpha)} \right\},$$

where $F_\alpha^{i,l} = \{x : \ \left|x - \frac{l}{2^i}\right| \leq \frac{1}{2^{\alpha i}}\}$. This set $S_c(\alpha)^{i,l}$ is a closed set and $S_c(\alpha)$ is a Borel set. Let f be in V, and let β_1 and β_2 be such that the functions $f + \sum \beta_1^i g^i$ and $f + \sum \beta_2^i g^i$ are in $S_c(\alpha)^{i,l}$. There exist two points x_1 and x_2 in $F_\alpha^{i,l}$ such that in the cone of influence above x_1 and x_2,

$$\left| c_\lambda + \sum \beta_1^i d_\lambda^i - \left(c_\lambda + \sum \beta_2^i d_\lambda^i \right) \right| \leq 2c 2^{-\gamma(\alpha) j}.$$

However,

$$\left| c_\lambda + \sum \beta_1^i d_\lambda^i - \left(c_\lambda + \sum \beta_2^i d_\lambda^i \right) \right| = \left| \sum \beta_1^i d_\lambda^i - \beta_2^i d_\lambda^i \right|$$

and with (27),

$$\left| \sum \beta_1^i d_\lambda^i - \beta_2^i d_\lambda^i \right| \geq \left| \sum \beta_1^i - \beta_2^i \right| \frac{c(A)}{j^a} 2^{-H(\alpha) j}.$$

Thus,

$$\|\beta_1 - \beta_2\|_{\mathbb{R}^N} \leq \tilde{c} j^a 2^{-\varepsilon j}.$$

So the Lebesgue measure in $\mathbb{R}^N$ of the set of β such that $f + \beta g$ belongs to $S_c(\alpha)^{i,l}$ is bounded by $(\tilde{c} j^a)^N 2^{-N\varepsilon j}$.

The Lebesgue measure of the set of β such that $f + \sum \beta^i g^i$ belongs to $S_c(\alpha)$ vanishes. Therefore $S_c(\alpha)$ is Haar-null.

By taking a countable union over $c_n > 0$ of sets $S_c(\alpha)$, the set of functions in V with a pointwise Hölder exponent greater than $\gamma(\alpha)$ in a point of F_α is also Haar-null. If $\varepsilon_n \to 0$, then by taking the union over ε_n it follows that for all $\alpha \geq 1$ the set of functions in V with a Hölder exponent greater than $H(\alpha)$ in some point of F_α is Haar-null.

Let α_n be a dense sequence in $(1, \infty)$. By a countable intersection

$$M = \{f \in V : \forall n \ \ \forall x \in F_{\alpha_n} \ \ h_f(x) \leq H(\alpha)\} \tag{28}$$

is prevalent. Let $f \in M$ and let $\alpha \geq 1$. There exists a subsequence $\alpha_{\phi(n)}$ which is nondecreasing and tends to α. If we denote by $\tilde{F}_\alpha$ the intersection of sets F_{α_n}, then it follows that $\tilde{F}_\alpha$ contains F_α. Furthermore, the Hausdorff dimension of $\tilde{F}_\alpha$ is greater than $\frac{d}{\alpha}$ and for all $x \in \tilde{F}_\alpha$, $h_f(x) \leq H(\alpha)$.

To conclude the second point of Theorem 1, we rewrite $H(\alpha)$ in the following form:

$$H(\alpha) = \frac{1}{\alpha} \inf_{a \geq \alpha} G(a),$$

where $G(a) = \sup_q (a(-qd + s(q)) + qd) = a \sup_q (qd(-1 + \frac{1}{a}) + s(q)) = a\ s^*\left(d\left(1 - \frac{1}{a}\right)\right)$. Here s^* is the Legendre transform of s. By definition of the Legendre transform, this is a convex function. Furthermore it satisfies

$$\begin{cases} s^*(h) = +\infty & \text{if } h < s'(+\infty), \\ s^*(h) = s(0) & \text{if } h > s'(0). \end{cases} \tag{29}$$

And if s^* is twice differentiable (we refer to [10] for a general case), G is also twice differentiable and its derivative is

$$G'(a) = s^*\left(d\left(1 - \frac{1}{a}\right)\right) + \frac{d}{a}(s^*)'\left(d\left(1 - \frac{1}{a}\right)\right)$$

and

$$G''(a) \geq 0.$$

Thus G is also convex and there exists a_0 such that $G(a_0) = \inf_{a \geq 0} G(a)$, a_0 being such that $G'(a_0) = 0$. We also deduce from (29) that

$$\begin{cases} G(a) = +\infty & \text{if } a < \frac{d}{d - s'(+\infty)}, \\ G(a) = as(0) & \text{if } a > \frac{d}{d - s'(0)}. \end{cases} \tag{30}$$

By definition of s and with the hypothesis that η is an admissible function we have $0 \leq s'(q) \leq d$ for all $q > 0$. It follows from (30) that a_0 belongs to the interval $(\frac{d}{d - s'(+\infty)}, \frac{d}{d - s'(0)})$, which is included in $[0, \infty)$.

Another way to treat G is to write $G(a) = \sup_q(\tilde{s}(q))$, where $\tilde{s}(q) = a(s(q) - qd) + qd$. And if s is also twice differentiable, $\tilde{s}'(q) = -da + as'(q) + d$ and $\tilde{s}''(q) = as''(q) < 0$. Thus $\tilde{s}$ is a concave function, and there exists an upper bound q_0 which satisfies $\tilde{s}'(q_0) = -da + as'(q_0) + d = 0$, and $s'(q_0) = \frac{da-d}{a}$. The value of q_0 also depends on a, so we can now write $q_0 = q(a)$.

We can finally write $G(a) = a(-q(a)d + s(q(a))) + q(a)d$. This function is twice differentiable and its derivative satisfies

$$G'(a) = s(q(a)) - dq(a).$$

If $a = a_0$ is the lower bound of G, we obtain $G'(a_0) = s(q(a_0)) - dq(a_0) = 0$, so that $s(q(a_0)) = dq(a_0)$, that is, $q(a_0) = q_c = 1/p_c$. Furthermore, G is decreasing for $a \leq a_0$ and increasing for $a \geq a_0$. The following cases are now possible:

(i) If $\alpha \geq \frac{d}{d-s'(0)}$, then for all $a \geq \alpha$, $G(a) = as(0)$ and $H(\alpha) = s(0)$. So,

$$\dim_H(\{x:\ h_f(x) \leq s(0)\}) = d - s'(0).$$

(ii) If $1 \leq \alpha \leq a_0$, then

$$\inf_{a\geq\alpha} G(a) = G(a_0) = (a_0(-q_c d + s(q_c)) + q_c d) = dq_c.$$

The corresponding value of H is

$$H(\alpha) = \frac{1}{\alpha}\inf_{a\geq\alpha} G(a) = \frac{dq_c}{\alpha}.$$

Thus, the spectrum of singularities is defined on the interval $[\frac{dq_c}{a_0}, dq_c]$ and for almost every function, and for all $H \in [\frac{dq_c}{a_0}, dq_c]$

$$\dim_H(\{x:\ h_f(x) \leq H\}) = \frac{H}{q_c}.$$

Furthermore, we have already seen that

$$s'(q(a)) = \frac{da-d}{a},$$

which is an increasing function. As s' is decreasing, the application $a \mapsto q(a)$ is itself decreasing. So, for $\alpha \leq a_0$, $q \geq q_c = \frac{1}{p_c}$ and

$$\dim_H(\{x:\ h_f(x) \leq H\}) \leq \inf_{p\geq p_c}(pH - \eta(p) + d).$$

(iii) If $a_0 \leq \alpha \leq \frac{d}{d-s'(0)}$, which is equivalent to

$$\inf_{a\geq\alpha} G(a) = G(\alpha) = \alpha\sup_q\left(-qd + s(q) + \frac{qd}{\alpha}\right),$$

we obtain

$$H(\alpha) = \sup_q\left(-qd + s(q) + \frac{qd}{\alpha}\right).$$

So for almost every function, for all $H \in [s(0), \frac{dq_c}{a_0}]$,

$$\dim_H(\{x:\ h_f(x) \leq H\}) = \inf_{p\geq p_c}(pH - \eta(p) + d) \leq \inf_{p\geq p_c}(pH - \eta(p) + d).$$

Furthermore (see [16]), the spectrum of singularities of all functions of V satisfies

$$d(H) \leq \inf_{p \geq p_c} (pH - \eta(p) + d). \tag{31}$$

This implies that the Hausdorff dimension of the set $\{x : h_f(x) < H\}$ is strictly less than $\frac{d}{\alpha}$. As proved in [14], there exists a measure m_α such that $m_\alpha(\{x : h_f(x) \leq H\}) > 0$. But by definition of the Hausdorff dimension, $m_\alpha(\{x : h_f(x) < H\}) = 0$. Then $m_\alpha(\{x : h_f(x) = H) > 0$ and

$$d(H) \geq \inf_{p \geq p_c} (pH - \eta(p) + d). \qquad \square$$

Proposition 18. *For almost every function f in V, the scaling function of f satisfies*

$$\eta_f(p) = \eta(p) \quad \forall 0 < p < \infty.$$

Proof. As we have $V = \bigcap_{\varepsilon>0, 0<p<\infty} B^{(\eta(p)-\varepsilon)/p,p}_{p,loc}$, for any $f \in V$ the scaling function is greater than $\eta(p)$ for all p. Let $\tau > 0$ and $p > 0$ be fixed. We denote $\tau(p) = \frac{\eta(p)}{p} + \tau$. We first prove that the set

$$M(p) = \{f \in V;\ f \in B^{\tau(p),\infty}_p\}$$

is a Haar-null set. Let g be the function with wavelet coefficients given by (25), and let P be the probe spanned by g. First, we check that g does not belong to $B^{\tau(p),\infty}_p$. We write $\beta_j = d(1 - \frac{j}{J})$, where J is defined as in (25). This term β_j takes discrete values, spaced by $\frac{d}{j}$ and between 0 and d. As s is a concave function and $0 \leq s'(q) \leq d$ for all q, there exists, for j large enough, a β_j close to $s'(q)$ such that the line given by $\tau(p) + \beta_j(\frac{1}{\tilde{p}} - \frac{1}{p})$ is always above the graph of s. Thus

$$\forall \tilde{p} > 0 \quad \tau(p) + \beta_j \left(\frac{1}{\tilde{p}} - \frac{1}{p}\right) > s(1/p).$$

But $a(j,k) = j \inf_p(\frac{\beta_j}{p} - s(1/p))$, and this infimum is attained for a $p_0 \in (0,\infty)$. Therefore,

$$a(j,k) \geq j\left(\frac{\beta_j}{p} - \tau(p)\right)$$

and

$$\|g\|_{B^{\tau(p),\infty}_p} \geq \sup_j j^{-ap} 2^{\tau j p}.$$

Thus $g \notin B^{\tau(p),\infty}_p$.

Let f be in V. Suppose that there exist α_1 and α_2 such that the functions $f_1 = f + \alpha_1 g$ and $f_2 = f + \alpha_2 g$ belong to $B^{\tau(p),\infty}_p$. Then $f_1 - f_2$ also belongs to $B^{\tau(p),\infty}_p$. But

$$f_1 - f_2 = (\alpha_1 - \alpha_2)g,$$

which is possible only if $\alpha_1 = \alpha_2$. Thus $M(p)$ is Haar-null.

Taking countable unions over $\tau \to 0$ and $p > 0$, we obtain that for almost every f in V, $\eta_f(p) \leq \eta(p)$ for all $p > 0$. $\square$

REFERENCES

[1] P. Abry, *Ondelettes et turbulences. Multirésolutions, algorithmes de décomposition, invariance d'échelle et signaux de pression*, Nouveaux Essais, Diderot, Paris, 1997.

[2] A. Arnéodo, E. Bacry, and J. Muzy, *The thermodynamics of fractals revisited with wavelets*, Phys. A., 213 (1995), pp. 232–275.

[3] S. Banach, *Über die Baire'sche Kategorie gewisser Funktionenmengen*, Studia Math., 3 (1931), pp. 174–179.

[4] J. Barral and B. B. Mandelbrot, *Multifractal products of cylindrical pulses*, Probab. Theory Related Fields, 124 (2002), pp. 409–430.

[5] Y. Benyamini and J. Lindenstrauss, *Geometric Nonlinear Functional Analysis. Volume* 1, Amer. Math. Soc. Colloq. Publ. 48, AMS, Providence, RI, 2000.

[6] G. Brown, G. Michon, and J. Peyrière, *On the multifractal analysis of measures*, J. Statist. Phys., 66 (1992), pp. 775–790.

[7] J. Christensen, *On sets of Haar measure zero in Abelian Polish groups*, Israel J. Math., 13 (1972), pp. 255–260.

[8] A. Fraysse and S. Jaffard, *How smooth is almost every function in a Sobolev space?*, Rev. Mat. Iber. Amer., 22 (2006), pp. 663–682.

[9] U. Frisch and G. Parisi, *On the singularity structure of fully developed turbulence*, in Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics, M. Ghil, R. Benzi, and G. Parisi, eds., North-Holland, Amsterdam, 1985, pp. 84–88.

[10] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms. Part* 1*: Fundamentals*, Springer-Verlag, Berlin, 1993.

[11] B. Hunt, T. Sauer, and J. Yorke, *Prevalence: A translation invariant "almost every" on infinite dimensional spaces*, Bull. Amer. Math. Soc., 27 (1992), pp. 217–238.

[12] B. Hunt, *The prevalence of continuous nowhere differentiable function*, Proc. Amer. Math. Soc., 122 (1994), pp. 711–717.

[13] S. Jaffard and Y. Meyer, *Wavelet Methods for Pointwise Regularity and Local Oscillations of Functions*, Mem. Amer. Math. Soc., 123 (1996).

[14] S. Jaffard, *Multifractal formalism for functions Part* I: *Results valid for all functions*, SIAM J. Math. Anal., 28 (1997), pp. 944–970.

[15] S. Jaffard, *Old friends revisited: The multifractal nature of some classical functions*, J. Fourier Anal. Appl., 3 (1997), pp. 1–22.

[16] S. Jaffard, *On the Frisch-Parisi conjecture*, J. Math. Pures Appl., 79 (2000), pp. 525–552.

[17] A. Kolmogoroff, *The local structure of turbulence in incompressible viscous fluid for every large Reynold's numbers*, C. R. (Doklady) Acad. Sci. URSS (N.S.), 30 (1941), pp. 301–305.

[18] Y. Meyer, *Ondelettes et opérateurs*, Hermann, Paris, 1990.

[19] L. Olsen, *A multifractal formalism*, Adv. Math., 116 (1995), pp. 82–196.

[20] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 608–626

CONTINUITY ESTIMATES FOR THE MONGE–AMPÈRE EQUATION*

HUAI-YU JIAN† AND XU-JIA WANG‡

Abstract. In this paper, we study the regularity of solutions to the Monge–Ampère equation. We prove the log-Lipschitz continuity for the gradient under certain assumptions. We also give a unified treatment for the continuity estimates of the second derivatives. As an application we show the local existence of continuous solutions to the semigeostrophic equation arising in meteorology.

Key words. Monge–Ampère equation, a priori estimates, semigeostrophic equation

AMS subject classifications. 35J60, 35Q35

DOI. 10.1137/060669036

1. Introduction.

In this paper we study the regularity of solutions to the Monge–Ampère equation

$$\det D^2 u = f \quad \text{in} \quad B_1(0), \tag{1.1}$$

where $B_1(0)$ is the unit ball in the Euclidean space R^n. We are mainly concerned with the log-Lipschitz continuity of the gradient Du,

$$|Du(x) - Du(y)| \le C|x-y|(1 + |\log|x-y||), \quad x, y \in B_{1/2}(0), \tag{1.2}$$

which has applications in the existence of continuous solutions to the semigeostrophic equation [E, L] or, more generally, in the optimal transportation problems [E, V]. We also study the continuity estimates of the second derivatives D^2u under appropriate conditions.

For the regularity of the Monge–Ampère equation, Caffarelli [C2] established the interior $W^{2,p}$ estimates (for any $p > 1$) for strictly convex solutions when f is positive and continuous. He also obtained the $C^{2,\alpha}$ estimate when $f > 0$, $f \in C^\alpha$, $\alpha \in (0, 1)$. In [C3] he proved the $C^{1,\alpha}$ estimate for strictly convex solutions if $C_1 \le f \le C_2$ for some positive constants C_1, C_2. By an example in [W2], the $C^{1,\alpha}$ regularity cannot be improved to $W^{2,p}$ for large p if f is not continuous.

For the Laplace equation

$$\Delta u = f, \tag{1.3}$$

the log-Lipschitz continuity of Du was established [Y] for $f \in L^\infty$; see also Theorem 3.9 in [GT]. The log-Lipschitz continuity plays a key role in the existence and uniqueness of global solutions to the 2-dimensional Euler equation [Y]. A simple proof of the log-Lipschitz continuity was recently found by the second author [W3]. Considering applications to the semigeostrophic equation [BB, C5, CuF, CRD, E, L], one wishes to

*Received by the editors September 4, 2006; accepted for publication (in revised form) March 21, 2007; published electronically August 1, 2007. This work was supported by the NSF of China grants 10631020 and 10428103 and Australian Research Council grant DP0664517.
http://www.siam.org/journals/sima/39-2/66903.html

†Department of Mathematics, Tsinghua University, Beijing 100084, China (hjian@math.tsinghua.edu.au).
‡Centre for Mathematics and Its Applications, Australian National University, Canberra, ACT 0200, Australia (wang@maths.anu.edu.au).

know when a solution to the Monge–Ampère equation (1.1) satisfies the log-Lipschitz continuity. By an example in [W2], the condition $C_1 \le f \le C_2$ is not enough; a stronger condition is necessary.

In this paper we first give a unified treatment for the continuity estimates of the second derivatives of solutions to the Monge–Ampère equation. Before stating our results, we first introduce the modulus of convexity for a convex function u, which is given by

$$m(t) = \inf\{u(x) - \ell_z(x) : \quad |x - z| > t\}, \tag{1.4}$$

where $t > 0$, ℓ_z is the tangent plane of u at z. Obviously m is a nonnegative function of $t > 0$. When u is strictly convex, it is a positive function.

We also denote

$$\omega_f(r) = \sup\{|f(x) - f(y)| : \quad |x - y| < r\}. \tag{1.5}$$

We say f is Dini continuous if

$$\int_0^1 \frac{\omega_f(r)}{r} dr < \infty. \tag{1.6}$$

THEOREM 1. *Let $u \in C^2$ be a strictly convex solution of* (1.1)*. Assume that f satisfies* (1.6) *and*

$$C_1 \le f \le C_2 \tag{1.7}$$

for some positive constants $C_1, C_2 > 0$. Then for all $x, y \in B_{1/2}(0)$, we have the estimate

$$|D^2u(x) - D^2u(y)| \le C\left[d + \int_0^d \frac{\omega_f(r)}{r} + d\int_d^1 \frac{\omega_f(r)}{r^2}\right], \tag{1.8}$$

where $d = |x - y|$, $C > 0$ depends only on n, m, and C_1, C_2. It follows that

(i) *if f is Dini continuous, then $u \in C^2(B_1)$;*

(ii) *if $f \in C^\alpha(B_1)$ and $\alpha \in (0, 1)$, then*

$$\|u\|_{C^{2,\alpha}(B_{1/2})} \le C\left[1 + \frac{\|f\|_{C^\alpha(B_1)}}{\alpha(1-\alpha)}\right]; \tag{1.9}$$

(iii) *if $f \in C^{0,1}(B_1)$, then*

$$|D^2u(x) - D^2u(y)| \le Cd\left[1 + \|f\|_{C^{0,1}}|\log d|\right]. \tag{1.10}$$

Note that in the estimate (1.8), the constant C depends also on $\sup_{B_1}(u - \ell_0)$, which is in turn determined by m, C_1, and C_2. Recall that if u is a convex solution of (1.1), vanishing in $\partial\Omega$, then u is strictly convex [C1]. The C^2 estimate in (i) and the $C^{2,\alpha}$ estimate in (ii) were proved in [W1] and [C2], respectively. See also section 6 of [TW]. Here we give a unified and shorter proof, using an idea from [W3], where a short and elementary proof of (1.8) for the Laplace and heat equations was given. Our argument was also inspired by the original idea of Caffarelli [C2].

The main estimate of the paper is the following log-Lipschitz continuity for the gradient Du.

THEOREM 2. *Let u be a strictly convex solution to the Monge–Ampère equation* (1.1)*, and suppose f is continuous and satisfies* (1.7)*. Then we have the estimate*

$$|Du(x) - Du(y)| \le Cd[1 + e^{-2\theta\psi(d)}] \tag{1.11}$$

for any $x, y \in B_{1/2}(0)$, where $d = |x - y|$, $C = C(n, m, C_1, C_2, \omega_f)$, θ is a positive constant, and

$$\psi(d) = -\int_d^1 \frac{\omega_{\log f}(r)}{r} dr. \tag{1.12}$$

In particular, we have $\theta < \frac{1}{2}$.

Remarks. (i) In Theorem 2 we relax the Dini continuity of f to continuity. If f is not continuous, we wouldn't obtain a log-Lipschitz continuous from (1.11) and (1.12). But we still have a related estimate; see (3.9) and Remark 3.2 below.

(ii) From (1.11) and (1.12), we see that if $\omega_{\log f}(r) \le \frac{c_0}{|\log r|}$ for some constant $c_0 \le 1/2\theta$, then Du is log-Lipschitz continuous. Our estimate $\theta < \frac{1}{2}$ implies that Du is log-Lipschitz continuous if, for small $r > 0$,

$$\omega_{\log f}(r) \le \frac{1}{|\log r|}. \tag{1.13}$$

(iii) Our estimate (1.11) should be optimal, that is, the log-Lipschitz continuity does not hold, if $\omega_{\log f}(r) \ge \frac{C}{|\log r|}$ for large C. See section 4 for discussion.

(iv) In the application to the semigeostrophic equation in section 4, it suffices to establish an estimate

$$|Du(x) - Du(y)| \le Cd[1 + \eta(d)] \tag{1.14}$$

for some positive function $\eta(t)$ satisfying

$$\int_0^{e^{-2}} \frac{1}{t\eta(t)} = \infty. \tag{1.15}$$

Examples satisfying (1.15) include the function $\eta(t) = |\log t| \log |\log t|$. However, from our argument one sees that there is not much room for inhomogeneous functions f such that Du satisfies (1.14) and (1.15) but is not log-Lipschitz continuous. This situation is similar to the Laplace equation (1.3).

Theorems 1 and 2 will be proved, respectively, in sections 2 and 3. We indicate an application of Theorem 2 on the local existence of continuous solutions to the semigeostrophic equation in section 4.

2. Proof of Theorem 1. First we collect some basic properties.

LEMMA 2.1. *Let Ω be a bounded convex domain in R^n. Then there is a unique minimum ellipsoid containing Ω, which attains the minimum volume among all ellipsoids containing Ω.*

We refer the reader to [G] for a proof. We say a convex set Ω is *normalized* if its minimum ellipsoid is a ball. When Ω is normalized, one has $B_{r/n} \subset \Omega \subset B_r$ for concentric balls $B_{r/n}$ and B_r [G].

Therefore for any bounded convex domain Ω, there is a unique unimodular linear transformation T (namely, $\det T = 1$) such that $T(\Omega)$ is normalized. Choose an appropriate coordinate system such that the minimum ellipsoid of Ω is given by $E =$

$\{\Sigma \frac{x_i^2}{a_i^2} < 1\}$, with $a_1 \geq \cdots \geq a_n$. Then T is determined by the matrix $\text{diag}(\lambda_1, \ldots, \lambda_n)$, with

$$\lambda_i = \frac{1}{a_i}(a_1 \ldots a_n)^{1/n}, 1, 2, \ldots, n.$$

Note that λ_1 and λ_n are the least and largest eigenvalues of T, respectively. For convenience we say in this paper that Ω has a *good shape* if

$$\lambda_n \leq c^* \lambda_1 \tag{2.1}$$

for some constant c^* under control. If Ω has a good shape, then there exist two concentric balls B_r and B_R, with $R \leq nc^* r$, such that $B_r \subset \Omega \subset B_R$.

Let u be a convex function defined in a bounded domain Ω. Following Caffarelli [C2], we denote, for any $y \in \Omega, h > 0$,

$$S^0_{h,u}(y) = \{x \in \Omega : \quad u(x) < \ell_y(x) + h\} \tag{2.2}$$

the level set of u and denote $S_{h,u}(y) = \partial S^0_{h,u}(y)$ its boundary, where ℓ_y is the tangent plane of u at y. When no confusion arises we will drop the subscript u, and when y is the minimum point of u, we will simply write the level set as S^0_h.

LEMMA 2.2. *Let u_i, $i = 1, 2$, be two convex solutions of* $\det D^2 u = 1$ *in $B_1(0)$. Suppose $\|u_i\|_{C^4} \leq C_0$. Then if $|u_1 - u_2| \leq \delta$ in $B_1(0)$ for some constant $\delta > 0$, we have, for $1 \leq k \leq 3$,*

$$|D^k(u_1 - u_2)| \leq C\delta \quad \text{in} \;\; B_{1/2}(0). \tag{2.3}$$

Proof. We have

$$\begin{aligned}\det D^2 u_2 - \det D^2 u_1 &= \int_0^1 \frac{d}{dt} \det[D^2 u_1 + t(D^2 u_2 - D^2 u_1)] dt \\ &= a_{ij}(x) \partial_i \partial_j (u_2 - u_1) = 0.\end{aligned}$$

Since u_i satisfies the equation det $D^2 u = 1$, it is uniformly convex by the assumption $\|u_i\|_{C^4} \leq C_0$. Hence the operator $L = a_{ij}(x)\partial_i\partial_j$ is linear and uniformly elliptic, with C^2 coefficients. By the Schauder estimates of linear elliptic equations, we obtain (2.3). □

We also need the following regularity for the Monge–Ampère equation [GT, P].

LEMMA 2.3. *Let Ω be a bounded convex domain in R^n. Let u be a convex solution of* $\det D^2 u = 1$ *in Ω, vanishing on $\partial\Omega$. If $B_r(0) \subset \Omega \subset B_R(0)$, then for any $\Omega' \subset\subset \Omega$, there is a constant $C > 0$, depending only on n, r, R, and dist$(\Omega', \partial\Omega)$, such that*

$$\|u\|_{C^4(\Omega')} \leq C. \tag{2.4}$$

From Lemma 2.3 we have the following.

LEMMA 2.4. *Let u be a convex solution of* $\det D^2 u = 1$ *in Ω which vanishes on $\partial\Omega$. Suppose u attains its minimum at the origin and $D^2 u(0)$ is the unit matrix (or uniformly bounded); then the domain Ω is of good shape.*

Indeed, if Ω does not have a good shape, one may make a unimodular linear transform $y = Tx$ to normalize Ω. Then the ratio of the largest and least eigenvalues of T will be large. By Lemma 2.3, $D^2_y u(0)$ is uniformly bounded. Hence $D^2_x u(0) = T' D^2_y(0) T$ cannot be the unit matrix (or uniformly bounded).

Proof of Theorem 1. By subtracting a linear function we suppose

$$u(0) = 0, \quad Du(0) = 0,$$

so that the origin is the minimum point of u. We consider the solution u in the level set S^0_h, where $h > 0$ is chosen small such that S^0_h is compactly supported in $B_1(0)$. By Lemma 2.1 there is a unique unimodular linear transform T_h such that $T_h(S^0_h)$ is normalized. Hence by making the change $x \to T_h x/\sqrt{h}$ and $u \to u/h$, we may suppose $h = 1$, S^0_1 is normalized, and

$$\int_0^1 \frac{\omega(r)}{r} \le \varepsilon, \tag{2.5}$$

where $\omega(r) = \omega_f(r)$, and ε can be as small as we want, provided h is sufficiently small. Note that the Monge–Ampère equation is invariant under the change $x \to Tx$ and $u \to (\det T)^{2/n} u$ for any nondegenerate linear transform T.

Let u_k, $k = 0, 1, \ldots$, be the solution of

$$\begin{aligned} \det D^2 u_k &= f(0) \quad \text{in } \ S^0_{4^{-k},u}, \\ u_k &= u \ (= 4^{-k}) \quad \text{on } \ \partial S^0_{4^{-k},u}. \end{aligned} \tag{2.6}$$

Denote

$$\begin{aligned} \nu(t) &= \sup_{z \in B_1} \{|f(x) - f(y)| : \quad x, y \in S^0_{t^2,u}(z)\}, \\ \nu_k &= \nu(2^{-k}), \end{aligned} \tag{2.7}$$

which is invariant under unimodular linear transformation of x. If $S^0_{t^2,u}$ has a good shape, then we have $\nu(t) \le \omega(Ct)$.

Since $S^0_{1,u}$ has a good shape, by Lemma 2.3, $\|u_0\|_{C^4(S^0_{3/4,u})} \le C$. Note that

$$\det D^2(1 - C\nu_0)u \le \det D^2 u_0 \le \det D^2(1 + C\nu_0)u \quad \text{in } \ S^0_{1,u}$$

for some constant C, and $u = u_0 = 1$ on the boundary. By the comparison principle, we have

$$(1 + C\nu_0)(u - 1) \le u_0 - 1 \le (1 - C\nu_0)(u - 1).$$

It follows that $|u - u_0| \le C\nu_0$. Similarly we have $|u - u_1| \le C\nu_1$. Hence we obtain $|u_1 - u_0| \le C\nu_0$. Since $S^0_{1,u}$ has a good shape, so does $S^0_{4^{-1},u}$. It follows that $\|u_1\|_{C^4(S^0_{3/16,u_1})} \le C$. Replacing the balls B_1 and $B_{1/2}$ in Lemma 2.2 by the level sets $S^0_{1/8,u_1}$ and $S^0_{1/16,u_1}$, we obtain

$$|D^k u_0(x) - D^k u_1(x)| \le C\nu_0 \tag{2.8}$$

for $x \in S^0_{4^{-2},u_1}$, where $1 \le k \le 3$. By Lemma 2.4, the estimate also implies that $S^0_{4^{-2},u_1}$ has a good shape.

By induction we assume that $S^0_{4^{-k-1},u}$ has a good shape, with the constant c^* (see (2.1)) independent of k. Hence $\nu_k \le \omega(C2^{-k})$ for some $C > 0$ independent of k (depending on c^*). Applying the same argument to $\hat{u}_0 := 4^k u_k(\frac{x}{2^k})$ and $\hat{u}_1 := 4^k u_{k+1}(\frac{x}{2^k})$, we obtain, for $x \in S^0_{4^{-k-2},u_{k+1}}$,

$$|Du_k(x) - Du_{k+1}(x)| \le C2^{-k}\nu_k, \tag{2.9a}$$

$$|D^2u_k(x) - D^2u_{k+1}(x)| \le C\nu_k, \tag{2.9b}$$

$$|D^3u_k(x) - D^3u_{k+1}(x)| \le C2^k\nu_k, \tag{2.9c}$$

where 2^k in (2.9) is the scaling constant. Hence

$$|D^2u_0(x) - D^2u_{k+1}(x)| \le C\sum_{i=0}^{k}\nu_i \le C\int_{2^{-k}}^{1}\frac{\omega(r)}{r}dr \tag{2.10}$$

for $x \in S^0_{4^{-k-2},u_{k+1}}$, where $C > 0$ is independent of k.

Estimate (2.10), together with (2.5) and Lemma 2.4, implies that $S^0_{4^{-k-2},u_{k+1}}$ has a good shape, with the constant c^* independent of k. Denote

$$\hat{u} = 4^{k+1}u\left(\frac{x}{2^{k+1}}\right), \quad \hat{u}_{k+1} = 4^{k+1}u_{k+1}\left(\frac{x}{2^{k+1}}\right). \tag{2.11}$$

Then $\hat{u}$ and $\hat{u}_{k+1}$ satisfy the equations $\det D^2\hat{u} = f(2^{-k-1}x)$ and $\det D^2\hat{u} = f(0)$, respectively. By the comparison principle,

$$|\hat{u} - \hat{u}_{k+1}| \le C\nu_{k+1}. \tag{2.12}$$

Hence $S^0_{4^{-1},\hat{u}}$ has a good shape, and so also $S^0_{4^{-k-2},u}$ has a good shape.

For any given point z near the origin,

$$\begin{aligned} |D^2u(z) - D^2u(0)| &\le I_1 + I_2 + I_3 =: \\ |D^2u_k(z) - D^2u_k(0)| &+ |D^2u_k(0) - D^2u(0)| + |D^2u(z) - D^2u_k(z)|. \end{aligned} \tag{2.13}$$

Let $k \ge 1$ such that $4^{-k-4} \le u(z) \le 4^{-k-3}$. Then by (2.9b) and recalling that $\nu(t) \le \omega(Ct)$,

$$I_2 \le C\sum_{j=k}^{\infty}\nu_j \le C\int_0^{|z|}\frac{\omega(r)}{r}. \tag{2.14}$$

We remark that in the second inequalities in (2.10) and (2.14), the integrand should be $\frac{1}{r}\omega(Cr)$ for some constant $C > 0$. But from the definition (1.5), we have $\omega(Cr) \le C\omega(r)$.

Next we estimate I_3. Let $u_{z,j}$ be the solution of

$$\begin{aligned} \det D^2u_{z,j} &= f(z) \quad \text{in } S^0_{4^{-j},u}(z), \\ u_{z,j} &= u \quad \text{on } \partial S^0_{4^{-j},u}(z). \end{aligned} \tag{2.15}$$

Let $j_k = \inf\{j : \ S^0_{4^{-j},u}(z) \subset S^0_{4^{-k},u}(0)\}$. Obviously $j_k \ge k$. We claim that $j_k \le k + l_0$ for some fixed l_0 independent of k. Indeed, by making the dilation $x \to 2^kx$ and $u \to 4^ku$, we may assume that $k = 0$ and $u(z) \le 4^{-3}$. From Caffarelli's strict convexity [C1], there exists a constant $l_0 > 0$ such that the tangent plane ℓ of u at z satisfies

$$\ell(x) \le u(x) - 4^{-l_0}$$

for any boundary point x of $S^0_{1,u}(0)$. In other words, we have $S^0_{4^{-l_0},u}(z) \subset S^0_{1,u}(0)$. Scaling back, we obtain $j_k \le k + l_0$. Note that $|u_k - u_{z,k+l_0}| \le C\nu_k$. Applying Lemma 2.2 to u_k and $u_{z,k+l_0}$ in $S^0_{4^{-k-l_0},u}(z)$, we have

$$|D^2u_k(z) - D^2u_{z,k+l_0}(z)| \le C\nu_k. \tag{2.16}$$

Similarly to (2.14) we have

$$|D^2u(z) - D^2u_{z,k+l_0}(z)| \le C \sum_{j=k+l_0}^{\infty} \nu_j \le C \int_0^{|z|} \frac{\omega(r)}{r}. \tag{2.17}$$

Combining (2.16) and (2.17), we obtain an estimate for I_3.

To estimate I_1, denote $h_j = u_j - u_{j-1}$. By (2.9c),

$$|D^2h_j(z) - D^2h_j(0)| \le C2^j\nu_j|z|. \tag{2.18}$$

Hence

$$\begin{aligned} I_1 &\le |D^2u_{k-1}(z) - D^2u_{k-1}(0)| + |D^2h_k(z) - D^2h_k(0)| \\ &\le |D^2u_0(z) - D^2u_0(0)| + \sum\nolimits_{j=1}^{k} |D^2h_j(z) - D^2h_j(0)| \\ &\le C|z|\left(1 + \sum\nolimits_{j=1}^{k} 2^j\nu_j\right) \\ &\le C|z|\left(1 + \int_{|z|}^1 \frac{\omega(r)}{r^2}\right). \end{aligned} \tag{2.19}$$

Hence we obtain (1.8). Note that (1.9) and (1.10) follow readily from (1.8). This completes the proof of Theorem 1. □

3. Proof of Theorem 2. The proof of Theorem 2 is divided into two parts. The first part is the proof of (1.11) with a large constant θ, and the second one is an estimate for θ.

The first part is a modification of the proof of Theorem 1. A difference is that in the proof of Theorem 1, due to (2.5) the level sets $S^0_{4^{-k},u}$ have a good shape for all $k > 0$, and we don't need to make linear transforms to normalize them. But in the proof of Theorem 2, due to the lack of (2.5), the level set will in general not have a good shape. Therefore we have to make linear transforms for every k to keep these level sets in good shape and estimate carefully the accumulation of all of the transforms. We also allow discontinuous f in our argument from (3.1) until (3.9).

3.1. By subtracting a linear function we suppose $u(0) = 0$, $Du(0) = 0$. By making a dilation of the axes, we may assume that $f(0) = 1$. Consider u in the level set S^0_h for some small $h > 0$ such that $S^0_h \subset\subset B_1(0)$. By making a linear transform as in section 2, we may assume that $h = 1$ and S^0_1 is normalized. Let $\nu_k = \nu(2^{-k})$ be as in (2.7). Let u_k, $k = 0, 1, \dots$, be the solution of

$$\begin{aligned} \det D^2u_k &= f_k \quad \text{in } S^0_{4^{-k},u}, \\ u_k &= u \quad \text{on } \partial S^0_{4^{-k},u}, \end{aligned} \tag{3.1}$$

where f_k is a constant:

$$f_k = \frac{1}{2}\left[\inf\{f(x) :\ x \in S^0_{4^{-k},u}\} + \sup\{f(x) :\ x \in S^0_{4^{-k},u}\}\right].$$

We choose such a special constant f_k to get a better (smaller) upper bound for the constant θ in (1.11). In this subsection we will assume $f_k = 1$ by making a dilation of the axes.

First we make a unimodular linear transform $x^{(0)} = T_0x$ such that $D^2_{x^{(0)}}u_0(0) = I$, where I denotes the unit matrix. Here and below we use D to denote derivatives in x and $D_{x^{(i)}}$ to denote derivatives in the new coordinates $x^{(i)}$. By Lemma 2.4, the level set $S^0_{\frac{1}{4},u}$ has a good shape.

We then make a unimodular linear transform $x^{(1)} = T_1 x^{(0)}$ such that $D^2_{x^{(1)}} u_1(0) = I$. Then the level set $S^0_{4^{-2},u}$ has a good shape, and estimates (2.9) (for $k = 1$) hold in the new coordinates $x^{(1)}$.

By induction we assume that $D^2_{x^{(k-1)}} u_{k-1}(0) = I$. Then by Lemma 2.4, the level set $S^0_{4^{-k+1},u_{k-1}}$ has a good shape. Hence, from the proof of Theorem 1 (see (2.11) and (2.12)), $S^0_{4^{-k+1},u}$ has a good shape. We then make a unimodular linear transform $x^{(k)} = T_k x^{(k-1)}$ such that $D^2_{x^{(k)}} u_k(0) = I$. More precisely, assume that, in the coordinates $x^{(k-1)}$, u_k has the expansion (after a rotation of axes such that $D^2_{x^{(k)}} u_k(0)$ is diagonal)

$$u_k(x) = u_k(0) + a_i x_i + \frac{1}{2} b_i x_i^2 + O(|x|^3).$$

Then the transform $x^{(k)} = T_k x^{(k-1)}$ is given by

$$T_k x = (b_1^{\frac{1}{2}} x_1, \ldots, b_n^{\frac{1}{2}} x_n),$$

so that, in $x^{(k)}$,

$$u_k(x) = u_k(0) + \frac{a_i}{b_i^{1/2}} x_i + \frac{1}{2} x_i^2 + O(|x|^3),$$

and the largest eigenvalue of T_k is

$$\lambda_{max}(T_k) = \max b_k^{1/2}. \tag{3.2}$$

Remark 3.1. Here we assume that the constant f_k in (3.1) is equal to 1, so that $\prod_{i=1}^n b_i = 1$. If $f_k \neq 1$, then T_k should be given by

$$T_k x = \frac{1}{f_k^{1/2n}} (b_1^{\frac{1}{2}} x_1, \ldots, b_n^{\frac{1}{2}} x_n),$$

so that T_k is unimodular, namely, $\det T_k = 1$.

After the transform T_k, the level set $S^0_{4^{-k},u_k}$ has a good shape, and estimates (2.9) hold in the new coordinates $x^{(k)}$. That is,

$$|D_{x^{(k)}} u_k(x) - D_{x^{(k)}} u_{k+1}(x)| \leq C 2^{-k} \nu_k, \tag{3.3a}$$

$$|D^2_{x^{(k)}} u_k(x) - D^2_{x^{(k)}} u_{k+1}(x)| \leq C \nu_k \tag{3.3b}$$

for $x \in S^0_{4^{-k-1},u_k}$.

For any given point z near the origin,

$$\begin{gathered} |Du(z) - Du(0)| \leq I_1 + I_2 + I_3 =: \\ |Du_k(z) - Du_k(0)| + |Du_k(0) - Du(0)| + |Du(z) - Du_k(z)|, \end{gathered}$$

where we choose $k = k_z \geq 1$ such that $4^{-k-4} \leq u(z) \leq 4^{-k-3}$. For the estimate of I_2, we have

$$I_2 = |Du_k(0) - Du(0)| \leq \sum_{i=k}^{\infty} |Du_i(0) - Du_{i+1}(0)|.$$

Denote $T^{(i)} = T_i \cdot T_{i-1} \ldots T_1 \cdot T_0$. Let λ_i be the largest eigenvalue of $T^{(i)}$. Then

$$|Du_i(0) - Du_{i+1}(0)| \leq \lambda_i |D_{x^{(i)}} u_i(0) - D_{x^{(i)}} u_{i+1}(0)|.$$

By (3.3a),

$$|D_{x^{(i)}}u_i(0) - D_{x^{(i)}}u_{i+1}(0)| \le C2^{-i}\nu_i.$$

It follows that

$$|Du_i(0) - Du_{i+1}(0)| \le C\lambda_i 2^{-i}\nu_i,$$

where C is independent of i. Hence we obtain

$$I_2 \le C\sum_{i=k}^{\infty} \lambda_i 2^{-i}\nu_i. \tag{3.4}$$

Similarly we have $I_3 \le C\sum_{i=k}^{\infty} \lambda_i 2^{-i}\nu_i$. To estimate I_1, by (3.3b) we have

$$|D^2_{x^{(i)}}u_i(x^{(i)}) - D^2_{x^{(i)}}u_{i+1}(x^{(i)})| \le C\nu_i$$

for any $i = 0, 1, \ldots, k$ and $x^{(i)} \in T^{(i)}(S^0_{4^{-i-2},u})$. Hence

$$|D^2u_i(x) - D^2u_{i+1}(x)| \le C\lambda_i^2\nu_i$$

for any $x \in S^0_{4^{-i-2},u}$.

Denote $h_i = u_i - u_{i-1}$. We have

$$|Dh_i(z) - Dh_i(0)| \le |D^2h_i|\,|z| \le C\lambda_i^2\nu_i|z|.$$

Hence

$$\begin{aligned} I_1 &\le |Du_{k-1}(z) - Du_{k-1}(0)| + |Dh_k(z) - Dh_k(0)| \\ &\le |Du_0(z) - Du_0(0)| + \sum\nolimits_{i=1}^{k}|Dh_i(z) - Dh_i(0)| \\ &\le |Du_0(z) - Du_0(0)| + C|z|\sum_{i=0}^{k}\lambda_i^2\nu_i \\ &\le C|z|\left[1 + \sum_{i=0}^{k}\lambda_i^2\nu_i\right]. \end{aligned} \tag{3.5}$$

We obtain

$$|Du(z) - Du(0)| \le C\sum_{i=k}^{\infty}\lambda_i 2^{-i}\nu_i + C|z|\left[1 + \sum_{i=0}^{k}\lambda_i^2\nu_i\right]. \tag{3.6}$$

Next we estimate λ_i. For a fixed i, denote

$$\hat{u} = 4^i u\left(\frac{x^{(i)}}{2^i}\right), \quad \hat{u}_i = 4^i u_i\left(\frac{x^{(i)}}{2^i}\right), \quad u^*_{i+1} = 4^i u_{i+1}\left(\frac{x^{(i)}}{2^i}\right). \tag{3.7}$$

Then $\hat{u}$, $\hat{u}_i$, and u^*_{i+1} satisfy, respectively, the equation $\det D^2\hat{u} = f(2^{-i}x^{(i)})$, f_i, and f_{i+1}. By definition, $\nu_i \ge \sup\{|f(2^{-i}x^{(i)}) - f(0)| : \ x^{(i)} \in S^0_{1,\hat{u}}\}$. Hence by the comparison principle,

$$|\hat{u} - \hat{u}_i| \le C\nu_i,$$
$$|\hat{u} - u^*_{i+1}| \le C\nu_{i+1}.$$

It follows that

$$|\hat{u}_i - u^*_{i+1}| \le C\nu_i.$$

Note that $D^2_{x^{(i)}} u_i(0) = I$. Hence by Lemma 2.2, $|D^2_{x^{(i)}} u_{i+1}(0) - I| \le C\nu_i$. We obtain

$$\lambda_{max}(T_i) \le 1 + \theta\nu_i \tag{3.8}$$

for some constant θ independent of i (but later we will give a more precise upper bound of θ for large i). Hence

$$\begin{aligned}\lambda_i &\le \prod\nolimits_{j=0}^{i} \lambda_{max}(T_j)\\ &\le \prod\nolimits_{j=0}^{i} (1 + \theta\nu_j)\\ &= e^{\sum_{j=0}^{i} \log(1+\theta\nu_j)}\\ &\le e^{\theta \sum_{j=0}^{i} \nu_j}.\end{aligned}$$

We have therefore established (recall that $Du(0) = 0$) that

$$\begin{aligned}|Du(z)| &\le C \sum_{i=k}^{\infty} 2^{-i}\nu_i e^{\theta \sum_{j=0}^{i} \nu_j} + C|z| \left[1 + \sum_{i=0}^{k} \nu_i e^{2\theta \sum_{j=0}^{i} \nu_j}\right]\\ &\le C \int_0^{2^{-k}} \nu(t) e^{\theta \int_t^1 \frac{\nu(s)}{s} ds} dt + C|z| \left[1 + \int_{2^{-k}}^1 \frac{\nu(t)}{t} e^{2\theta \int_t^1 \frac{\nu(s)}{s} ds} dt\right].\end{aligned} \tag{3.9}$$

Remark 3.2. From (3.9) we see that if $\lim_{t\to 0} \nu(t) < \varepsilon$ for some small $\varepsilon > 0$, then $u \in C^{1,\alpha}$ for some $\alpha \in (1 - 2\theta\varepsilon, 1)$, which also follows from Caffarelli's $W^{2,p}$ estimate. If, furthermore, $\int_0^1 \frac{\nu(t)}{t} < \infty$, the above estimate implies that Du is Lipschitz continuous; namely, D^2u is uniformly bounded. In the following we assume $\lim_{t\to 0} \nu(t) = 0$.

The right-hand side of (3.9) can be simplified as follows. Denote

$$\varphi(t) = -\int_t^1 \frac{\nu(s)}{s} ds.$$

Assume that $\nu(t) \to 0$ at $t \to 0$, so that $\varphi(t) = o(|\log t|)$ as $t \to 0$. The first integral on the right-hand side of (3.9) is equal to

$$\begin{aligned}\int_0^r \nu(t) e^{\theta \int_t^1 \frac{\nu(s)}{s} ds} dt &= \int_0^r t\varphi'(t) e^{-\theta\varphi(t)} dt\\ &= \frac{-r}{\theta} e^{-\theta\varphi(r)} + \frac{1}{\theta} \int_0^r e^{-\theta\varphi(t)},\end{aligned}$$

where $r = 2^{-k}$. The second integral on the right-hand side of (3.9) is equal to

$$\int_r^1 \varphi'(t) e^{-2\theta\varphi(t)} dt = \frac{1}{2\theta} [e^{-2\theta\varphi(r)} - e^{-2\theta\varphi(1)}].$$

We claim that if $\varphi(0) = -\infty$,

$$\int_0^r e^{-\theta\varphi(t)} = O(re^{-\theta\varphi(r)}) \ \text{ as } \ r \to 0.$$

Indeed, noting that $r\varphi'(r) = \nu(r) = o(1)$, we have

$$\begin{aligned}\lim_{r\to 0} \frac{\int_0^r e^{-\theta\varphi(t)}}{re^{-\theta\varphi(r)}} &= \lim_{r\to 0} \frac{(\int_0^r e^{-\theta\varphi(t)})'}{(re^{-\theta\varphi(r)})'}\\ &= \lim_{r\to 0} \frac{e^{-\theta\varphi(r)}}{e^{-\theta\varphi(r)}(1-\theta r\varphi'(r))}\\ &= 1.\end{aligned}$$

Therefore from (3.9),

$$|Du(z)| \le C2^{-k}[1 + e^{-\theta\varphi(2^{-k})}] + C|z|[1 + e^{-2\theta\varphi(2^{-k})}]. \tag{3.10}$$

Claim.

$$2^{-k}[1 + e^{-\theta\varphi(2^{-k})}] \le C|z|[1 + e^{-2\theta\varphi(2^{-k})}]. \tag{3.11}$$

Indeed, (3.11) is obvious if $2^{-k} \le |z|$. If $|z| \le 2^{-k}$, denote $h(t) = u(t\frac{z}{|z|})$, $\alpha = 2^{-k}/|z|$, $\beta = e^{-\theta\varphi(2^{-k})}$. Since $h(0) = h'(0) = 0$, by convexity, and since $u(z) \ge 4^{-k-4}$, we have

$$h'(t) \ge \frac{1}{t}[h(t) - h(0)] \ge \frac{4^{-k-4}}{|z|} = 2^{-k-8}\alpha \quad \text{at} \quad t = |z|.$$

From (3.10),

$$h'(t) \le C2^{-k}(1+\beta) + C|z|(1+\beta^2).$$

Combining the above two inequalities, we obtain

$$\alpha \le C(1+\beta) + \frac{C}{\alpha}(1+\beta^2).$$

Hence $\alpha \le C(1+\beta)$, and so

$$2^{-k}(1+\beta) = |z|\alpha(1+\beta) \le C|z|(1+\beta^2).$$

We also obtain (3.11).

We have therefore proved that

$$|Du(z) - Du(0)| \le C|z|[1 + e^{-2\theta\varphi(2^{-k})}]. \tag{3.12}$$

Note that estimate (3.12) still holds and Du is Lipschitiz continuous if $\varphi(0) > -\infty$. This is the case treated in section 2.

To obtain (1.11) from (3.12), note that, by Remark 3.2, $u \in C^{1,\alpha}$ for any α close to 1. Hence for any $\varepsilon > 0$, the level set $S^0_{t^2,u}(y)$ is contained in the ball $B_{t^{1-\varepsilon}}(y)$ provided $t > 0$ is sufficiently small. In particular, we have $2^{-k} \ge |z|^{1+\varepsilon}$ and $\nu(t) \le \omega(t^{1-\varepsilon})$. With $d = |z|$ we then obtain

$$\begin{aligned}|\varphi(2^{-k})| \le \int_{d^{1+\varepsilon}}^1 \frac{\nu(t)}{t}dt &\le \int_{d^{1+\varepsilon}}^1 \frac{\omega(t^{1-\varepsilon})}{t}dt\\ &= \frac{1}{1-\varepsilon}\int_{d^{1+\varepsilon}}^1 \frac{\omega(t^{1-\varepsilon})}{t^{1-\varepsilon}}dt^{1-\varepsilon}\\ &\le \frac{1}{1-\varepsilon}\int_{d^{1-\varepsilon^2}}^1 \frac{\omega(t)}{t}dt\\ &= \frac{1}{1-\varepsilon}|\hat\psi(d^{1-\varepsilon^2})| \le \frac{1}{1-\varepsilon}|\hat\psi(d)|,\end{aligned}$$

where

$$\hat{\psi}(d) = -\int_d^1 \frac{\omega_f(r)}{r} dr.$$

We obtain (1.11) at $y = 0$ with $\omega_{\log f}$ replaced by ω_f but under the assumption that $f(0) = 1$.

To see that estimate (1.11) is actually determined by $\omega_{\log f}$, consider $w = \frac{u}{f^{1/n}(0)}$ such that $\det D^2 w = \hat{f} =: \frac{f}{f(0)}$. We have

$$\begin{aligned} |\hat{f}(x) - \hat{f}(0)| &= \left| \frac{f(x)}{f(0)} - 1 \right| \\ &\le (1+\varepsilon) \log \frac{f(x)}{f(0)} \\ &= (1+\varepsilon)[\log f(x) - \log f(0)] \end{aligned}$$

for x near 0, namely, $\omega_{\hat{f}} \le (1+\varepsilon)\omega_{\log f}$ for some constant $\varepsilon > 0$. The constant ε can be as small as we want, provided x is sufficiently close to 0. Recall that we allow that the constant C in (1.11) depends on the upper and lower bounds of f and the modulus of continuity of f.

From (3.12) we see that if

$$\nu(t) \le \frac{1}{2\theta\,|\log t|} \tag{3.13}$$

for $t > 0$ small, then Du is log-Lipschitz continuous.

3.2. To finish the proof of Theorem 2, it remains to prove that $\theta < \frac{1}{2}$.

LEMMA 3.1. *Assume* $\lim_{x\to 0} f(x) = 1$. *For any* $\varepsilon > 0$, *there exists* $h_\varepsilon > 0$ *such that, when* $0 < h < h_\varepsilon$, $S_{h,u}(0)$ *is in the* $\varepsilon h^{\frac{1}{2}}$*-neighborhood of a sphere of radius* $(2h)^{1/2}$ *after normalization.*

This was proved in [C2, Lemma 7]. Moreover, the condition $\lim_{x\to 0} f(x) = 1$ can be relaxed to

$$\overline{\lim}_{x\to 0} |f(x) - 1| \le \delta$$

for some $\delta > 0$ depending on ε. We note that Lemma 3.1 can also be proved by a blowup argument, as a convex solution of $\det D^2 u = 1$ must be a quadratic function if its graph is complete.

LEMMA 3.2. *Let* $f(0) = 1$ *and* $\nu = osc_{S^0_{1,u}} f$. *Let* $\overline{u} \in C^4_{loc}(S^0_{1,u})$ *be the solution of*

$$\det D^2 v = \overline{f} \quad in \quad S^0_{1,u}, \quad v = 1 \quad on \quad \partial S^0_{1,u}, \tag{3.14}$$

where $\overline{f}$ *is a constant,* $\overline{f} = \frac{1}{2}(f_{min} + f_{max})$, $f_{min} = \inf\{f(x) : \ x \in S^0_{1,u}\}$, *and* $f_{max} = \sup\{f(x) : \ x \in S^0_{1,u}\}$. *Then*

$$|u - \overline{u}| \le \frac{3\nu}{8n} + C\nu^2 \quad on \ \partial S^0_{\frac{1}{4},u}. \tag{3.15}$$

Proof. Let u_{min} (u_{max}, respectively) be the solution of $\det D^2 v = f_{min}$ (f_{max}, respectively) in $S^0_{1,u}$ such that $v = 1$ on $\partial S^0_{1,u}$. Then

$$u_{max} - 1 = \left(\frac{f_{max}}{f_{min}}\right)^{\frac{1}{n}} (u_{min} - 1).$$

Observe that $\frac{f_{max}}{f_{min}} = 1 + \nu + O(\nu^2)$. We have $\left(\frac{f_{max}}{f_{min}}\right)^{\frac{1}{n}} = 1 + \frac{\nu}{n} + O(\nu^2)$. By the comparison principle,

$$u_{max} \le u \le u_{min}.$$

Hence on $S_{\frac{1}{4},u}$,

$$|u_{max}| \le |u_{min}| + \frac{3\nu}{4n} + C\nu^2. \tag{3.16}$$

By our choice, $\overline{f} = \frac{1}{2}(f_{max} + f_{min})$. Hence

$$\overline{u} = \frac{1}{2}\big(u_{min} + u_{max}\big) + O(\nu^2).$$

We obtain (3.15). Alternatively we have

$$\overline{u} - 1 = \left(\frac{\overline{f}}{f_{min}}\right)^{\frac{1}{n}} (u_{min} - 1),$$

and, similarly to (3.16), $|\overline{u}| \le |u_{min}| + \frac{3\nu}{8n} + C\nu^2$. We also obtain (3.15). ◻

Denote

$$\beta_n = \sup |D^2 u(0)|, \tag{3.17}$$

where $|D^2 u| = \max_{|\xi|=1} u_{\xi\xi}$, and the sup is taken among all harmonic functions in the unit ball $B_1(0) \subset \mathrm{R}^n$ satisfying $|u| \le 1$ on ∂B_1.

Let $\hat{u}$, $\hat{u}_i$, and u^*_{i+1} be the functions given in (3.7). Then $\hat{u}_i$ (u^*_{i+1}, respectively) satisfies

$$\det D^2 u = f_i \quad (f_{i+1}, \text{respectively}) \;\; \text{in } S^0_{1,\hat{u}} \;\; (\text{in } S^0_{\frac{1}{4},\hat{u}}, \text{respectively}),$$

where, by Lemma 3.1, the set $S^0_{\frac{r^2}{2},\hat{u}}$ is a small perturbation of a ball of radius r. As f_{i+1} may differ from f_i, we introduce a new function $\hat{v}_{i+1}$, which is the solution of

$$\det D^2 v = f_i \quad \text{in} \;\; S^0_{\frac{1}{4},\hat{u}}, \quad v = \hat{u} \quad \text{on} \;\; S_{\frac{1}{4},\hat{u}}.$$

Then

$$\hat{v}_{i+1} - \frac{1}{4} = \left(\frac{f_i}{f_{i+1}}\right)^{\frac{1}{n}} (u^*_{i+1} - \frac{1}{4}). \tag{3.18}$$

Let

$$v = \frac{\hat{u}_i - \hat{v}_{i+1}}{\|\hat{u}_i - \hat{v}_{i+1}\|_{L^\infty(S^0_{1/4,\hat{u}})}}.$$

Then v satisfies a linearized Monge–Ampère equation, that is,

$$\begin{aligned} \det D^2\hat{v}_{i+1} - \det D^2\hat{u}_i &= \int_0^1 \frac{d}{dt}\det[D^2\hat{u}_i + t(D^2\hat{v}_{i+1} - D^2\hat{u}_i)]dt \\ &= a_{ij}(x)\partial_i\partial_j(\hat{v}_{i+1} - \hat{u}_i) = 0. \end{aligned} \tag{3.19}$$

Notice that, by Lemma 3.1, both $\hat{u}_i$ and $\hat{v}_{i+1}$ converge to $\frac{1}{2}|x|^2$ as $i \to \infty$. By the regularity of the Monge–Ampère equation (Lemma 2.3), the matrix $\{a_{ij}\}$ converges to the unit matrix. Hence from (3.17),

$$|D^2 v(0)| \le 2(\beta_n + \varepsilon), \tag{3.20}$$

where $\varepsilon > 0$ can be as small as we want, provided i is large enough. The coefficient 2 is due to the fact that v_{i+1} is defined in $S^0_{\frac{1}{4},\hat{u}}$, which is a small perturbation of $B_{1/\sqrt{2}}$.

By the homogeneous equation (3.19), $\sup\{|\hat{u}_i - \hat{v}_{i+1}|(x) : \ x \in S^0_{\frac{1}{4},\hat{u}}\}$ is attained on the boundary $S_{\frac{1}{4},\hat{u}}$. Hence by Lemma 3.2,

$$\begin{aligned}\sup\{|\hat{u}_i - \hat{v}_{i+1}|(x) : \ x \in S^0_{\frac{1}{4},\hat{u}}\} &\le \sup\{|\hat{u}_i - \hat{u}|(x) : \ x \in S_{\frac{1}{4},\hat{u}}\} \\ &\le \frac{3\nu_i}{8n} + O(\nu_i^2).\end{aligned}$$

Recall that $D^2\hat{u}_i(0) = I$. Hence by (3.20),

$$|D^2\hat{v}_{i+1}(0) - I| \le (\beta_n + \varepsilon)\frac{3\nu_i}{4n} + O(\nu_i^2). \tag{3.21}$$

Note that, by (3.18),

$$D^2\hat{v}_{i+1} = \left(\frac{f_i}{f_{i+1}}\right)^{\frac{1}{n}} D^2 u^*_{i+1}.$$

Hence by a dilation $x \to \big(f_i/f_{i+1}\big)^{-1/2n}x$, we may cancel the coefficient $\big(\frac{f_i}{f_{i+1}}\big)^{\frac{1}{n}}$. It is obvious that the dilation does not affect the eigenvalues of the mapping T_i in (3.8) (because T_i is unimodular). Hence by (3.2) and (3.21),

$$\lambda_{max}(T_i) \le \left(1 + (\beta_n + \varepsilon)\frac{3\nu_i}{4n}\right)^{\frac{1}{2}}.$$

Therefore we obtain an upper bound for the constant θ in (3.8) (for large i)

$$\theta \le \frac{3\beta_n}{8n} + \varepsilon. \tag{3.22}$$

Next we give an upper bound for β_n.

Lemma 3.3. *Let β_n be given in* (3.17). *Then we have the estimate*

$$\beta_n = \frac{4(n+2)\omega_{n-1}}{\omega_n\sqrt{n}}\left(\frac{n-1}{n}\right)^{\frac{n-1}{2}}, \tag{3.23}$$

where ω_n is the area of the unit sphere $S^{n-1} \subset \mathrm{R}^n$.

Proof. For any small $\varepsilon > 0$, let u be a harmonic function satisfying

$$u_{nn}(0) \ge \sup|D^2 v(0)| - \varepsilon,$$

where the sup is taken among all harmonic functions v in the unit ball with $|v| \le 1$, and $|D^2 v|$ denotes the largest eigenvalue of the matrix $D^2 v$. By a rotation of axes we assume that $D^2u(0)$ is diagonal and $|D^2u(0)| = u_{nn}(0)$. By Green's representation,

$$u(x) = \frac{1 - |x|^2}{\omega_n}\int_{\partial B_1}\frac{g(y)}{|x - y|^n}dy,$$

where g is the boundary value of u on ∂B_1. Hence

$$u_{nn}(0)=\frac{n+2}{\omega_n}\left[\int_{\partial B_1} ny_n^2 g-\int_{\partial B_1} g\right]. \tag{3.24}$$

To compute the above integrals, we make a rearrangement of the function g, which keeps the integral $\int_{\partial B_1} g$ invariant, such that g is rotationally symmetric in $x'=(x_1,\dots,x_{n-1})$, even in x_n, and is monotone increasing in x_n for $x_n\in(0,1)$. It is easy to see that the rearrangement will increase the value $u_{nn}(0)$.

After the arrangement, g is a function of x_n. There exists a constant $t\in(0,1)$ such that $g>0$ when $x_n>t$ and $g\le 0$ when $x_n\le t$. If g is strictly positive or negative, we take $t=0$ or 1. Let h be a function on ∂B_1 which is rotationally symmetric in x', even in x_n, increasing in x_n for $x_n\in(0,1)$, and satisfies $\int_{\partial B_1} h=0$. Then $\int_{\partial B_1} y_n^2 h\ge 0$ and

$$\int_{\partial B_1} ny_n^2(g+h)-\int_{\partial B_1}(g+h)\ge\int_{\partial B_1} ny_n^2 g-\int_{\partial B_1} g.$$

Hence to compute $\sup|D^2v(0)|$, we may assume furthermore that

$$\begin{aligned} g&=1 \quad \text{when } x_n\in(t,1],\\ g&=-1 \quad \text{when } x_n\in[0,t) \end{aligned} \tag{3.25}$$

for a different $t\in(0,1)$. We have now the family of functions $\{g=g_t\}_{|t\in(0,1)}$. From the integrand in (3.24), one easily verifies that, among all of the functions $g=g_{t|t\in(0,1)}$, the sup is attained when $t=\frac{1}{\sqrt{n}}$. Therefore

$$u_{nn}(0)\le\frac{n+2}{\omega_n}\left\{\int_{S^{n-1}\cap\{|x_n|>\frac{1}{\sqrt{n}}\}}(ny_n^2-1)-\int_{S^{n-1}\cap\{|x_n|<\frac{1}{\sqrt{n}}\}}(ny_n^2-1)\right\}. \tag{3.26}$$

Notice that $u_{nn}(0)$ is invariant if we add a constant to g. Hence

$$\int_{S^{n-1}}(ny_n^2-1)=0.$$

We obtain

$$u_{nn}(0)\le\frac{2(n+2)}{\omega_n}\int_{S^{n-1}\cap\{|x_n|>\frac{1}{\sqrt{n}}\}}(ny_n^2-1).$$

Denote $r=|x'|$, $a=\sqrt{1-\frac{1}{n}}$. Then $y_n^2=1-r^2$ and

$$\int_{S^{n-1}\cap\{|x_n|>\frac{1}{\sqrt{n}}\}}(ny_n^2-1)=2\omega_{n-1}\int_0^a\frac{n(1-r^2)-1}{\sqrt{1-r^2}}r^{n-2}dr.$$

We have

$$\int_0^a\frac{r^n}{\sqrt{1-r^2}}=\frac{n-1}{n}\int_0^a\frac{r^{n-2}}{\sqrt{1-r^2}}-\frac{1}{n}a^{n-1}\sqrt{1-a^2}.$$

Hence

$$\int_0^a\frac{n(1-r^2)-1}{\sqrt{1-r^2}}r^{n-2}dr=a^{n-1}\sqrt{1-a^2}.$$

We obtain

$$\begin{aligned} u_{nn}(0) &\leq \frac{4(n+2)\omega_{n-1}}{\omega_n} a^{n-1}\sqrt{1-a^2} \\ &= \frac{4(n+2)\omega_{n-1}}{\omega_n\sqrt{n}}\left(\frac{n-1}{n}\right)^{\frac{n-1}{2}}. \end{aligned}$$

This completes the proof. ∎

The upper bound in (3.23) can be simplified.

LEMMA 3.4. *Let* β_n *be given in* (3.17). *Then* $\beta_2 = \frac{4}{\pi}$, $\beta_3 = \frac{20\sqrt{3}}{9}$, $\beta_4 = \frac{9\sqrt{3}}{\pi}$, $\beta_5 = \frac{168}{25}\sqrt{\frac{4}{5}}$, *and, for* $n \geq 6$,

$$\beta_n < n+2. \tag{3.27}$$

Proof. Since $\omega_2 = 2\pi$, $\omega_3 = 4\pi$, we have obviously $\beta_2 = \frac{4}{\pi}$, $\beta_3 = \frac{20\sqrt{3}}{9}$.

For $n \geq 4$, denote

$$\beta_n^* := \frac{\beta_n}{n+2} = \frac{4}{\sqrt{n-1}}\frac{\omega_{n-1}}{\omega_n}\left(\frac{n-1}{n}\right)^{\frac{n}{2}}. \tag{3.28}$$

We have

$$\begin{aligned} \omega_n &= 2\int_{\{|x'|<1\}} \frac{dx'}{\sqrt{1-|x'|^2}} \\ &= 2\omega_{n-1}\int_0^1 \frac{r^{n-2}}{\sqrt{1-r^2}}dr, \end{aligned}$$

where $x' = (x_1, \ldots, x_{n-1})$. Integration by parts gives

$$\int_0^1 \frac{r^{n-2}}{\sqrt{1-r^2}}dr = \frac{n-3}{n-2}\int_0^1 \frac{r^{n-4}}{\sqrt{1-r^2}}dr.$$

Hence we have

$$\int_0^1 \frac{r^{n-2}}{\sqrt{1-r^2}}dr = \frac{n-3}{n-2}\cdots\frac{2}{3} \tag{3.29}$$

if n is odd, and

$$\int_0^1 \frac{r^{n-2}}{\sqrt{1-r^2}}dr = \frac{n-3}{n-2}\cdots\frac{1}{2}\frac{\pi}{2}$$

if n is even. By direct computation, we obtain $\beta_4 = \frac{9\sqrt{3}}{\pi}$, $\beta_5 = \frac{168}{25}\sqrt{\frac{4}{5}}$, and (3.27) for $n \leq 10$. When $n > 10$, by (3.28) we have

$$\beta_n^* < \frac{4}{\sqrt{e}\sqrt{n-1}}\frac{\omega_{n-1}}{\omega_n}.$$

If $n = 2k > 10$ is even, then

$$\frac{\omega_{n-1}}{\omega_n} = \frac{1}{\pi}\frac{(n-2)(n-4)\ldots 2}{(n-3)(n-5)\ldots 1} := \frac{1}{\pi}I_n.$$

Since $\frac{k+1}{k} \le \frac{k}{k-1}$ for all k,

$$I_n < \left(\frac{n-2}{n-3}\frac{n-3}{n-4}\cdots\frac{10}{9}\frac{9}{8}\right)^{1/2}\frac{8}{7}\frac{6}{5}\frac{4}{3}\frac{2}{1} = \frac{\sqrt{n-2}}{\sqrt{8}}\frac{8}{7}\frac{6}{5}\frac{4}{3}\frac{2}{1}.$$

We obtain

$$\beta_n^* \le \sqrt{\frac{2}{e}}\frac{8}{7}\frac{6}{5}\frac{4}{3}\frac{2}{1}\frac{1}{\pi} < 1.$$

If $n = 2k+1$ is odd, then similarly

$$\beta_n^* \le \frac{2}{3\sqrt{e}}\frac{9}{8}\frac{7}{6}\frac{5}{4}\frac{3}{2} < 1.$$

This proves (3.27). $\square$

Let $\varepsilon > 0$ be sufficiently small; from (3.22) and Lemma 3.4 we have

$$\theta < \frac{1}{2} \tag{3.30}$$

in all dimensions. From (3.30) and (3.13), we see that Du is log-Lipschitz continuous when $\nu(t) \le \frac{1}{|\log t|}$. This completes the proof of Theorem 2. $\square$

4. Remarks. The log-Lipschitz continuity in Theorem 2 can be used to prove the local existence of continuous solutions to the Cauchy problem of the semigeostrophic equation, which is the transport equation

$$\partial_t \rho + \nabla(v\rho) = 0, \quad v = \nabla^{\perp} u, \tag{4.1}$$

coupled with the Monge–Ampère equation

$$\det(D^2 u + I) = \rho, \tag{4.2}$$

where $\nabla^{\perp} u = (-u_{x_2}, u_{x_1})$ in R^2 or $\nabla^{\perp} u = (-u_{x_2}, u_{x_1}, 0)$ in R^3 [BB, C5, CRD, CuF, E, L]. We assume the initial condition $\rho(\cdot, 0) = \rho_0 \in C^0(R^n)$ and the boundary condition $T_u(\mathrm{R}^n) = \Omega^*$, where $T_u(x) = Du(x) + x$ and Ω^* is a given bounded, convex domain. As in [L] we consider the periodic case. That is, for any $\vec{p} \in Z^n$, $\rho_0(x + \vec{p}) = \rho_0(x)$.

As the quasi-geostrophic equation, which is the transport equation (4.1) coupled with the equation $u = (-\Delta)^{-\frac{1}{2}}\rho$ [CF], the semigeostrophic equation has also been used as an approximation to the Euler equation [BB, C5, CRD, CuF, E, L]. Recall that for the 2-dimensional Euler equation, which can be written as a system of the transport equation (4.1) and the Poisson equation $u = (-\Delta)^{-1}\rho$, the global existence and uniqueness of smooth solutions were first derived by using the log-Lipschitz continuity for the Poisson equation [Y]. For the semigeostrophic equation, since the log-Lipschitz estimate (1.11) depends on the modulus of continuity of $\log f$, our Theorem 2 implies the local existence of continuous solutions when the initial data ρ_0 satisfies

$$\omega_{\log \rho_0}(r) < \frac{1}{|\log r|} \tag{4.3}$$

for small r. The proof is similar to that in [L], where it is assumed that f is Dini continuous. The local existence of continuous solutions was first obtained in [L] when ρ_0 is Dini continuous. We also refer the reader to [Br, C4] for the existence and regularity of solutions to the Monge–Ampère equation (4.2) subject to the boundary condition $T_u(\Omega) = \Omega^*$, where Ω is a convex domain in R^n.

Next we remark that the L^∞-oscillation $\omega_{\log f}$ in the estimate (3.9) can be replaced by the L^1-oscillation, using the following L^∞-estimate for the Monge–Ampère equation.

LEMMA 4.1. *Let u (respectively, $\overline{u}$) be a solution to the Monge–Ampère equation $\det D^2 u = f \geq 0$ (respectively, $\overline{f}$) in Ω. Suppose $u = \overline{u}$ on $\partial\Omega$. Then we have the estimate*

$$\sup_\Omega |u - \overline{u}| \leq C|f - \overline{f}|^{1/n}_{L^1(\Omega)}, \tag{4.4}$$

where C depends only on n and $\mathrm{diam}(\Omega)$.

Proof. Let φ be the convex solution to

$$\begin{aligned} \det D^2\varphi &= |f^{1/n} - \overline{f}^{1/n}|^n \quad \text{in} \quad \Omega, \\ \varphi &= 0 \quad \text{on} \quad \partial\Omega. \end{aligned}$$

Then we have the estimate

$$|\inf \varphi|^n \leq C \int_\Omega |f - \overline{f}|. \tag{4.5}$$

Estimate (4.5) can be established easily as follows. Suppose $\inf\varphi$ is attained at $x_0 \in \Omega$. Consider a convex function ψ whose graph is the convex cone with a vertex at $(x_0, \varphi(x_0))$ and a base $\partial\Omega \times \{x_{n+1} = 0\}$. Then $\psi = \varphi$ on $\partial\Omega$ and $\psi \geq \varphi$ in Ω. Denote by N_φ the normal mapping of φ [P]. Then $N_\psi(\Omega) \subset N_\varphi(\Omega)$. It follows that

$$|\inf \varphi|^n \leq C|N_\psi(\Omega)| \leq C|N_\varphi(\Omega)| = C \int_\Omega |f^{1/n} - \overline{f}^{1/n}|^n.$$

Noting that $|f^{1/n} - \overline{f}^{1/n}|^n \leq |f - \overline{f}|$, we obtain

$$|\inf \varphi|^n \leq C \int_\Omega |f - \overline{f}|,$$

where C depends only on n and $\mathrm{diam}(\Omega)$.

Next, observing that the Monge–Ampère operator $M(u) = \det^{1/n} D^2 u$ is concave, we have $M(\frac{u+v}{2}) \geq \frac{1}{2}(M(u) + M(v))$ for any convex function u, v. Hence

$$\det{}^{1/n} D^2(\overline{u} + \varphi) \geq \det{}^{1/n} D^2\overline{u} + \det{}^{1/n} D^2\varphi \geq f^{1/n}.$$

Similarly, $\det^{1/n} D^2(u + \varphi) \geq \overline{f}^{1/n}$. By the comparison principle, $|u - \overline{u}| \leq |\varphi|$. We obtain the estimate. □

Finally we would like to point out that estimate (1.11) should be optimal, in the sense that the gradient Du may not be log-Lipschitz continuous if $\omega_{\log f} \geq \frac{C}{|\log r|}$ for some large C. Indeed, consider the case when u and f are even functions in dimension two such that u and u_i (solutions of (3.1)) attain their minimum at 0. Then for an appropriate f, there is a positive constant $c_0 > 0$ independent of i such that $|\hat{u}_i - u^*_{i+1}| \geq c_0\nu_i$ and $\lambda_{max}(T_i) \geq 1 + c_0\nu_i$ (see (3.8) and the formula before it).

REFERENCES

[BB] J.D. Benamou and Y. Brenier, *Weak existence for the semigeostrophic equations formulated as a coupled Monge-Ampère/transport problem*, SIAM J. Appl. Math., 58 (1998), pp. 1450–1461.

[Br] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.

[C1] L.A. Caffarelli, *A localization property of viscosity solutions to the Monge-Ampère equation and their strict convexity*, Ann. of Math., 131 (1990), pp. 129–134.

[C2] L.A. Caffarelli, *Interior $W^{2,p}$ estimates for solutions of Monge-Ampère equations*, Ann. of Math., 131 (1990), pp. 135–150.

[C3] L.A. Caffarelli, *Some regularity properties of solutions of Monge-Ampère equation*, Comm. Pure Appl. Math., 44 (1991), pp. 965–969.

[C4] L.A. Caffarelli, *The regularity of mapping with a convex potential*, J. Amer. Math. Soc., 5 (1992), pp. 99–104.

[C5] L.A. Caffarelli, *Nonlinear elliptic theory and the Monge-Ampère equation*, in Proceedings of the International Congress of Mathematicians, 2002, Beijing Higher Education Press, Beijing, 2002, Vol. I, pp. 179–188.

[CF] D. Cordoba and C. Fefferman, *Growth of solutions for quasi-geostrophic and 2D Euler equations*, J. Amer. Math. Soc., 15 (2002), pp. 665–670.

[CuF] M. Cullen and M. Feldman, *Lagrangian solutions of semigeostrophic equations in physical space*, SIAM J. Math. Anal., 37 (2006), pp. 1371–1395.

[CRD] M. Cullen, L. Road and R. J. Douglas, *Applications of the Monge-Ampère equation and Monge transport problem to meteorology and oceanography*, Contemp. Math., 226 (1999), pp. 33–53.

[E] L.C. Evans, *Partial differential equations and Monge-Kantorovich mass transfer*, in Current Development in Mathematics, International Press, Boston, 1999, pp. 65–126.

[GJS] B. Guan, H.-Y. Jian, and R. Schoen, *Entire spacelike hypersurfaces of prescribed Gauss curvature in Minkowski space*, J. Reine Angew. Math., 595 (2006), pp. 167–188.

[GT] D. Gilbarg and N.S. Trudinger, *Elliptic partial differential equations of second order*, Springer, New York, 1983.

[G] M. de Guzman, *Differentiation of Integrals in* R^n, Lecture Notes in Math. 481, Springer, New York, 1975.

[L] G. Loeper, *A fully nonlinear version of the Euler incompressible equations: The semi-geostrophic system*, SIAM J. Math. Anal., 38 (2006), pp. 795–823

[P] A.V. Pogorelov, *The Multidimensional Minkowski Problems*, Wiley, New York, 1978.

[TW] N.S. Trudinger and X.-J. Wang, *Boundary regularity for the Monge-Ampère and affine maximal surface equations*, Ann. of Math., to appear.

[V] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, Providence, RI, 2003.

[W1] X.-J. Wang, *Remarks on the Regularity of Monge-Ampère equations*, in Proceedings of the International Conference on Nonlinear PDE, G.C. Dong and F.H. Lin, eds., Scientific Press, Beijing, 1992, pp. 257–263.

[W2] X.-J. Wang, *Some counterexamples to the regularity of Monge-Ampère equations*, Proc. Amer. Math. Soc., 123 (1995), pp. 841–845.

[W3] X.-J. Wang, *Schauder estimate for elliptic and parabolic equations*, Chinese Ann. Math. Ser. B, 27 (2006), pp. 637–642.

[Y] V. Yudovich, *Non-stationary flows of an ideal incompressible fluid*, Zh. Vychisl. Mat. Mat. Fiz., 3 (1963), pp. 1032–1066 (in Russian).

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 627–641

REMARKS ON THE MASS CONSTRAINT FOR KP-TYPE EQUATIONS*

L. MOLINET†, J. C. SAUT‡, AND N. TZVETKOV§

Abstract. For a rather general class of equations of Kadomtsev–Petviashvili type, we prove that the zero-mass (in x) constraint is satisfied at any nonzero time even if it is not satisfied at initial time zero. Our results are based on a precise analysis of the fundamental solution of the linear part and its anti-x-derivative.

Key words. Kadomtsev–Petviashvili equation, zero-mass (in x) solution, conservation laws

AMS subject classifications. 35B05, 35Q53

DOI. 10.1137/060654256

1. Introduction.

Kadomtsev–Petviashvili (KP) equations are universal for the modeling of the propagation of long weakly dispersive waves which propagate essentially in one direction with weak transverse effects. As explained in the pioneering paper of Kadomtsev and Petviashvili [7], they are (formally) obtained in the following way. We start from a one-dimensional long-wave dispersive equation which is of Korteweg–de Vries (KdV) type, i.e.,

$$u_t + u_x + uu_x - Lu_x = 0, \qquad u = u(t,x),\ x \in \mathbb{R},\ t \geq 0\,. \tag{1.1}$$

In (1.1) L is a (possibly nonlocal) operator, defined in a Fourier variable by

$$\widehat{Lf}(\xi) = c(\xi)\widehat{f}(\xi), \tag{1.2}$$

where c is a real function which is linked to the phase velocity. For instance, the case $c(\xi) = \pm\xi^2$ ($L = \mp\partial_x^2$) corresponds to the classical KdV equation. In the context of water waves, the sign of $c(\xi)$ depends on the surface tension parameter. The case $c(\xi) = |\xi|$ ($L = H\partial_x$) corresponds to the Benjamin–Ono equation, etc. We will consider here only the case of homogeneous symbols.

As observed in [7] the correction to (1.1) due to weak transverse effects is independent of the dispersion in x and is related only to the finite propagation speed properties of the transport operator $M = \partial_t + \partial_x$. Recall that M gives rise to one-directional waves moving to the right with speed one; i.e., a profile $\varphi(x)$ evolves under the flow of M as $\varphi(x-t)$. A weak transverse perturbation of $\varphi(x)$ is a two-dimensional function $\psi(x,y)$ close to $\varphi(x)$, localized in the frequency region $\left|\frac{\eta}{\xi}\right| \ll 1$, where ξ and η are the Fourier modes corresponding to x and y, respectively. We look for a two-dimensional perturbation $\widetilde{M} = \partial_t + \partial_x + \omega(D_x, D_y)$ of M such that, similarly to above, the profile of $\psi(x,y)$ does not change much when evolving under the flow of $\widetilde{M}$. Here $\omega(D_x, D_y)$

*Received by the editors March 14, 2006; accepted for publication (in revised form) February 20, 2007; published electronically August 10, 2007.
http://www.siam.org/journals/sima/39-2/65425.html

†L.A.G.A., Institut Galilée, Université Paris 13, 93430 Villetaneuse, France (molinet@math.univ-paris13.fr).
‡Université de Paris-Sud, UMR de Mathématiques, Bât. 425, 91405 Orsay Cedex, France (jean-claude.saut@math.u-psud.fr).
§Département de Mathématiques, Université Lille I, 59 655 Villeneuve d'Ascq Cedex, France (nikolay.tzvetkov@math.univ-lille1.fr).

denotes the Fourier multiplier with symbol the real function $\omega(\xi,\eta)$. Natural generalizations of the flow of M in two dimensions are the flows of the wave operators $\partial_t \pm \sqrt{-\Delta}$ which enjoy the finite propagation speed property. Since

$$\xi + \frac{1}{2}\xi^{-1}\eta^2 \sim \pm\sqrt{\xi^2+\eta^2}, \quad \text{when} \quad \left|\frac{\eta}{\xi}\right| \ll 1,$$

we deduce that

$$\partial_t + \partial_x + \frac{1}{2}\partial_x^{-1}\partial_y^2 \sim \partial_t \pm \sqrt{-\Delta},$$

which leads to the correction $\omega(D_x, D_y) = \frac{1}{2}\partial_x^{-1}\partial_y^2$ in (1.1).

Of course when the transverse effects are two-dimensional, the correction is $\frac{1}{2}\partial_x^{-1}\Delta_\perp$, where $\Delta_\perp = \partial_y^2 + \partial_z^2$.

We are thus led to the following model in two dimensions:

$$u_t + u_x + uu_x - Lu_x + \partial_x^{-1}\partial_y^2 u = 0. \tag{1.3}$$

In (1.3), it is implicitly assumed that the operator $\partial_x^{-1}\partial_y^2$ is well defined, which a priori imposes a constraint on the solution u, which, in some sense, has to be an x-derivative. This is achieved, for instance, if $u \in \mathcal{S}'(\mathbb{R}^2)$ is such that

$$\xi_1^{-1}\,\xi_2^2\,\widehat{u}(t,\xi_1,\xi_2) \in \mathcal{S}'(\mathbb{R}^2)\,, \tag{1.4}$$

thus in particular if $\xi_1^{-1}\,\widehat{u}(t,\xi_1,\xi_2) \in \mathcal{S}'(\mathbb{R}^2)$. Another possibility to fulfill the constraint is to write u as

$$u(t,x,y) = \frac{\partial}{\partial x}v(t,x,y), \tag{1.5}$$

where v is a continuous function having a classical derivative with respect to x, which, for any fixed y and $t \neq 0$, vanishes when $x \to \pm\infty$. Thus one has

$$\int_{-\infty}^{\infty} u(t,x,y)dx = 0, \qquad y \in \mathbb{R}, \ \ t \neq 0, \tag{1.6}$$

in the sense of generalized Riemann integrals. Of course the differentiated version of (1.3), namely

$$(u_t + u_x + uu_x - Lu_x)_x + \partial_y^2 u = 0, \tag{1.7}$$

can make sense without any constraint of type (1.4) or (1.6) on u, and so does the Duhamel integral representation of (1.3),

$$u(t) = S(t)u_0 - \int_0^t S(t-s)(u(s)u_x(s))ds, \tag{1.8}$$

where $S(t)$ denotes the (unitary in all Sobolev spaces $H^s(\mathbb{R}^2)$) group associated with (1.3),

$$S(t) = e^{-t(\partial_x - L\partial_x + \partial_x^{-1}\partial_y^2)}\,. \tag{1.9}$$

Let us notice, at this point, that models alternative to KdV-type equations (1.1) are the equations of Benjamin–Bona–Mahony (BBM) type

$$u_t + u_x + uu_x + Lu_t = 0 \tag{1.10}$$

with corresponding two-dimensional "KP–BBM-type models" (in the case $c(\xi) \geq 0$)

$$u_t + u_x + uu_x + Lu_t + \partial_x^{-1}\partial_y^2 u = 0 \tag{1.11}$$

or

$$(u_t + u_x + uu_x + Lu_t)_x + \partial_y^2 u = 0 \tag{1.12}$$

and free group

$$S(t) = e^{-t(I+L)^{-1}[\partial_x + \partial_x^{-1}\partial_y^2]} .$$

In view of the above discussion, all the results established for the Duhamel form of KP-type equations (e.g., those of Bourgain [5] and Saut and Tzvetkov [14]) do not need any constraint on the initial data u_0. It is then possible (see, for instance, [10]) to check that the solution u will satisfy (1.7) or (1.12) in the distributional sense but not a priori the integrated forms (1.3) or (1.11).

On the other hand, a constraint has to be imposed when using the Hamiltonian formulation of the equation. In fact, the Hamiltonian for (1.7) is

$$\frac{1}{2}\int \left[-u\,Lu + (\partial_x^{-1}u_y)^2 + u^2 + \frac{u^3}{3}\right] \tag{1.13}$$

and the Hamiltonian associated with (1.12) is

$$\frac{1}{2}\int \left[(\partial_x^{-1}u_y)^2 + u^2 + \frac{u^3}{3}\right] . \tag{1.14}$$

Therefore, the global well-posedness results for KP-I obtained in [11, 8] do need that the initial data satisfy (in particular) the constraint $\partial_x^{-1}\partial_y u_0 \in L^2(\mathbb{R}^2)$, and this constraint is preserved by the flow. Actually, the global results of [11, 8] make use of the next conservation law of the KP-I equation whose quadratic part contains the L^2-norms of u_{xx}, u_y, and $\partial_x^{-2}u_{yy}$. The constraint $\partial_x^{-2}\partial_{yy}u_0 \in L^2(\mathbb{R}^2)$ is thus also clearly needed, and one can prove that it is preserved by the flow.

On the other hand, as noticed in [11] there is a serious drawback with the higher conservation laws of both KP-I and KP-II equations, starting with that involving the L^2-norm of $\partial_x^3 u$ in its quadratic part. The problem is that this conservation law contains the L^2-norm of $(\partial_x^{-1}\partial_y)u^2$, whereas this expression is meaningless for $u \in H^3(\mathbb{R}^2)$. Indeed, for $u \in H^3(\mathbb{R}^2)$, using the Lebesgue theorem it is not to hard to see that if $(\partial_x^{-1}\partial_y)u^2 \in L^2(\mathbb{R}^2)$, then the integral $\int_{\mathbb{R}} \partial_y(u^2)(x,y)\,dx = \partial_y \int_{\mathbb{R}} u^2(x,y)\,dx$ must vanish for every $y \in \mathbb{R}$. Since $u \in L^2(\mathbb{R}^2)$ this clearly forces u to be identically zero.

The goal of this paper is to investigate the behavior of a solution to the general KP-type equations (1.7), (1.12) which initially does not satisfy the zero-mass constraint. We will show that, in fact, the zero-mass constraint is satisfied at any nonzero time t.

At this point we should mention the papers [1, 3, 4] and especially [6], where the inverse scattering transform machinery is used to solve the Cauchy problem for KP-I and KP-II without the constraint. The more complete and rigorous results are obtained in [6] (see also [17]). In the present work we consider a rather general class of KP or KP–BBM equations and emphasize the key point which concerns only the linear part: the fundamental solution of KP-type equations is a x derivative of a C^1

with respect to an x-continuous function which, for fixed $t \neq 0$ and y, tends to zero as $x \to \pm\infty$. Thus its generalized Riemann integral in x vanishes for all values of the transverse variable y and of $t \neq 0$. A similar property can be established for the solution of the nonlinear problem.

The paper is organized as follows. Section 2 deals with the KP-type equations, while section 3 focuses on KP–BBM-type equations. Section 4 reviews briefly some extensions: the three-dimensional case and nonhomogeneous dispersion relations.

In what follows, different harmless numerical constants will be denoted by c.

2. KP-type equations.

2.1. The linear case. We consider two-dimensional linear KP-type equations

$$(u_t - Lu_x)_x + u_{yy} = 0, \quad u(0,x,y) = \varphi(x,y)\,, \tag{2.1}$$

where

$$\widehat{Lf}(\xi) = \varepsilon|\xi|^\alpha \widehat{f}(\xi), \quad \xi \in \mathbb{R}, \tag{2.2}$$

where $\varepsilon = 1$ (KP-II-type equations) or $\varepsilon = -1$ (KP-I-type equations). We denote by G the fundamental solution

$$G(t,x,y) = \mathcal{F}^{-1}_{(\xi,\eta)\to(x,y)}\big[e^{it(\varepsilon\xi|\xi|^\alpha - \eta^2/\xi)}\big].$$

A priori, we have only that $G(t,\cdot,\cdot) \in \mathcal{S}'(\mathbb{R}^2)$. Actually, for $t \neq 0$, $G(t,\cdot,\cdot)$ has a very particular form which is the main result of this section.

THEOREM 2.1. *Suppose that $\alpha > 1/2$ in (2.2). Then for $t \neq 0$,*

$$G(t,\cdot,\cdot) \in C(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2).$$

Moreover, for $t \neq 0$, there exists

$$A(t,\cdot,\cdot) \in C(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2) \cap C^1_x(\mathbb{R}^2)$$

($C^1_x(\mathbb{R}^2)$ denotes the space of continuous functions on $\mathbb{R}^2$ which have a continuous derivative with respect to the first variable) such that

$$G(t,x,y) = \frac{\partial A}{\partial x}(t,x,y).$$

In addition, for $t \neq 0$, $y \in \mathbb{R}$, $\varphi \in L^1(\mathbb{R}^2)$,

$$\lim_{|x|\to\infty} (A \star \varphi)(t,x,y) = 0.$$

As a consequence, the solution of (2.1) with data $\varphi \in L^1(\mathbb{R}^2)$ is given by

$$u(t,\cdot,\cdot) \equiv S(t)\varphi = G \star \varphi$$

and

$$u(t,\cdot,\cdot) = \frac{\partial}{\partial x}\big(A \star \varphi\big).$$

One therefore has

$$\int_{-\infty}^{\infty} u(t,x,y)\,dx = 0 \quad \forall y \in \mathbb{R}, \;\; \forall t \neq 0$$

in the sense of generalized Riemann integrals.

REMARK 2.1. *It is worth noticing that the result of Theorem* 2.1 *is related to the infinite speed of propagation of the KP free evolutions. Let us also notice that the assumption* $\alpha > 1/2$ *can be relaxed if we assume that a sufficient number of derivatives of* φ *belong to* L^1. *Such an assumption is, however, not natural in the context of the KP equations.*

REMARK 2.2. *In the case of the classical KP-II equation* ($\alpha = 2$, $\varepsilon = +1$), *Theorem* 2.1 *follows from an observation of Redekopp* [13]. *Namely, one has*

$$G(t,x,y) = -\frac{1}{3t}\mathrm{Ai}(\zeta)\,\mathrm{Ai}'(\zeta),$$

where Ai *is the Airy function and*

$$\zeta = c_1\frac{x}{t^{1/3}} + c_2\frac{y^2}{t^{4/3}}$$

for some real constants $c_1 > 0$ *and* $c_2 > 0$. *Thus* $G(t,x,y) = \frac{\partial}{\partial x}A(t,x,y)$ *with*

$$A(t,x,y) = -\frac{1}{6c_1 t^{2/3}}\,\mathrm{Ai}^2\Big(c_1\frac{x}{t^{1/3}} + c_2\frac{y^2}{t^{4/3}}\Big)$$

and

$$u = \frac{\partial A}{\partial x}\star\varphi = \frac{\partial}{\partial x}\big(A\star\varphi\big),$$

which proves the claim for the KP-II equation (the fact that $\lim_{|x|\to\infty} A(t,x,y) = 0$ *results from a well-known decay property of the Airy function). A similar explicit computation does not seem to be valid for the classical KP-I equation or for KP-type equations with general symbols.*

Proof of Theorem 2.1. We will consider only the case $\varepsilon = 1$ in (2.1). The analysis in the case $\varepsilon = -1$ is analogous. It is plainly sufficient to consider only the case $t > 0$. We have

$$G(t,x,y) = (2\pi)^{-2}\int_{\mathbb{R}^2} e^{i(x\xi+y\eta)+it(\xi|\xi|^\alpha-\eta^2/\xi)}d\xi d\eta\,, \tag{2.3}$$

where the last integral has the usual interpretation of a generalized Riemann integral. We first check that $G(t,x,y)$ is a continuous function of x and y. By the change of variables

$$\eta' = \frac{t^{1/2}}{|\xi|^{1/2}}\eta,$$

we obtain

$$\begin{aligned}
G(t,x,y) &= \frac{c}{t^{1/2}}\int_{\mathbb{R}_\xi}|\xi|^{1/2}\left(\int_{\mathbb{R}_\eta} e^{i(y/t^{1/2})|\xi|^{1/2}\eta - i\mathrm{sgn}(\xi)\eta^2}d\eta\right)e^{ix\xi+it\xi|\xi|^\alpha}d\xi\\
&= \frac{c}{t^{1/2}}\int_{\mathbb{R}} e^{-i(\mathrm{sgn}(\xi))\frac{\pi}{4}}\,|\xi|^{1/2}e^{iy^2\xi/4t}\,e^{ix\xi+it\xi|\xi|^\alpha}d\xi\\
&= \frac{c}{t^{1/2}}\int_{\mathbb{R}} e^{-i(\mathrm{sgn}(\xi))\frac{\pi}{4}}\,|\xi|^{1/2}e^{i\xi(x+y^2/4t)}\,e^{it\xi|\xi|^\alpha}d\xi\\
&= \frac{c}{t^{\frac12+\frac{3}{2(\alpha+1)}}}\int_{\mathbb{R}} e^{-i(\mathrm{sgn}(\xi))\frac{\pi}{4}}\,|\xi|^{1/2}\exp\left(i\xi\left(\frac{x}{t^{\frac{1}{\alpha+1}}}+\frac{y^2}{4t^{\frac{\alpha+2}{\alpha+1}}}\right)\right)e^{i\xi|\xi|^\alpha}d\xi\,.
\end{aligned}$$

Let us define

$$H(\lambda) = c\int_{\mathbb{R}} e^{-i(\mathrm{sgn}(\xi))\frac{\pi}{4}}\,|\xi|^{1/2}e^{i\lambda\xi}\,e^{i\xi|\xi|^{\alpha}}d\xi\,.$$

Then H is continuous in λ. We will consider only the worst case $\lambda \le 0$. The phase $\varphi(\xi) = i(\lambda\xi + \xi|\xi|^{\alpha})$ then has two critical points $\pm\xi_{\alpha}$, where $\xi_{\alpha} = \left(\frac{\mu}{\alpha+1}\right)^{1/\alpha}$, $\mu = -\lambda$. We write, for $\varepsilon > 0$ small enough

$$H(\lambda) = \int_{-\infty}^{-\xi_{\alpha}-\varepsilon} + \int_{-\xi_{\alpha}-\varepsilon}^{\xi_{\alpha}+\varepsilon} + \int_{\xi_{\alpha}+\varepsilon}^{\infty} := I_1(\lambda) + I_2(\lambda) + I_3(\lambda)\,.$$

Clearly $I_2(\lambda)$ is a continuous function of λ. We consider only $I_3(\lambda)$,

$$\begin{aligned} I_3(\lambda) &= c\int_{\xi_{\alpha}+\varepsilon}^{\infty} \frac{\xi^{1/2}}{\varphi'(\xi)}\,\frac{d}{d\xi}\left[e^{\varphi(\xi)}\right]d\xi \\ &= c\left[\frac{\xi^{1/2}e^{\varphi(\xi)}}{\lambda+\xi^{\alpha}(\alpha+1)}\right]_{\xi_{\alpha}+\varepsilon}^{\infty} + c\int_{\xi_{\alpha}+\varepsilon}^{\infty}\left[\frac{1}{2(\lambda+\xi^{\alpha}(\alpha+1))\xi^{1/2}} - \frac{\alpha(\alpha+1)\xi^{\alpha-1/2}}{(\lambda+(\alpha+1)\xi^{\alpha})^2}\right]e^{\varphi(\xi)}d\xi, \end{aligned}$$

which for $\alpha > 1/2$ defines a continuous function of λ. Hence the integral (2.3) is a continuous function of (x, y) which coincides with the inverse Fourier transform (in $\mathcal{S}'(\mathbb{R}^2)$) of $\exp(it(\xi|\xi|^{\alpha} - \eta^2/\xi))$.

We next set for $t > 0$

$$A(t,x,y) \equiv (2\pi)^{-2}\int_{\mathbb{R}^2}\frac{1}{i\xi}e^{i(x\xi+y\eta)+it(\xi|\xi|^{\alpha}-\eta^2/\xi)}d\xi d\eta\,.$$

The last integral is clearly not absolutely convergent not only at infinity but also for ξ near zero. Nevertheless, the oscillations involved in its definition will allow us to show that $A(t, x, y)$ is in fact a continuous function. By the change of variables

$$\eta' = \frac{t^{1/2}}{|\xi|^{1/2}}\eta,$$

we obtain

$$\begin{aligned} A(t,x,y) &= \frac{c}{t^{1/2}}\int_{\mathbb{R}_{\xi}}\frac{\mathrm{sgn}(\xi)}{|\xi|^{1/2}}\left(\int_{\mathbb{R}_{\eta}}e^{i(y/t^{1/2})|\xi|^{1/2}\eta - i\mathrm{sgn}(\xi)\eta^2}d\eta\right)e^{ix\xi+it\xi|\xi|^{\alpha}}d\xi \\ &= \frac{c}{t^{1/2}}\int_{\mathbb{R}}\frac{(\mathrm{sgn}(\xi))e^{-i(\mathrm{sgn}(\xi))\frac{\pi}{4}}}{|\xi|^{1/2}}e^{iy^2\xi/4t}\,e^{ix\xi+it\xi|\xi|^{\alpha}}d\xi \\ &= \frac{c}{t^{\frac{\alpha+2}{2(\alpha+1)}}}\int_{\mathbb{R}}\frac{(\mathrm{sgn}(\xi))e^{-i(\mathrm{sgn}(\xi))\frac{\pi}{4}}}{|\xi|^{1/2}}\exp\left(i\xi\left(\frac{x}{t^{\frac{1}{\alpha+1}}} + \frac{y^2}{4t^{\frac{\alpha+2}{\alpha+1}}}\right)\right)e^{i\xi|\xi|^{\alpha}}d\xi\,. \end{aligned}$$

We now need the following lemma.

LEMMA 2.1. *Let for $\alpha > 0$*

$$F(\lambda) = \int_{\mathbb{R}}\frac{(\mathrm{sgn}(\xi))e^{-i(\mathrm{sgn}(\xi))\frac{\pi}{4}}}{|\xi|^{1/2}}e^{i\lambda\xi+i\xi|\xi|^{\alpha}}d\xi\,.$$

Then F is a continuous function which tends to zero as $|\lambda| \to +\infty$.

Proof. Write F as

$$F(\lambda) = \int_{|\xi|\leq 1} + \int_{|\xi|\geq 1} := F_1(\lambda) + F_2(\lambda)\,.$$

Since $|\xi|^{-1/2}$ in integrable near the origin, by the Riemann–Lebesgue lemma $F_1(\lambda)$ is continuous and

$$\lim_{|\lambda|\to\infty} F_1(\lambda) = 0\,.$$

We consider two cases in the analysis of $F_2(\lambda)$.

Case 1. $\lambda \geq -1$.

After an integration by parts, we obtain that

$$\begin{aligned}(2.4)\quad F_2(\lambda) = &\frac{c\cos(\lambda+1-\frac{\pi}{4})}{\lambda+\alpha+1}\\ &+ c\int_1^\infty \cos\Big(\lambda\xi+\xi^{\alpha+1}-\frac{\pi}{4}\Big)\,\frac{\lambda+(\alpha+1)(2\alpha+1)\xi^\alpha}{\xi^{3/2}(\lambda+(\alpha+1)\xi^\alpha)^2}\,d\xi\,.\end{aligned}$$

The first term is clearly a continuous function of λ which tends to zero as $\lambda\to\infty$. Observing that

$$0 \leq \frac{\lambda+(\alpha+1)(2\alpha+1)\xi^\alpha}{\xi^{3/2}(\lambda+(\alpha+1)\xi^\alpha)^2} \leq C_\alpha \xi^{-3/2}$$

uniformly with respect to $\xi\geq 1$ and $\lambda\geq -1$, we deduce from the dominated convergence theorem that the right-hand side of (2.4) is a continuous function of λ for $\lambda\geq -1$. On the other hand, for $\lambda\geq 1$,

$$\frac{\lambda+(\alpha+1)(2\alpha+1)\xi^\alpha}{\xi^{3/2}(\lambda+(\alpha+1)\xi^\alpha)^2} \leq \frac{2\alpha+1}{\lambda\xi^{3/2}},$$

and thus the right-hand side of (2.4) tends to zero as $\lambda\to+\infty$.

Case 2. $\lambda\leq -1$.

Set $\lambda=-\mu$ with $\mu\geq 1$. In the integral over $|\xi|\geq 1$ defining $F_2(\lambda)$, we consider only the integration over $[1,+\infty[$. The integration over $]-\infty,-1]$ can be treated in a completely analogous way. We perform the changes of variables

$$\xi \longrightarrow \xi^2$$

and

$$\xi \longrightarrow \mu^{\frac{1}{2\alpha}}\xi$$

to conclude that

$$\widetilde{F_2}(\lambda) := c\int_1^\infty \frac{1}{\xi^{1/2}} e^{i\lambda\xi+i\xi|\xi|^\alpha}\,d\xi = c\mu^{\frac{1}{2\alpha}}\int_{\mu^{-\frac{1}{2\alpha}}}^\infty e^{i\mu^{1+\frac{1}{\alpha}}\left[\xi^{2(\alpha+1)}-\xi^2\right]}\,d\xi.$$

Let us set

$$\varphi(\xi) = \xi^{2(\alpha+1)} - \xi^2\,.$$

Then

$$\varphi'(\xi) = 2\xi[(\alpha+1)\xi^{2\alpha} - 1]\,.$$

Let us split

$$\widetilde{F_2}(\lambda) = c\mu^{\frac{1}{2\alpha}} \int_{\mu^{-\frac{1}{2\alpha}}}^{1} + c\mu^{\frac{1}{2\alpha}} \int_{1}^{\infty} := I_1(\mu) + I_2(\mu)\,.$$

Since $\varphi'(\xi)$ does not vanish for $\xi \geq 1$, we can integrate by parts, which gives

$$I_2(\mu) = \frac{1}{2i\mu^{1+\frac{1}{2\alpha}}} \left(\frac{c}{\alpha} + c\int_1^\infty e^{i\mu^{1+\frac{1}{\alpha}}\left[\xi^{2(\alpha+1)}-\xi^2\right]}\ \frac{(\alpha+1)(2\alpha+1)\xi^{2\alpha}-1}{\xi^2((\alpha+1)\xi^{2\alpha}-1)^2} d\xi \right),$$

which is a continuous function of $\mu \geq 1$ thanks to the dominated convergence theorem. Moreover, it clearly tends to zero as $\mu \to +\infty$.

Let us next analyze $I_1(\mu)$. We first observe that thanks to the dominated convergence theorem, $I_1(\mu)$ is a continuous function of μ. It remains to prove that $I_1(\mu) \to 0$ as $\mu \to \infty$. For $\xi \in [\mu^{-\frac{1}{2\alpha}}, 1]$, the phase φ has a critical point, and a slightly more delicate argument is needed. Compute

$$\varphi''(\xi) = 2[(\alpha+1)(2\alpha+1)\xi^{2\alpha} - 1]\,.$$

Observe that $\varphi'(\xi)$ is vanishing only at zero and

$$\xi_1(\alpha) := \left(\frac{1}{\alpha+1}\right)^{\frac{1}{2\alpha}} \quad .$$

Next, we notice that $\varphi''(\xi)$ is vanishing at

$$\xi_2(\alpha) := \left(\frac{1}{(\alpha+1)(2\alpha+1)}\right)^{\frac{1}{2\alpha}} \quad .$$

Clearly $\xi_2(\alpha) < \xi_1(\alpha) < 1$, and we choose a real number δ such that

$$\xi_2(\alpha) < \delta < \xi_1(\alpha) < 1\,.$$

For $\mu \gg 1$, we can split

$$I_1(\mu) = c\mu^{\frac{1}{2\alpha}} \int_{\mu^{-\frac{1}{2\alpha}}}^{\delta} + c\mu^{\frac{1}{2\alpha}} \int_{\delta}^{1} := J_1(\mu) + J_2(\mu)\ .$$

For $\xi \in [\mu^{-\frac{1}{2\alpha}}, \delta]$, we have the lower bound

$$|\varphi'(\xi)| \geq c\mu^{-\frac{1}{2\alpha}} > 0,$$

and an integration by parts shows that

$$J_1(\mu) = \mu^{\frac{1}{\alpha}}\,\mathcal{O}(\mu^{-1-\frac{1}{\alpha}}) \leq C\mu^{-1},$$

which clearly tends to zero as $\mu \to \infty$. For $\xi \in [\delta, 1]$, we have the minoration

$$|\varphi''(\xi)| \geq c > 0,$$

and therefore we can apply the Van der Corput lemma (see [16, Proposition 2]) to conclude that

$$J_2(\mu) = \mu^{\frac{1}{2\alpha}} \mathcal{O}(\mu^{-\frac{1}{2}-\frac{1}{2\alpha}}) \le C\mu^{-\frac{1}{2}},$$

which tends to zero as $\mu \to \infty$. This completes the proof of Lemma 2.1. ∎

It is now easy to check that $\partial_x A = G$ in the sense of distributions. Since both A and G are continuous, we deduce that A has a classical derivative with respect to x which is equal to G. Finally, since $\varphi \in L^1(\mathbb{R}^2)$, applying Lemma 2.1 and the Lebesgue theorem completes the proof of Theorem 2.1. ∎

2.2. The nonlinear case. After a change of frame we can eliminate the u_x term and reduce the Cauchy problem for (1.7) to

$$(u_t + uu_x - Lu_x)_x + u_{yy} = 0, \quad u(0,x,y) = \varphi(x,y). \tag{2.5}$$

In order to state our result concerning (2.5), for $k \in \mathbb{N}$, we denote by $H^{k,0}(\mathbb{R}^2)$ the Sobolev space of $L^2(\mathbb{R}^2)$ functions $u(x,y)$ such that $\partial_x^k u \in L^2(\mathbb{R}^2)$.

Theorem 2.2. *Assume that $\alpha > 1/2$. Let $\varphi \in L^1(\mathbb{R}^2) \cap H^{2,0}(\mathbb{R}^2)$ and*

$$u \in C([0,T]\,;\, H^{2,0}(\mathbb{R}^2)) \tag{2.6}$$

be a distributional solution of (2.5). Then, for every $t \in (0,T]$, $u(t,\cdot,\cdot)$ is a continuous function of x and y which satisfies

$$\int_{-\infty}^{\infty} u(t,x,y)dx = 0 \quad \forall y \in \mathbb{R},\ \ \forall t \in (0,T]$$

in the sense of generalized Riemann integrals. Moreover, $u(t,x,y)$ is the derivative with respect to x of a C_x^1 continuous function which vanishes as $x \to \pm\infty$ for every fixed $y \in \mathbb{R}$ and $t \in [0,T]$.

Remark 2.3. *The case $\alpha = 2$ corresponds to the classical KP-I, KP-II equations. In the case of the KP-II, we have global solutions for data in $L^1(\mathbb{R}^2) \cap H^{2,0}(\mathbb{R}^2)$ (see [5]). Thus Theorem 2.2 displays a striking smoothing effect of the KP-II equation: $u(t,\cdot,\cdot)$ becomes a continuous function of x and y (with zero mean in x) for $t \neq 0$ (note that $L^1(\mathbb{R}^2) \cap H^{2,0}(\mathbb{R}^2)$ is not included in $C^0(\mathbb{R}^2)$). A similar comment is valid for the local solutions of the KP-I equation in [10] and more especially in [12].*

Remark 2.4. *The numerical simulations in [9] display clearly the phenomena described in Theorem 2.2 in the case of the KP-I equation.*

Proof of Theorem 2.2. Under our assumption on u, one has the Duhamel representation

$$u(t) = S(t)\varphi - \int_0^t S(t-s)\big(u(s)u_x(s)\big)ds, \tag{2.7}$$

where

$$\int_0^t S(t-s)\big(u(s)u_x(s)\big)ds = \int_0^t \partial_x \left(\int_{\mathbb{R}^2} A(x-x', y-y', t-s)(uu_x)(x',y',s)dx'dy' \right) ds\,.$$

From Theorem 2.1, it suffices to consider only the integral term in the right-hand side of (2.7). Using the notations of Lemma 2.1,

$$A(x-x',y-y',t-s)=\frac{c}{(t-s)^{\frac{\alpha+2}{2(\alpha+1)}}}F\left(\frac{x-x'}{(t-s)^{\frac{1}{\alpha+1}}}+\frac{(y-y')^2}{4(t-s)^{\frac{\alpha+2}{\alpha+1}}}\right).$$

Recall that F is a continuous and bounded function on $\mathbb{R}$. Next, we set

$$I(x,y,t-s,s)\equiv\partial_x\left(\int_{\mathbb{R}^2}A(x-x',y-y',t-s)(uu_x)(x',y',s)dx'dy'\right).$$

Using the Lebesgue differentiation theorem and the assumption (2.6), we can write

$$I(x,y,t-s,s)=\int_{\mathbb{R}^2}A(x-x',y-y',t-s)\partial_x(uu_x)(x',y',s)dx'dy'.$$

Moreover, for $\alpha>0$,

$$\frac{\alpha+2}{2(\alpha+1)}<1,$$

and therefore I is integrable in s on $[0,t]$. Therefore, by the Lebesgue differentiation theorem,

$$\int_0^t\int_{\mathbb{R}^2}A(x-x',y-y',t-s)(uu_x)(x',y',s)dx'dy'ds \tag{2.8}$$

is a C_x^1 function and

$$\int_0^t S(t-s)\big(u(s)u_x(s)\big)ds$$
$$=\partial_x\left(\int_0^t\int_{\mathbb{R}^2}A(x-x',y-y',t-s)(uu_x)(x',y',s)dx'dy'ds\right).$$

Let us finally show that for fixed y and t the function (2.8) tends to zero as x tends to $\pm\infty$. For that purpose, it suffices to apply the Lebesgue dominated convergence theorem to the integral in s,x',y'. Indeed, for fixed s,x',y', the function under the integral tends to zero as x tends to $\pm\infty$ thanks to the linear analysis. On the other hand, using Lemma 2.1, we can write

$$|A(x-x',y-y',t-s)(uu_x)(x',y',s)|\le\frac{c}{(t-s)^{\frac{\alpha+2}{2(\alpha+1)}}}|(uu_x)(x',y',s)|.$$

Thanks to the assumptions on u, the right-hand side of the above inequality is integrable in s,x',y' and independent of x. Thus we can apply the Lebesgue dominated convergence theorem to conclude that the function (2.8) tends to zero as x tends to $\pm\infty$. This completes the proof of Theorem 2.2. □

REMARK 2.5. *If $\alpha>2$, the assumptions can be weakened to $\varphi\in L^1(\mathbb{R}^2)\cap H^{1,0}(\mathbb{R}^2)$ and $u\in C([0,T];H^{1,0}(\mathbb{R}^2))$. This result follows from the fact that the fundamental solution G writes*

$$G(t,x,y)=\frac{c}{t^{1/2+3/2(\alpha+1)}}B(t,x,y),$$

where $B\in L^\infty$ and $1/2+3/(2(\alpha+1))<1$ for $\alpha>2$.

3. KP–BBM-type equations.

3.1. The linear case. We consider the Cauchy problem

$$(u_t + u_x + Lu_t)_x + u_{yy} = 0, \quad u(0,x,y) = \varphi(x,y)\,, \tag{3.1}$$

where L is given by (2.2) with $\varepsilon = 1$. A simple computation shows that the fundamental solution is given by

$$G(t,x,y) = \mathcal{F}^{-1}\big[e^{-i\frac{t}{1+|\xi|^\alpha}(\xi+\eta^2/\xi)}\big]\,.$$

Due to the bad oscillatory properties of the phase, we have to modify slighty the statement of Theorem 2.1.

THEOREM 3.1. *Assume that $\alpha > 0$. Let φ be such that $(I-\partial_x^2)^{\beta/2}\varphi \in L^1(\mathbb{R}^2)$ with $\beta > (\alpha+3)/2$. Then*

$$u(t,\cdot,\cdot) \equiv S(t)\varphi = G \star \varphi$$

can be written as

$$\frac{\partial}{\partial x}\big(A \star (I-\partial_x^2)^{\beta/2}\varphi\big),$$

where

$$A = \partial_x^{-1}(I-\partial_x^2)^{-\beta/2}G$$

and

$$A(t,\cdot,\cdot) \in C(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2) \cap C_x^1(\mathbb{R}^2)\,.$$

Moreover, $(A \star (I-\partial_x^2)^{\beta/2}\varphi)(t,x,y)$ is, for fixed $t \neq 0$, continuous in x and y and satisfies

$$\lim_{|x|\to\infty}(A \star (I-\partial_x^2)^{\beta/2}\varphi)(t,x,y) = 0$$

for any $y \in \mathbb{R}$ and $t \neq 0$. Thus,

$$\int_{-\infty}^{\infty} u(t,x,y)\,dx = 0 \quad \forall y \in \mathbb{R}, \;\; \forall t \neq 0$$

in the sense of generalized Riemann integrals.

REMARK 3.1. *The assumption $(I-\partial_x^2)^{\beta/2}\varphi \in L^1(\mathbb{R}^2)$ is natural in the context of KP–BBM problems, in view of the weak dispersive properties of the BBM free evolution.*

Proof. We set

$$\tilde{G}(t,x,y) = \int_{\mathbb{R}^2} \frac{1}{(1+|\xi|^2)^{\beta/2}}\; e^{-\frac{it}{1+|\xi|^\alpha}(\xi+\eta^2/\xi)} e^{ix\xi+iy\eta} d\xi d\eta\,.$$

Setting

$$\eta' = \frac{t^{1/2}\eta}{|\xi|^{1/2}(1+|\xi|^\alpha)^{1/2}}\,,$$

we obtain

$$
\begin{aligned}
&\tilde{G}(t,x,y)\\
&= \frac{c}{t^{1/2}} \int_{\mathbb{R}_\xi} |\xi|^{1/2} \frac{(1+|\xi|^\alpha)^{1/2}}{(1+|\xi|^2)^{\beta/2}} \left(\int_{\mathbb{R}_\eta} e^{-i\operatorname{sgn}(\xi)\eta^2} e^{i(y\eta/t^{1/2})|\xi|^{1/2}(1+|\xi|^\alpha)^{1/2}} d\eta \right) e^{ix\xi - i\frac{t\xi}{1+|\xi|^\alpha}} d\xi\\
&= \frac{c}{t^{1/2}} \int_{\mathbb{R}} |\xi|^{1/2} e^{-i(\operatorname{sgn}(\xi))\frac{\pi}{4}} \frac{(1+|\xi|^\alpha)^{1/2}}{(1+|\xi|^2)^{\beta/2}} e^{-i\frac{t\xi}{1+|\xi|^\alpha}} e^{i\lambda\xi} e^{i(y^2/4t)\xi|\xi|^\alpha} d\xi\,,
\end{aligned}
$$

where $\lambda = x + y^2/4t$. Setting

$$
a_t(\xi) = |\xi|^{1/2} e^{-i(\operatorname{sgn}(\xi))\frac{\pi}{4}} \frac{(1+|\xi|^\alpha)^{1/2}}{(1+|\xi|^2)^{\beta/2}} e^{-it\frac{\xi}{1+|\xi|^\alpha}}\,,
$$

we clearly have

$$
\tilde{G}(t,x,y) = \frac{c}{t^{1/2}} \int a_t(\xi) e^{i\lambda\xi} e^{i(y^2/4t)\xi|\xi|^\alpha} d\xi\,.
$$

We have the following lemma.

Lemma 3.1. *Let us fix $y \in \mathbb{R}$ and $t > 0$. Set*

$$
F(\lambda) := \int a_t(\xi) e^{i\lambda\xi} e^{i(y^2/4t)\xi|\xi|^\alpha} d\xi\,.
$$

Then F is a continuous function such that

$$
\lim_{|\lambda|\to\infty} F(\lambda) = 0\,.
$$

Proof. It suffices to apply the Riemann–Lebesgue lemma since $a \in L^1(\mathbb{R})$ when $\beta > (\alpha+3)/2$. $\square$

Next, we set

$$
\tilde{A}(t,x,y) = -i \int_{\mathbb{R}^2} \frac{1}{\xi(1+|\xi|^2)^{\beta/2}}\ e^{-\frac{it}{1+|\xi|^\alpha}(\xi+\eta^2/\xi)} e^{ix\xi+iy\eta} d\xi d\eta\,.
$$

Similarly to above, we set

$$
\eta' = \frac{\eta t^{1/2}}{|\xi|^{1/2}(1+|\xi|^\alpha)^{1/2}}\,,
$$

and therefore

$$
\begin{aligned}
&\tilde{A}(t,x,y)\\
&= \frac{c}{t^{1/2}} \int_{\mathbb{R}_\xi} \frac{\operatorname{sgn}(\xi)}{|\xi|^{1/2}} \frac{(1+|\xi|^\alpha)^{1/2}}{(1+|\xi|^2)^{\beta/2}} \left(\int_{\mathbb{R}_\eta} e^{-i\operatorname{sgn}(\xi)\eta^2} e^{iy\eta/t^{1/2}|\xi|^{1/2}(1+|\xi|^\alpha)^{1/2}} d\eta \right) e^{ix\xi - i\frac{t\xi}{1+|\xi|^\alpha}} d\xi\\
&= \frac{c}{t^{1/2}} \int_{\mathbb{R}} \frac{\operatorname{sgn}(\xi)}{|\xi|^{1/2}} e^{-i(\operatorname{sgn}(\xi))\frac{\pi}{4}} \frac{(1+|\xi|^\alpha)^{1/2}}{(1+|\xi|^2)^{\beta/2}} e^{-it\frac{\xi}{1+|\xi|^\alpha}} e^{i\lambda\xi} e^{i(y^2/4t)\xi|\xi|^\alpha} d\xi\,,
\end{aligned}
$$

where $\lambda = x + y^2/4t$. Setting

$$
\tilde{a}_t(\xi) = \frac{\operatorname{sgn}(\xi)}{|\xi|^{1/2}} e^{-i(\operatorname{sgn}(\xi))\frac{\pi}{4}} \frac{(1+|\xi|^\alpha)^{1/2}}{(1+|\xi|^2)^{\beta/2}} e^{-i\frac{t\xi}{1+|\xi|^\alpha}}\,,
$$

we clearly have

$$\tilde{A}(t,x,y) = \frac{c}{t^{1/2}} \int \tilde{a}_t(\xi) e^{i\lambda\xi}\, e^{i(y^2/4t)\xi|\xi|^\alpha} d\xi\,.$$

LEMMA 3.2. *Let us fix $y \in \mathbb{R}$ and $t > 0$. Set*

$$F_1(\lambda) := \int \tilde{a}_t(\xi) e^{i\lambda\xi}\, e^{i(y^2/4t)\xi|\xi|^\alpha} d\xi\,.$$

Then F_1 is a continuous function such that

$$\lim_{|\lambda|\to\infty} F_1(\lambda) = 0\,.$$

Proof. It suffices to apply the Riemann–Lebesgue lemma since $\tilde{a}_t \in L^1(\mathbb{R})$. ∎

The proof of Theorem 3.1 is now straightforward. ∎

3.2. The nonlinear case. We investigate the Cauchy problem

$$(u_t + u_x + uu_x + Lu_t)_x + u_{yy} = 0, \quad u(0,x,y) = \varphi(x,y)\,. \tag{3.2}$$

THEOREM 3.2. *Let $\alpha > 0$, $k > \frac{\alpha+3}{4}$. Assume that $(I - \partial_x^2)^k \varphi \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$. Let u be a solution of (3.2) such that*

$$u \in C([0,T]\,;\, H^{2k+1,0}(\mathbb{R}^2))\,.$$

Then, for any $t \in (0,T]$, $u(t,x,y)$ is a continuous function in x and y and satisfies

$$\int_{-\infty}^{\infty} u(t,x,y)dx = 0 \quad \forall y \in \mathbb{R},\ \ \forall t \in (0,T]$$

in the sense of generalized Riemann integrals. In fact $u(t,x,y)$ is the derivative with respect to x of a C_x^1 continuous function which vanishes as $x \to \pm\infty$ for any fixed y and $t \in (0,T]$.

Proof. Again we use the Duhamel formula

$$u(t) = S(t)\varphi - \int_0^t S(t-s)u(s)u_x(s)ds,$$

where

$$S(t) = e^{it(I+L)^{-1}(\partial_x + \partial_x^{-1}\partial_y^2)}\,.$$

By Theorem 3.1, it suffices to consider the integral term in the Duhamel formula. We have

$$\int_0^t S(t-s)u(s)u_x(s)ds = \frac{1}{2}\int_0^t \frac{\partial}{\partial x}\Big(S(t-s)u^2(s)\Big)ds = \frac{1}{2}\frac{\partial}{\partial x}\int_0^t S(t-s)u^2(s)ds.$$

To justify the last equality, we have to check that $S(t-s)u(s)u_x(s)$ is dominated by a $L^1(0,t)$ function uniformly in (x,y). We write

$$S(t-s)(u(s)u_x(s)) = \tilde{A}(x,y,t-s) \star (I - \partial_x^2)^k u(s)u_x(s), \tag{3.3}$$

where

$$\tilde{A}(x,y,t-s) = \int_{\mathbb{R}^2} \frac{1}{(1+|\xi|^2)^k} e^{-\frac{i(t-s)}{1+|\xi|^\alpha}(\xi+\eta^2/\xi)} e^{ix\xi+iy\eta} d\xi d\eta \,.$$

Proceeding as in the beginning of the proof of Theorem 3.1, it follows that

$$\tilde{A}(x,y,t-s) = \frac{c}{(t-s)^{1/2}} \int_{\mathbb{R}} \frac{e^{-i(\mathrm{sgn}(\xi))\frac{\pi}{4}} |\xi|^{1/2}(1+|\xi|^\alpha)^{1/2}}{(1+\xi^2)^k} e^{-i\frac{\xi(t-s)}{1+|\xi|^\alpha}} e^{i\lambda\xi} e^{i(y^2/4(t-s))\xi|\xi|^\alpha} d\xi \,,$$

where $\lambda = x + y^2/4(t-s)$. Since $k > \frac{\alpha+3}{4}$, the integral in ξ defines a continuous bounded function in x, y, t, s by the Riemann–Lebesgue theorem. It follows that

$$|S(t-s)u(s)u_x(s)| \le \frac{c}{(t-s)^{1/2}} \|(I-\partial_x^2)^k(uu_x)\|_{L^1(\mathbb{R}^2)} \le \frac{c}{(t-s)^{1/2}}$$

since $u \in C([0,T]\,;\, H^{2k+1,0}(\mathbb{R}^2))$. Since the function

$$\frac{e^{-i(\mathrm{sgn}(\xi))\frac{\pi}{4}} |\xi|^{1/2}(1+|\xi|^\alpha)^{1/2}}{(1+\xi^2)^k} e^{-i\frac{\xi(t-s)}{1+\xi^\alpha}}$$

belongs to $L^1(\mathbb{R}_\xi)$, we can use the Riemann–Lebesgue lemma to obtain that for fixed y, t, and s the function $\tilde{A}(x,y,t-s)$ tends to zero as x tends to $\pm\infty$. Moreover, the absolute value of $\tilde{A}(x,y,t-s)$ is bounded by $c|t-s|^{-1/2}$. Thus as in the proof of Theorem 2.2, we can apply the Lebesgue dominated convergence theorem to conclude that

$$\lim_{x\to\pm\infty} \int_0^t S(t-s)u^2(s)ds = 0$$

for any fixed $y \in \mathbb{R}$ and $t \in (0,T]$. This achieves the proof of Theorem 3.2. $\square$

REMARK 3.2. *For large values of α, one can relax the assumptions on k in the hypothesis for u by simply using the H^s unitary property of $S(t)$.*

4. Extensions. With the price of some technicalities, one could consider symbols $c(\xi)$ in (1.2) which behave like $|\xi|^\alpha$ at infinity but which are not homogeneous.

Let us finally comment briefly on the three-dimensional case. For simplicity, we consider only the KP-type equations:

$$(u_t - Lu_x)_x + u_{yy} + u_{zz} = 0, \quad u(0,x,y,z) = \varphi(x,y,z) \tag{4.1}$$

with L given by (2.2). Following the lines of the proof of Theorem 2.1, we find that the fundamental solution G can be expressed as $G = \partial_x A$, where

$$A(t,x,y,z) = \frac{c}{t^{1+\frac{2}{\alpha+2}}} \int_{\mathbb{R}} \mathrm{sgn}(\xi)\, e^{-i(\mathrm{sgn}(\xi))\frac{\pi}{4}} e^{i\xi(x/t^{1/(\alpha+1)}+(y^2+z^2)/4t)} \, e^{i\xi|\xi|^\alpha} \, d\xi \,.$$

We first notice that, when $\alpha > 1$, G is a well-defined continuous function of (x,y,z). Actually the proof follows the same lines as the two-dimensional case. Let

$$F(\lambda) = \int_{\mathbb{R}} \mathrm{sgn}(\xi) e^{-i(\mathrm{sgn}(\xi))\frac{\pi}{4}} e^{i\lambda\xi} \, e^{i\xi|\xi|^\alpha} \, d\xi \,.$$

By a result of [15], $F(\lambda)$ is a continuous function which tends to zero as $|\lambda| \to \infty$, provided $\alpha > 1$, (notice that this excludes the case $\alpha = 1$ which would correspond to the three-dimensional generalizations of the Benjamin–Ono equation). We thus obtain the exact counterpart of Theorem 2.1 in the three-dimensional case when $\alpha > 1$ (this includes the three-dimensional usual KP equations).

REFERENCES

[1] M. J. Ablowitz and J. Villarroel, *On the Kadomtsev-Petviashvili equation and associated constraints*, Stud. Appl. Math., 85 (1991), pp. 195–213.

[2] M. J. Ablowitz and J. Villarroel, *Initial value problems and solutions of the Kadomtsev-Petviashvili equations*, in New Trends in Integrability and Partial Solvability, NATO Sci. Ser. II Math. Phys. Chem. 132, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004, pp. 1–47.

[3] M. Boiti, F. Pempinelli, and A. Pogrebkov, *On the KP-I equation with small initial data*, Inverse Problems, 10 (1994), pp. 505–519.

[4] M. Boiti, F. Pempinelli, and A. Pogrebkov, *Properties of solutions of KP-I equation*, J. Math. Phys., 35 (1994), pp. 4683–4718.

[5] J. Bourgain, *On the Cauchy problem for the Kadomtsev-Petviashvili equation*, Geom. Funct. Anal., 3 (1993), pp. 315–341.

[6] A. S. Fokas and L.-Y. Sung, *The Cauchy problem for the KP-I equation without the zero mass constraint*, Math. Proc. Cambridge Philos. Soc., 125 (1999), pp. 113–138.

[7] B. B. Kadomtsev and V. I. Petviashvili, *On the stability of solitary waves in weakly dispersive media*, Soviet Phys. Dokl., 15 (1970), pp. 539–541.

[8] C. Kenig, *On the local and global well-posedness for the KP-I equation*, Ann. Inst. H. Poinaré Anal. Non Linéaire, 21 (2004), pp. 827–838.

[9] C. Klein, C. Sparber, and P. Markowich, *Numerical study of oscillatory regimes in the Kadomtsev-Petviashvili equation*, J. Nonlinear Sci., to appear; also available online from http://arxiv.org/abs/math-ph/0601025.

[10] L. Molinet, J. C. Saut, and N. Tzvetkov, *Well-posedness and ill-posedness results for the Kadomtsev-Petviashvili-I equation*, Duke Math. J., 115 (2002), pp. 353–384.

[11] L. Molinet, J. C. Saut, and N. Tzvetkov, *Global well-posedness for the KP-I equation*, Math. Ann., 324 (2002), pp. 255–275. Correction: Math. Ann., 328 (2004), pp. 707–710.

[12] L. Molinet, J. C. Saut, and N. Tzvetkov, *Global well-posedness for the KP-I equation on the background of a non localized solution*, Comm. Math. Phys., 272 (2007), pp. 775–810.

[13] L. Redekopp, *Similarity solutions of some two dimensional nonlocal wave evolution equations*, Stud. Appl. Math., 63 (1980), pp. 185–207.

[14] J. C. Saut and N. Tzvetkov, *Global well-posedness for the KP-BBM equations*, AMRX Appl. Math. Res. Express, 1 (2004), pp. 1–16.

[15] A. Sidi, C. Sulem, and P. L. Sulem, *On the long time behavior of a generalized KdV equation*, Acta Appl. Math., 7 (1986), pp. 35–47.

[16] E. Stein, *Oscillatory integrals in Fourier analysis*, in Beijing Lectures in Harmonic Analysis, Princeton University Press, Princeton, NJ, 1986, pp. 307–355.

[17] L.-Y. Sung, *Square integrability and uniqueness of the solutions of the KP-I equation*, Math. Phys. Anal. Geom., 3 (1999), pp. 1–24.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 642–668

TWO-SCALE HOMOGENIZATION FOR EVOLUTIONARY VARIATIONAL INEQUALITIES VIA THE ENERGETIC FORMULATION*

ALEXANDER MIELKE† AND AIDA M. TIMOFTE‡

Abstract. This paper is devoted to two-scale homogenization for a class of rate-independent systems described by the energetic formulation or equivalently by an evolutionary variational inequality. In particular, we treat the classical model of linearized elastoplasticity with hardening. The associated nonlinear partial differential inclusion has periodically oscillating coefficients, and the aim is to find a limit problem for the case in which the period tends to 0. Our approach is based on the notion of *energetic solutions*, which is phrased in terms of a stability condition and an energy balance of an energy-storage functional and a dissipation functional. Using the recently developed method of *weak and strong two-scale convergence* via *periodic unfolding*, we show that these two functionals have a suitable two-scale limit, but now involving the macroscopic variable in the physical domain as well as the microscopic variable in the periodicity cell. Moreover, relying on an abstract theory of Γ-convergence for the energetic formulation using so-called *joint recovery sequences*, it is possible to show that the solutions of the problem with periodicity converge to the energetic solution associated with the limit functionals.

Key words. weak two-scale convergence, strong two-scale convergence, two-scale Γ-limit, rate-independent evolution, energetic formulation, elastoplasticity with hardening

AMS subject classifications. 35J85, 49J40, 74C05, 74Q05

DOI. 10.1137/060672790

1. Introduction.

Our aim is to provide homogenization results for evolutionary variational inequalities of the following type:

$$\forall\, v \in \mathcal{Q}: \ \langle \mathcal{A}q - \ell(t), v - \dot{q} \rangle + \mathcal{R}(v) - \mathcal{R}(\dot{q}) \geq 0. \tag{1.1}$$

Here $\mathcal{Q}$ is a Hilbert space with dual $\mathcal{Q}^*$, and the continuous linear operator $\mathcal{A} : \mathcal{Q} \to \mathcal{Q}^*$ is symmetric and positive definite on the cone on which $\mathcal{R}$ is finite. The forcing ℓ lies in $\mathrm{C}^1([0,T], \mathcal{Q}^*)$, and the dissipation functional $\mathcal{R} : \mathcal{Q} \to [0, \infty)$ is convex, lower semicontinuous, and positively homogeneous of degree 1, i.e., $\mathcal{R}(\gamma q) = \gamma \mathcal{R}(q)$ for all $\gamma \geq 0$ and $q \in \mathcal{Q}$. The latter property of $\mathcal{R}$ leads to rate independence.

Problem (1.1) has many different equivalent formulations. We have chosen the *energetic formulation* for rate-independent hysteresis problems (cf. [MT99, MT04, Mie05]), since it is more flexible in treating general rate-independent energetic material models, including nonconvexities and strong nonlinearities; see [Mie03, Mie04, MT06, FM06]. We hope that our methods simplify and clarify the theory of two-scale convergence and thus provide ideas and tools for solving more general prob-

*Received by the editors October 19, 2006; accepted for publication (in revised form) March 8, 2007; published electronically August 15, 2007. This research was supported by the European Union via the Research Training Network HPRN-CT-2002-00284 *Smart Systems: New Materials, Adaptive Systems and Their Nonlinearities. Modelling, Control and Numerical Simulation.*

http://www.siam.org/journals/sima/39-2/67279.html

†Weierstraß-Institut für Angewandte Analysis und Stochastik, Mohrenstraße 39, 10117 Berlin, Germany, and Institut für Mathematik, Humboldt-Universität zu Berlin, Rudower Chaussee 25, 12489 Berlin (Adlershof) (mielke@wias-berlin.de).

‡Weierstraß-Institut für Angewandte Analysis und Stochastik, Mohrenstraße 39, 10117 Berlin, Germany. Current address: IMAR Institute of Mathematics "Simion Stoilow" of the Romanian Academy, 21, Calea Grivitei Street, 010702-Bucharest, Sector 1, Romania (aida.timofte@imar.ro).
This author was partially supported by the DFG under MATHEON project C18.

lems, including nonlinear elastic effects. A similar development was initiated in [Alb00, Alb03, Nes06, Nes07] based on the theory of monotone operators. This theory allows for rate-dependent effects, but is restricted to linearized elasticity.

The energetic formulation is based on the storage functional $\mathcal{E} : [0,T]\times\mathcal{Q} \to \mathbb{R}$ defined via $\mathcal{E}(t,q) = \frac{1}{2}\langle \mathcal{A}q, q\rangle - \langle \ell(t), q\rangle$ and the dissipation functional $\mathcal{R}$. Thus, homogenization of an evolutionary problem can be reduced to some extent to the homogenization of functionals. A function $q : [0,T] \to \mathcal{Q}$ is called an *energetic solution* associated with the functionals $\mathcal{E}$ and $\mathcal{R}$ if for all $t \in [0,T]$ it satisfies the *global stability condition* (S) and the *energy balance* (E):

$$\text{(S)} \quad \forall\, q \in \mathcal{Q} :\ \mathcal{E}(t,q(t)) \le \mathcal{E}(t,q) + \mathcal{R}(q{-}q(t));$$

$$\text{(E)} \quad \mathcal{E}(t,q(t)) + \int_0^t \mathcal{R}(\dot q(s))\,\mathrm{d}s = \mathcal{E}(0,q(0)) - \int_0^t \langle \dot\ell(s), q(s)\rangle\,\mathrm{d}s.$$

We also say that q solves the *energetic formulation* (S) and (E) associated with $\mathcal{E}$ and $\mathcal{R}$.

Note that rate-independent hysteresis phenomena occur in quite general systems like hyperbolic PDEs for electromagnetism; cf. [Vis94, BS96]. Then hysteresis operators are seen as constitutive relations and are coupled to general balance equations. Here, as a first step, we treat only purely rate-independent processes which are relevant in elastoplasticity and other material models.

In this paper we consider a family of energy functionals $(\mathcal{E}_\varepsilon)_\varepsilon$ and of dissipation functionals $(\mathcal{R}_\varepsilon)_\varepsilon$ which are defined as integrals over a domain $\Omega \subset \mathbb{R}^d$ and where the densities depend periodically on x with a period proportional to ε. More precisely, for a periodicity lattice Λ we denote by $\mathcal{Y} = \mathbb{R}^d/_{\Lambda}$ the periodicity torus. For a tensor-valued mapping $\mathbb{A} : \mathcal{Y} \to \mathrm{Lin}(\mathbb{R}^{d\times d}_{\mathrm{sym}}\times\mathbb{R}^m)$ and a function $\rho : \mathcal{Y}\times\mathbb{R}^m \to [0,\infty]$ we define the functionals

$$\mathcal{E}_\varepsilon(t,u,z) = \int_\Omega \frac{1}{2} \left\langle\!\!\left\langle \mathbb{A}\left(\frac{x}{\varepsilon}\right) \begin{pmatrix} \boldsymbol{e}(u) \\ z \end{pmatrix}, \begin{pmatrix} \boldsymbol{e}(u) \\ z \end{pmatrix} \right\rangle\!\!\right\rangle \mathrm{d}x - \langle \ell(t), u\rangle,$$
$$\mathcal{R}_\varepsilon(z) = \int_\Omega \rho\left(\frac{x}{\varepsilon}, z(x)\right) \mathrm{d}x$$

on the space $\mathcal{Q} = \mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d\times\mathrm{L}^2(\Omega)^m$.

The task is now to describe the limiting behavior of the associated energetic solutions. Because of the nonsmoothness and the hysteretic behavior of the evolution of the systems, it will not be possible to find a homogenized limit equation in the classical sense. This would mean finding limiting functionals defined on Ω again. Instead we will need the so-called two-scale homogenization that decomposes solutions into macroscopic and microscopic behavior.

The classical notion of two-scale convergence was introduced by Nguetseng in 1989 [Ngu89] and further developed by Allaire in 1992 [All92]. It aimed for a better description of sequences of oscillating functions and thus for the derivation of a new homogenization method. In [LNW02], an overview of the main homogenization problems which have been studied by this technique is given. This concept is applied in a variety of quite different applications in continuum mechanics; see, e.g., [BM93, HJM94, Vis96, BLM96, Vis97, Alb00, EKK02, MS02]. Moreover, even in engineering this method is used extensively for numerical simulations. There the unit periodicity cell is usually called a “representative unit cell.”

To explain our results in some detail we introduce a few new notions. The two-scale method relies on a micro-macro-decomposition of points $x \in \mathbb{R}^d$ via

$$x = \mathcal{N}_\varepsilon(x) + \varepsilon \mathcal{R}_\varepsilon(x) \quad \text{with } \mathcal{N}_\varepsilon(x) = \varepsilon \left[\frac{x}{\varepsilon}\right]_\Lambda \text{ and } \mathcal{R}_\varepsilon(x) = \left\{\frac{x}{\varepsilon}\right\}_Y,$$

where $[\widetilde{x}]_\Lambda$ is the closest lattice point to $\widetilde{x}$ and $\{\widetilde{x}\}_Y$ is the remainder; see section 2.1 for the exact details. The decomposition of functions is then done by the so-called *periodic unfolding* introduced in [CDG02, CDD04, CDD06]:

$$(\mathcal{T}_\varepsilon u)(x, y) = u_{\text{ex}}(\mathcal{N}_\varepsilon(x) + \varepsilon y),$$

where u_{ex} is the extension of $u : \Omega \to \mathbb{R}$ by 0 to all of $\mathbb{R}^d$. Thus, functions in $\mathrm{L}^p(\Omega)$ are mapped to functions $U = \mathcal{T}_\varepsilon u \in \mathrm{L}^p(\mathbb{R}^d \times \mathcal{Y})$. Similar ideas, called "dilation operators," were used before in [ADH90]. See also [LSB99] for an application in electrical networks.

In section 2.2 we discuss this periodic unfolding operator together with a newly introduced *folding operator* $\mathcal{F}_\varepsilon : \mathrm{L}^p(\mathbb{R}^d \times \mathcal{Y}) \to \mathrm{L}^p(\Omega)$, which is a kind of pseudoinverse as well as the adjoint operator (when taking the conjugate exponent p). In particular, we give special care to the complications arising from the mismatch of Ω and a finite union of small cells of type $\varepsilon(\lambda + Y)$.

In section 2.3 we introduce our notion of *weak and strong two-scale convergences*:

$$u_\varepsilon \overset{\text{w2}}{\rightharpoonup} U \quad \Longleftrightarrow \quad \mathcal{T}_\varepsilon u_\varepsilon \rightharpoonup U_{\text{ex}} \ \text{ in } \mathrm{L}^p(\mathbb{R}^d \times \mathcal{Y}),$$
$$u_\varepsilon \xrightarrow{\text{s2}} U \quad \Longleftrightarrow \quad \mathcal{T}_\varepsilon u_\varepsilon \to U_{\text{ex}} \ \text{ in } \mathrm{L}^p(\mathbb{R}^d \times \mathcal{Y}).$$

This definition is an adaptation of the definitions in [Vis04] to the case that Ω has a boundary; see also [Vis06b]. Nevertheless, the convergences on the right-hand side are asked to occur in $\mathrm{L}^p(\mathbb{R}^d \times \mathcal{Y})$, since the support of $\mathcal{T}_\varepsilon u$ is in general not contained in $\Omega \times \mathcal{Y}$. We relate our definitions to the ones which are used in [Ngu89, All92, CD99, LNW02] and show that our strengthening makes many relations more natural. In particular, the notions of weak and strong two-scale convergences allow us to prove the convergence of scalar products, namely,

$$u_n \overset{\text{w2}}{\rightharpoonup} U, \ v_n \xrightarrow{\text{s2}} V \quad \Longrightarrow \quad \int_\Omega u_n v_n \, \mathrm{d}x \ \to \ \int_\Omega \int_{\mathcal{Y}} U(x, y) V(x, y) \, \mathrm{d}y \, \mathrm{d}x.$$

Using classical two-scale convergence on bounded domains only allows for a weaker statement (cf. [LNW02, Thm. 11]), since at the boundary losses may occur. Exactly this result will be crucial for our limit procedure in section 4.

In section 2.4 we recall known results on the two-scale limits of sequences of gradients and construct a *gradient folding operator* $\mathcal{G}_\varepsilon : \mathrm{H}^1_0(\Omega) \times \mathrm{L}^2(\Omega, \mathrm{H}^1_{\text{av}}(\mathcal{Y})) \to \mathrm{H}^1_0(\Omega)$ such that for all (u_0, U_1) we have $\nabla \mathcal{G}_\varepsilon(u_0, U_1) \xrightarrow{\text{s2}} \nabla_x u_0 + \nabla_y U_1$ and $\mathcal{G}_\varepsilon(u_0, U_1) \rightharpoonup u_0$ in $\mathrm{H}^1_0(\Omega)$. Based on these results we provide the relevant two-scale Γ-limit results for the functionals $\mathcal{E}_\varepsilon(t, \cdot)$ and $\mathcal{R}_\varepsilon$. Under simple additional assumptions, the two-scale limits are

$$\boldsymbol{E}(t, u_0, U_1, Z) = \int_{\Omega \times \mathcal{Y}} \frac{1}{2} \left\langle\!\!\left\langle \mathbb{A}(y) \begin{pmatrix} \boldsymbol{e}_x(u_0) + \boldsymbol{e}_y(U_1) \\ Z \end{pmatrix}, \begin{pmatrix} \boldsymbol{e}_x(u_0) + \boldsymbol{e}_y(U_1) \\ Z \end{pmatrix} \right\rangle\!\!\right\rangle \mathrm{d}y \, \mathrm{d}x - \langle \ell(t), u_0 \rangle$$
$$\boldsymbol{R}(Z) = \int_{\Omega \times \mathcal{Y}} \rho(y, Z(x, y)) \, \mathrm{d}y \, \mathrm{d}x.$$

The convergence of $\mathcal{E}_\varepsilon$ and $\mathcal{R}_\varepsilon$ to $\boldsymbol{E}$ and $\boldsymbol{R}$ can be seen as a type of two-scale Mosco convergence, i.e., Γ-convergence in the weak and in the strong topology; see [MRS07]. Recovery sequences (also called *realizing sequences* in [JKO94]) in the strong two-scale convergence sense are obtained via our explicit operators $\mathcal{F}_\varepsilon$ and $\mathcal{G}_\varepsilon$.

In section 3 we formulate our rate-independent evolution systems and provide existence and uniqueness theorems for energetic formulations associated with $\mathcal{E}_\varepsilon$ and $\mathcal{R}_\varepsilon$ on the one hand and with $\boldsymbol{E}$ and $\boldsymbol{R}$ on the other hand. We obtain uniform a priori Lipschitz bounds for the energetic solutions $q_\varepsilon = (u_\varepsilon, z_\varepsilon) : [0,T] \to \mathcal{Q}$. The solutions $Q = (u_0, U_1, Z) : [0,T] \to \boldsymbol{Q}$ are defined on the space $\boldsymbol{Q} = \boldsymbol{H}\times\boldsymbol{Z}$ with

$$\boldsymbol{H} = \mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d\times\mathrm{L}^2(\Omega;\mathrm{H}^1_{\mathrm{av}}(\mathcal{Y}))^d, \quad \boldsymbol{Z} = \mathrm{L}^2(\Omega;\mathrm{L}^2(\mathcal{Y}))^m = \mathrm{L}^2(\Omega\times\mathcal{Y})^m,$$

with $\mathrm{H}^1_{\mathrm{av}}(\mathcal{Y}) = \{\, U \in \mathrm{H}^1(\mathcal{Y}) \mid \int_{\mathcal{Y}} U(y)\,\mathrm{d}y = 0\,\}$.

The final section, section 4, establishes the relation between the solutions q_ε and Q. The main result is Theorem 4.3, which states that if the initial data $q_\varepsilon(0)$ strongly two-scale cross-converge to Q^0, written as $q_\varepsilon(0) \xrightarrow{\text{s2c}} Q^0$ and defined as

$$\begin{aligned} &u_\varepsilon \rightharpoonup u_0 \text{ in } \mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d, \qquad \nabla u_\varepsilon \xrightarrow{\text{s2}} \nabla_x u_0+\nabla_y U_1 \text{ in } \mathrm{L}^2(\Omega,\mathrm{H}^1_{\mathrm{av}}(\mathcal{Y})),\\ &z_\varepsilon \xrightarrow{\text{s2}} Z \text{ in } \mathrm{L}^2(\Omega\times\mathcal{Y}),\end{aligned}$$

then for all $t \in [0,T]$ we have $q_\varepsilon(t) \xrightarrow{\text{s2c}} Q(t)$, where Q is the unique energetic solution associated with $\boldsymbol{E}$ and $\boldsymbol{R}$ with the initial value $Q(0) = Q^0$. In terms of evolutionary variational inequalities this means that the solutions $q_\varepsilon = (u_\varepsilon, z_\varepsilon)$ of

$$\langle \mathrm{D}\mathcal{E}_\varepsilon(t,q_\varepsilon), v-\dot q_\varepsilon\rangle + \mathcal{R}_\varepsilon(v) - \mathcal{R}_\varepsilon(\dot q_\varepsilon) \geq 0 \quad \text{ for all } v \in \mathcal{Q}$$

strongly two-scale cross-converge to the solution $Q = (u_0, U_1, Z)$ of

$$\langle \mathrm{D}\boldsymbol{E}(t,Q), V-\dot Q\rangle + \boldsymbol{R}(V) - \boldsymbol{R}(\dot Q) \geq 0 \quad \text{ for all } V \in \boldsymbol{Q}$$

if the initial conditions satisfy $q_\varepsilon(0) \xrightarrow{\text{s2c}} Q(0)$ for $\varepsilon \to 0$.

The crucial tool for proving this convergence is the abstract Γ-convergence theory developed in [MRS07]. The main difficulty in the theory is to show that weak (two-scale) limits of stable states are again stable. In [MRS07, eqn. (2.16)] a sufficient condition is provided that asks for the existence of a *joint recovery sequence* $(\widetilde q_\varepsilon)_\varepsilon$ such that

$$\limsup_{\varepsilon\to 0} \mathcal{E}_\varepsilon(t,\widetilde q_\varepsilon)+\mathcal{R}_\varepsilon(\widetilde z_\varepsilon-z_\varepsilon)-\mathcal{E}_\varepsilon(t,q_\varepsilon) \leq \boldsymbol{E}(t,\widetilde Q)+\boldsymbol{R}(\widetilde Z-Z)-\boldsymbol{E}(t,Q) \ \text{ and } \ \widetilde q_\varepsilon \overset{\text{w2c}}{\rightharpoonup} \widetilde Q,$$

where q_ε is a given family of stable states with $q_\varepsilon \overset{\text{w2c}}{\rightharpoonup} Q$ and $\widetilde Q$ is an arbitrary test state; cf. Proposition 4.5. In our situation this condition can be fulfilled by exploiting the quadratic nature of the energies, which leads to some cancellation of differences of the energies, namely, $\mathcal{E}_\varepsilon(t,\widetilde q_\varepsilon) - \mathcal{E}_\varepsilon(t,q_\varepsilon)$ converges to $\boldsymbol{E}(t,\widetilde Q) - \boldsymbol{E}(t,Q)$ if $q_\varepsilon \overset{\text{w2c}}{\rightharpoonup} Q$ and $\widetilde q_\varepsilon - q_\varepsilon \xrightarrow{\text{s2c}} \widetilde Q - Q$ strong. Here it is important that our notion of weak and strong convergences allows us to conclude convergence of scalar products; see Proposition 2.4(d).

As far as we know, this is the first homogenization work for a nonlinear and nonsmooth evolutionary problems except for [Alb03, Nes06, Nes07] and [Vis06a]. The former works treat more general quasi-static evolution laws and are not restricted to the rate-independent setting; however, they are more restrictive in the constitutive laws and prove the convergence only in an averaged sense over microscopic phase shifts of the cells. The latter work considers hyperbolic systems coupled with hysteresis operators. Similar variational inequalities are treated in [BM93, CS04, Yos01], but with different constraints and without time dependence.

2. Two-scale convergence. We recall here the definition of the two-scale convergence and several important results concerning this notion (see [Ngu89, All92, CD99, LNW02]). In particular, the presented results are based on [CDG02, Vis04], where the notions of *periodic unfolding* (also called "two-scale decomposition" in the latter work) and *periodic folding*, which is called "averaging operator" in [CDG02, sect. 5]. In the following subsections we take special care of the problems that are associated with the fact that we want to work on a bounded domain Ω and that this is only approximately compatible with microscopic periodicity. This gives rise to a certain notational complication but allows us a very precise and efficient definition of weak and strong two-scale convergences in section 2.3. Note also Example 2.7, which shows that such special care is necessary to avoid problems at the boundary.

2.1. Basic definitions of the two-scale variables. Let $d \in \mathbb{N}$ be the space dimension. The periodicity in $\mathbb{R}^d$ is expressed by a d-dimensional periodicity lattice

$$\Lambda = \left\{ \lambda = \sum_{j=1}^{d} k_j b_j \;\middle|\; k = (k_1, k_2, \dots, k_d) \in \mathbb{Z}^d \right\},$$

where $\{b_1, \dots, b_d\}$ is an arbitrary basis in $\mathbb{R}^d$. The associated unit cell is $Y = \{\, x = \sum_1^d \gamma_j b_j \mid \gamma_j \in [-1/2, 1/2) \,\} \subset \mathbb{R}^d$, such that $\mathbb{R}^d$ is the disjoint union of the translated cells $\lambda + Y$ if λ runs through all of Λ. Following [Vis04], we distinguish the unit cell from the periodicity cell $\mathcal{Y}$, which is obtained by identifying the opposite faces of $\overline{Y}$, or we may set $\mathcal{Y} = \mathbb{R}^d/_\Lambda$. Thus, $\mathcal{Y}$ has the structure of a torus. For most applications one may assume that $\Lambda = \mathbb{Z}^d$, $Y = [-1/2, 1/2)^d$, and $\mathcal{Y} = \mathbb{R}^d/_{Z^d} = \mathbb{T}^d$, the d-dimensional standard torus. However, our theory covers the general case. Yet, we will be slightly inconsistent and use y to denote elements of Y and $\mathcal{Y}$ simultaneously by relying on the natural identification between $y + \Lambda \in \mathcal{Y}$ and $y \in Y$.

On $\mathbb{R}^d$ we define the mappings $[\,\cdot\,]_\Lambda$ and $\{\cdot\}_Y$ such that

$$[\,\cdot\,]_\Lambda : \mathbb{R}^d \to \Lambda, \quad \{\cdot\}_Y : \mathbb{R}^d \to Y, \quad x = [x]_\Lambda + \{x\}_Y \text{ for all } x \in \mathbb{R}^d.$$

We also use the notation $\{\,\cdot\,\}_\mathcal{Y}$ such that $\{x\}_\mathcal{Y} = x \bmod \Lambda \in \mathcal{Y}$. Obviously a function f defined on R^d is Λ-periodic if $f(x) = f(\{x\}_Y)$ for $x \in \mathbb{R}^d$, and we may identify f with a function $\widetilde{f}$ defined on $\mathcal{Y}$. Note that $\mathrm{L}^p(Y)$ and $\mathrm{L}^p(\mathcal{Y})$ may be identified in contrast to $\mathrm{C}^k(Y)$ and $\mathrm{C}^k(\mathcal{Y}) = \mathrm{C}^k_{\mathrm{per}}(\overline{Y})$. Similarly, we use $\mathrm{H}^1(\mathcal{Y}) = \mathrm{H}^1_{\mathrm{per}}(\overline{Y})$, which is different from $\mathrm{H}^1(Y)$. A nonstandard space, which we will need in what follows, is

$$\mathrm{H}^1_{\mathrm{av}}(\mathcal{Y}) := \left\{ f \in \mathrm{H}^1(\mathcal{Y}) \;\middle|\; \int_\mathcal{Y} f(y)\,\mathrm{d}y = 0 \right\}. \tag{2.1}$$

We now introduce a small length-scale parameter $\varepsilon > 0$ and want to study functions which have fast periodic oscillations on the microscopic periodicity cell εY. We *decompose* the points $x \in \Omega \subset \mathbb{R}^d$ such that

$$x = \mathcal{N}_\varepsilon(x) + \varepsilon \mathcal{R}_\varepsilon(x) \quad \text{with } \mathcal{N}_\varepsilon(x) = \varepsilon \left[\frac{x}{\varepsilon}\right]_\Lambda \text{ and } \mathcal{R}_\varepsilon(x) = \left\{\frac{x}{\varepsilon}\right\}_Y.$$

Thus, $\mathcal{N}_\varepsilon \in \varepsilon\Lambda$ denotes the macroscopic center of the small cell $\mathcal{N}_\varepsilon(x)+\varepsilon Y$ that contains x, and $\mathcal{R}_\varepsilon$ denotes the fine-scale part of x. With this we define a *decomposition map* $\mathcal{D}_\varepsilon$ and a *composition map* $\mathcal{S}_\varepsilon$ (cf. [Vis04]) as follows:

$$\mathcal{D}_\varepsilon : \begin{cases} \mathbb{R}^d & \to & \mathbb{R}^d\times\mathcal{Y}, \\ x & \mapsto & (\mathcal{N}_\varepsilon(x), \mathcal{R}_\varepsilon(x)), \end{cases} \qquad \mathcal{S}_\varepsilon : \begin{cases} \mathbb{R}^d\times\mathcal{Y} & \to & \mathbb{R}^d, \\ (x,y) & \mapsto & \mathcal{N}_\varepsilon(x)+\varepsilon y, \end{cases}$$

where in the last sum some $y \in \mathcal{Y}$ is identified with $y \in Y \subset \mathbb{R}^d$. For the construction of a periodic unfolding operator and a folding operator in the next subsection, the following simple properties of $\mathcal{D}_\varepsilon$ and $\mathcal{S}_\varepsilon$ are essential:

$$\mathcal{D}_\varepsilon(\mathcal{S}_\varepsilon(x,y)) = (\mathcal{N}_\varepsilon(x), y) \quad \text{and} \quad \mathcal{S}_\varepsilon(\mathcal{D}_\varepsilon(x)) = x \qquad \text{for all } (x,y) \in \mathbb{R}^d\times\mathcal{Y}. \tag{2.2}$$

If Ω does not coincide with $\mathbb{R}^d$, then certain technicalities arise from the fact that the image of $\mathcal{D}_\varepsilon$ is not contained in $\Omega\times\mathcal{Y}$. Similarly, we note that $\mathcal{S}_\varepsilon(\Omega\times\mathcal{Y})$ is not contained in Ω. To handle this, we introduce, for a fixed open domain Ω, the following subsets of Λ:

$$\Lambda_\varepsilon^- = \{\, \lambda \in \Lambda \mid \varepsilon(\lambda+Y) \subset \overline{\Omega} \,\} \quad \text{and} \quad \Lambda_\varepsilon^+ = \{\, \lambda \in \Lambda \mid \varepsilon(\lambda+Y) \cap \Omega \neq \emptyset \,\}.$$

Using this, we define the domains Ω_ε^- and Ω_ε^+ via $\Omega_\varepsilon^\pm = \operatorname{int}\big(\cup_{\lambda\in\Lambda_\varepsilon^\pm} \varepsilon(\lambda+Y)\big)$. Clearly, we have $\Omega_\varepsilon^- \subset \Omega \subset \Omega_\varepsilon^+$. Moreover, we have $[\Omega_\varepsilon^\pm]_\varepsilon^\pm = \Omega_\varepsilon^\pm$, $\Omega \subset N_{\varepsilon \operatorname{diam}(Y)}(\Omega_\varepsilon^-)$, and $\Omega_\varepsilon^+ \subset N_{\varepsilon \operatorname{diam}(Y)}(\Omega)$, where $\operatorname{diam}(Y)$ is the diameter of Y and $N_\delta(A)$ denotes the δ-neighborhood of the set A.

Moreover, we set $[\Omega\times\mathcal{Y}]_\varepsilon = \mathcal{S}_\varepsilon^{-1}(\Omega) = \{\, (x,y) | \mathcal{S}_\varepsilon(x,y) \in \Omega \,\}$ and note the relations

$$\Omega_\varepsilon^- \times\mathcal{Y} \subset [\Omega\times\mathcal{Y}]_\varepsilon \subset \overline{\Omega_\varepsilon^+}\times\mathcal{Y}, \tag{2.3}$$

which will significantly be used later on. From now on we will assume that Ω satisfies

$$\Omega \text{ is open and bounded} \quad \text{and} \quad |\partial\Omega| = 0. \tag{2.4}$$

This guarantees that $|\Omega\backslash\Omega_\varepsilon^-| + |\Omega_\varepsilon^+\backslash\Omega| \to 0$ for $\varepsilon \to 0$, which will be used later. To see this, denote by ϕ_ε the characteristic function of the set $N_{\varepsilon \operatorname{diam}(Y)}(\partial\Omega)$, then $\Omega\backslash\Omega_\varepsilon^- \cup \Omega_\varepsilon^+\backslash\Omega \subset N_{\varepsilon \operatorname{diam}(Y)}(\partial\Omega)$, and for all $x \notin \partial\Omega$ we have $\phi_\varepsilon(x) \to 0$ for $\varepsilon \to 0$. Hence, we conclude $|\Omega\backslash\Omega_\varepsilon^-| + |\Omega_\varepsilon^+\backslash\Omega| \leq |N_{\varepsilon \operatorname{diam}(Y)}(\partial\Omega)| = \int_{\mathbb{R}^d} \phi_\varepsilon \,\mathrm{d}x \to 0$ for $\varepsilon \to 0$. The second condition in (2.4) is certainly satisfied if Ω has a Lipschitz boundary.

2.2. Folding and periodic unfolding operators. The notion of two-scale convergence is intrinsically linked with a suitable "two-scale embedding" of the function space $\mathrm{L}^p(\Omega)$ into the two-scale space $\mathrm{L}^p(\Omega\times\mathcal{Y})$. Such a mapping will be called a *periodic unfolding operator*. Moreover, for a two-scale function U defined on $\Omega\times\mathcal{Y}$ it is desirable to find a function u_ε defined on Ω that has the corresponding microscopic behavior. A mapping from the two-scale space into the original function space $\mathrm{L}^p(\Omega)$ will be called a *folding operator*.

The natural candidate for the periodic unfolding operator was introduced in [CDG02] and reads

$$\mathcal{T}_\varepsilon : \mathrm{L}^p(\Omega) \to \mathrm{L}^p(\mathbb{R}^d\times\mathcal{Y});\ v \mapsto v_{\mathrm{ex}} \circ \mathcal{S}_\varepsilon, \tag{2.5}$$

where $v_{\mathrm{ex}} \in \mathrm{L}^p(\mathbb{R}^d)$ is obtained from v by extending it by 0 outside of Ω. By definition, we immediately have the product rule:

$$\frac{1}{p}+\frac{1}{q} = \frac{1}{r} \leq 1,\ u \in \mathrm{L}^p(\Omega),\ v \in \mathrm{L}^q(\Omega) \implies \mathcal{T}_\varepsilon(uv) = (\mathcal{T}_\varepsilon u)\,(\mathcal{T}_\varepsilon v) \in \mathrm{L}^r(\Omega\times\mathcal{Y}). \tag{2.6}$$

In general, the support of $\mathcal{T}_\varepsilon v$ is $\overline{[\Omega\times\mathcal{Y}]_\varepsilon}$, which is not contained in $\Omega\times\mathcal{Y}$. This discrepancy in support is the main reason why we repeat the definitions of the operators and the different versions of two-scale convergence in detail. Most previous work either deals with $\Omega = \mathbb{R}^d$ or is not very precise about the supports. However, as was noted in [LNW02] (see also our Examples 2.3 and 2.7), we need to be careful here.

A variant of $\mathcal{T}_\varepsilon$ that maps continuous functions u into continuous ones can be found in [Vis04].

For a function space F, to be specified later, simple choices for folding operators are given in the form

$$\widehat{F}_\varepsilon\colon \mathrm{F}(\Omega\times\mathcal{Y}) \to \mathrm{F}(\mathbb{R}^d);\ U \mapsto U\circ\mathcal{D}_\varepsilon \text{ and } F_\varepsilon\colon \mathrm{F}(\Omega\times\mathcal{Y}) \to \mathrm{F}(\mathbb{R}^d);\ U \mapsto U\circ D_\varepsilon, \tag{2.7}$$

where D_ε is the simple decomposition $D_\varepsilon : x \mapsto (x, \{\frac{x}{\varepsilon}\}_\mathcal{Y})$. Neither of these choices is suitable if for the function space "F" we choose L^p since the image of Ω under $\mathcal{D}_\varepsilon$ and D_ε, respectively, is a set of measure 0 in $\mathbb{R}^d\times\mathcal{Y}$. However, the folding operator F_ε is well-defined as a mapping from $\mathrm{C}^k(\mathbb{R}^d\times\mathcal{Y})$ into $\mathrm{C}^k(\mathbb{R}^d)$ and has the big advantage that the image of $\Omega\times\mathcal{Y}$ under D_ε is equal to Ω. In fact, this is the basis of the classical definition of two-scale convergence; see (2.9).

The main point in this subsection is that we use a very particular folding operator $\mathcal{F}_\varepsilon$ that is well adapted to the classical L^p-spaces, namely,

$$\mathrm{L}^p(\Omega\times\mathcal{Y}) = \mathrm{L}^p(\Omega;\mathrm{L}^p(\mathcal{Y})) = \mathrm{L}^p(\mathcal{Y};\mathrm{L}^p(\Omega)) \quad \text{for } p\in[1,\infty).$$

These are the relevant ones for elliptic PDEs and our aim is to avoid spaces involving continuous functions like $\mathrm{L}^p(\Omega,\mathrm{C}(\mathcal{Y}))$ (on which $\widehat{F}_\varepsilon$ is well-defined). Our folding operator is a variant of the averaging operator $\mathcal{U}_\varepsilon$ defined in [CDG02, sect. 5], since we take special care on the domain Ω.

On $\mathrm{L}^p(\mathbb{R}^d\times\mathcal{Y})$ we first define the classical projector to piecewise constant functions on each $\varepsilon(\lambda+Y)$ via

$$(\mathcal{P}_\varepsilon U)(x,y) = \fint_{\mathcal{N}_\varepsilon(x)+\varepsilon Y} U(\xi,y)\,\mathrm{d}\xi,$$

where $\fint_A$ denotes the average over A, i.e., $\fint_A g(a)\,\mathrm{d}a = \frac{1}{|A|}\int_A g(a)\,\mathrm{d}a$. Clearly $(\mathcal{P}_\varepsilon)^2 = \mathcal{P}_\varepsilon$, $\|\mathcal{P}_\varepsilon U\|_p \le \|U\|_p$, and $\mathcal{P}_\varepsilon U \to U$ in $\mathrm{L}^p(\Omega\times\mathcal{Y})$ for all $U\in\mathrm{L}^p(\Omega\times\mathcal{Y})$.

Our folding operator $\mathcal{F}_\varepsilon$ is now defined as follows:

$$\mathcal{F}_\varepsilon : \mathrm{L}^p(\mathbb{R}^d\times\mathcal{Y}) \to \mathrm{L}^p(\Omega);\ U \mapsto \big(\mathcal{P}_\varepsilon(\chi_\varepsilon U)\circ\mathcal{D}_\varepsilon\big)|_\Omega \quad \text{with } \chi_\varepsilon = \chi_{[\Omega\times\mathcal{Y}]_\varepsilon}. \tag{2.8}$$

Note that the folding operator is defined for functions on the full space $\mathbb{R}^d\times\mathcal{Y}$ and takes values in the functions on Ω. The construction with the characteristic function $\chi_\varepsilon : \mathbb{R}^d\times\mathcal{Y} \to \{0,1\}$ guarantees that $\chi_\varepsilon = \mathcal{P}_\varepsilon\chi_\varepsilon$ and $\mathrm{sppt}(\chi_\varepsilon\circ\mathcal{D}_\varepsilon) = \overline{\Omega}$, which follows from the definition of $[\Omega\times\mathcal{Y}]_\varepsilon$ and from (2.2).

The following proposition summarizes the properties of the folding operator and the periodic unfolding operator. We restrict ourselves to the case $p\in(1,\infty)$ and leave the generalizations for $p=1$ and $p=\infty$ to the reader. In fact, in our application we will only use $p=p'=2$, which is especially nice.

PROPOSITION 2.1. *Let $p\in(1,\infty)$ and $p'=p/(p-1)$. Then the folding operator $\mathcal{F}_\varepsilon : \mathrm{L}^p(\mathbb{R}^d\times\mathcal{Y}) \to \mathrm{L}^p(\Omega)$ and the periodic unfolding operators $\mathcal{T}_\varepsilon : \mathrm{L}^p(\Omega) \to \mathrm{L}^p(\mathbb{R}^d\times\mathcal{Y})$ and $\widetilde{\mathcal{T}}_\varepsilon : \mathrm{L}^{p'}(\Omega) \to \mathrm{L}^{p'}(\mathbb{R}^d\times\mathcal{Y})$ satisfy*

(a) $\|\mathcal{T}_\varepsilon u\|_{\mathrm{L}^{p'}(\mathbb{R}^d\times\mathcal{Y})} = \|u\|_{\mathrm{L}^{p'}(\Omega)}$ *and* $\mathrm{sppt}(\mathcal{T}_\varepsilon u) \subset \overline{[\Omega\times\mathcal{Y}]_\varepsilon}$ *for all* $u\in\mathrm{L}^{p'}(\Omega)$*;*

(b) $\|\mathcal{F}_\varepsilon U\|_{\mathrm{L}^p(\Omega)} \leq \|U\|_{\mathrm{L}^p(\mathbb{R}^d\times\mathcal{Y})}$ *for all* $U \in \mathrm{L}^p(\mathbb{R}^d\times\mathcal{Y})$;

(c) $\mathcal{F}_\varepsilon$ *is the adjoint of* $\widehat{\mathcal{T}}_\varepsilon$, *i.e.*, $\mathcal{F}_\varepsilon = \big(\widehat{\mathcal{T}}_\varepsilon\big)'$;

(d) $\mathcal{F}_\varepsilon \circ \mathcal{T}_\varepsilon = \mathrm{id}_{\mathrm{L}^p(\Omega)}$ *and* $(\mathcal{T}_\varepsilon \circ \mathcal{F}_\varepsilon)^2 = \mathcal{T}_\varepsilon \circ \mathcal{F}_\varepsilon = \chi_\varepsilon \mathcal{P}_\varepsilon$.

All these identities can be obtained by elementary calculations via decomposing $\mathbb{R}^d$ into $\cup_{\lambda\in\Lambda}\varepsilon(\lambda+Y)$.

2.3. Weak and strong two-scale convergences. Following [Ngu89, All92, CD99, LNW02], a family $(u_\varepsilon)_\varepsilon$ in $\mathrm{L}^p(\Omega)$ is called *two-scale convergent* to a function $U \in \mathrm{L}^p(\Omega\times\mathcal{Y})$ and write $u_\varepsilon \xrightharpoonup{2} U$ if for all test functions $\psi : \Omega\times\mathcal{Y} \to \mathbb{R}$ we have

$$\lim_{\varepsilon\to 0} \int_\Omega u_\varepsilon(x)\psi\left(x, \left\{\frac{x}{\varepsilon}\right\}_{\mathcal{Y}}\right) \mathrm{d}x = \int_\Omega \int_{\mathcal{Y}} U(x,y)\psi(x,y)\,\mathrm{d}y\,\mathrm{d}x \quad \text{for all } \psi \in \Psi. \tag{2.9}$$

The choice of the set of test functions Ψ is important here; cf. [LNW02]. The weakest notion occurs if we take $\Psi = \mathrm{C}_\mathrm{c}^\infty(\Omega\times\mathcal{Y})$, which corresponds to a kind of distributional convergence. If $p' = p/(p-1)$ denotes the dual exponent to $p \in (1,\infty)$, the choice $\Psi = \mathrm{L}^{p'}(\Omega, \mathrm{C}(\mathcal{Y}))$ is advocated in [LNW02], since it guarantees weak convergence of $(u_\varepsilon)_\varepsilon$ to $\int_{\mathcal{Y}} U(\cdot,y)\,\mathrm{d}y$ in $\mathrm{L}^p(\Omega)$. Note that two-scale convergence can also be defined using the folding operator F_ε defined in (2.7):

$$u_\varepsilon \xrightharpoonup{2} U \quad \iff \quad \langle u_\varepsilon, F_\varepsilon\psi\rangle_\Omega = \langle u_\varepsilon, \psi\circ D_\varepsilon\rangle_\Omega \to \langle U, \psi\rangle_{\Omega\times\mathcal{Y}}.$$

Here we follow some notions from [Vis04], but modify them to fit the case $\Omega \subsetneq \mathbb{R}^d$, for defining weak and strong two-scale convergences via the periodic unfolding operators $\mathcal{T}_\varepsilon$.

Definition 2.2. *Let* $(u_\varepsilon)_{\varepsilon\in(0,\varepsilon_0)}$ *be a family in* $\mathrm{L}^p(\Omega)$ *with* $p \in (1,\infty)$.

(a) *We say that* u_ε weakly two-scale converges *to* $U \in \mathrm{L}^p(\Omega\times\mathcal{Y})$ *and write "*$u_\varepsilon \xrightharpoonup{\mathrm{w2}} U$ *in* $\mathrm{L}^p(\Omega\times\mathcal{Y})$*" if* $\mathcal{T}_\varepsilon u_\varepsilon \rightharpoonup U_{\mathrm{ex}}$ *in* $\mathrm{L}^p(\mathbb{R}^d\times\mathcal{Y})$.

(b) *We say that* u_ε strongly two-scale converges *to* $U \in \mathrm{L}^p(\Omega\times\mathcal{Y})$ *and write "*$u_\varepsilon \xrightarrow{\mathrm{s2}} U$ *in* $\mathrm{L}^p(\Omega\times\mathcal{Y})$*" if* $\mathcal{T}_\varepsilon u_\varepsilon \to U_{\mathrm{ex}}$ *(strongly) in* $\mathrm{L}^p(\mathbb{R}^d\times\mathcal{Y})$.

As the supports of $\mathcal{T}_\varepsilon u_\varepsilon$ are contained in $\overline{[\Omega\times\mathcal{Y}]_\varepsilon} \subset \overline{\Omega_\varepsilon^+}\times\mathcal{Y}$, it is clear that any possible accumulation point U of $(\mathcal{T}_\varepsilon)_\varepsilon$ has its support in $\overline{\Omega}\times\mathcal{Y}$. Because of $|\partial\Omega| = 0$ we have $\mathrm{L}^p(\Omega\times\mathcal{Y}) = \mathrm{L}^p(\overline{\Omega}\times\mathcal{Y})$, and hence accumulation points of $(\mathcal{T}_\varepsilon u_\varepsilon)_\varepsilon$ can be uniquely described by elements in $\mathrm{L}^p(\Omega\times\mathcal{Y})$. Nevertheless, it is important that our definition involves a convergence statement in $\mathrm{L}^p(\mathbb{R}^d\times\mathcal{Y})$; i.e., we need to consider functions outside of $\Omega\times\mathcal{Y}$. If the convergence was asked only for the restrictions on $\Omega\times\mathcal{Y}$, then different notions would occur.

Example 2.3. *We choose* $\Omega = (0,1)$ *and* $Y = [0,1)$. *Along the sequence* $\varepsilon_k = (k^3-1)/k^4 \to 0$ *we consider the functions*

$$u_{\varepsilon_k}(x) = a_k \text{ for } x \in (1-1/k^2, 1) \qquad \text{and } 0 \text{ otherwise},$$

which satisfy $\|u_{\varepsilon_k}\|_{\mathrm{L}^2(\Omega)} = |a_k|/k$. *The periodic unfolding* $U_k = \mathcal{T}_{\varepsilon_k}u_{\varepsilon_k} \in \mathrm{L}^2(\mathbb{R}\times\mathcal{Y})$ *reads*

$$U_k(x,y) = a_k \text{ if } \big(x \in (1-1/k^2, 1+(k-1)/k^2) \text{ and } y \in (0,1/k)\big) \qquad \text{and } 0 \text{ otherwise}.$$

The support of U_k *has only a small part in* $\Omega\times\mathcal{Y}$, *while the main part is in* $(\Omega_{\varepsilon_k}^+\setminus\Omega)\times\mathcal{Y}$. *Hence,* $U_k\big|_{\Omega\times Y}$ *has a much smaller norm, namely,* $\|U_k\big|_{\Omega\times\mathcal{Y}}\|_{\mathrm{L}^2(\Omega\times\mathcal{Y})} = |a_k|/k^{3/2}$. *Thus, for* $a_k = o(k^{3/2})$ *we have* $U_k\big|_{\Omega\times Y} \to 0$ *strongly in* $\mathrm{L}^2(\Omega\times\mathcal{Y})$, *which implies* $u_k \xrightharpoonup{2} 0$ *in* $\mathrm{L}^2(\Omega\times\mathcal{Y})$.

However, $u_{\varepsilon_k} \overset{\mathrm{w2}}{\rightharpoonup} U$ *holds if and only if* $a_k = O(k)$ *and then* $U \equiv 0$*. Moreover,* $u_{\varepsilon_k} \xrightarrow{\mathrm{s2}} U$ *if and only if* $a_k = o(k)$ *and* $U \equiv 0$ *then. (A similar example is given in* [Gri05, p. 15]*.)*

Using the fact that the folding operator is the adjoint of the periodic unfolding operator, we may equivalently define weak two-scale convergence in a way similar to the classical definition (2.9), namely,

$$\begin{aligned} (2.10) \qquad & u_\varepsilon \overset{\mathrm{w2}}{\rightharpoonup} U \text{ in } \mathrm{L}^p(\Omega\times\mathcal{Y}) \\ & \iff \quad \forall V \in \mathrm{L}^{p'}(\Omega\times\mathcal{Y}): \int_\Omega u_\varepsilon\, \mathcal{F}_\varepsilon V \,\mathrm{d}x \to \int_\Omega \int_{\mathcal{Y}} U\, V \,\mathrm{d}y\,\mathrm{d}x. \end{aligned}$$

Note that we have simply replaced the folding operator $F_\varepsilon : U \mapsto U \circ D_\varepsilon$ by the more sophisticated version $\mathcal{F}_\varepsilon$, which allows us to take general L^p functions. Moreover, the test functions V are allowed to have a support bigger than $\overline{\Omega}\times\mathcal{Y}$. As we are interested in $\varepsilon \to 0$, it suffices to consider $V \in \mathrm{L}^{p'}(N_\delta(\Omega)\times\mathcal{Y})$ for any $\delta > 0$, whereas $\delta = 0$ will lead to a strictly weaker notion of convergence.

The definitions of weak and strong two-scale convergences are obtained by transferring convergence to the classical weak and strong convergences in the classical space $\mathrm{L}^p(\Omega\times\mathcal{Y})$.

PROPOSITION 2.4. *Let* $p \in (1,\infty)$ *and* $p' = p/(p-1)$ *and assume that* Ω *satisfies* (2.4)*.*

(a) *If* $u_\varepsilon \overset{\mathrm{w2}}{\rightharpoonup} U$ *in* $\mathrm{L}^p(\Omega\times\mathcal{Y})$*, then* $\|u_\varepsilon\|_{\mathrm{L}^p(\Omega)}$ *is bounded for* $\varepsilon \to 0$*.*

(b) *If* $u_\varepsilon \overset{\mathrm{w2}}{\rightharpoonup} U$ *in* $\mathrm{L}^p(\Omega\times\mathcal{Y})$*, then* $u_\varepsilon \overset{2}{\rightharpoonup} U$*. (The reverse implication is in general not true.)*

(c) $\big(u_\varepsilon \overset{\mathrm{w2}}{\rightharpoonup} U$ *and* $\|u_\varepsilon\|_{\mathrm{L}^p(\Omega)} \to \|U\|_{\mathrm{L}^p(\Omega\times\mathcal{Y})}\big) \iff u_\varepsilon \xrightarrow{\mathrm{s2}} U$*.*

(d) *If* $u_\varepsilon \overset{\mathrm{w2}}{\rightharpoonup} U$ *in* $\mathrm{L}^p(\Omega\times\mathcal{Y})$ *and* $v_\varepsilon \xrightarrow{\mathrm{s2}} V$ *in* $\mathrm{L}^{p'}(\Omega\times\mathcal{Y})$*, then* $\langle u_\varepsilon, v_\varepsilon\rangle_\Omega \to \langle U, V\rangle_{\Omega\times\mathcal{Y}}$*.*

(e) *For each* $U \in \mathrm{L}^p(\Omega\times\mathcal{Y})$ *there exists a family* $(u_\varepsilon)_\varepsilon$ *such that* $u_\varepsilon \xrightarrow{\mathrm{s2}} U$ *in* $\mathrm{L}^p(\Omega\times\mathcal{Y})$ *(simply take* $u_\varepsilon = \mathcal{F}_\varepsilon U_{\mathrm{ex}}$*).*

(f) *For each* $w \in \mathrm{L}^p(\Omega)$ *we have* $\mathcal{T}_\varepsilon w \xrightarrow{\mathrm{s2}} Ew$ *in* $\mathrm{L}^p(\Omega\times\mathcal{Y})$*, where* $E : \mathrm{L}^p(\Omega) \to \mathrm{L}^p(\Omega\times\mathcal{Y})$ *is defined via* $Ev(x,y) = v(x)$*.*

(g) *For* $p \in (1,\infty)$*,* $q \in (1,\infty]$*, and* $\frac{1}{p} + \frac{1}{q} = \frac{1}{r} \le 1$*, let* $u_\varepsilon \overset{\mathrm{w2}}{\rightharpoonup} U$ *in* $\mathrm{L}^p(\Omega\times\mathcal{Y})$ *and* $v_\varepsilon \xrightarrow{\mathrm{s2}} V$ *in* $\mathrm{L}^q(\Omega\times\mathcal{Y})$*; then* $u_\varepsilon v_\varepsilon \overset{\mathrm{w2}}{\rightharpoonup} UV$ *in* $\mathrm{L}^r(\Omega\times\mathcal{Y})$*. If additionally* $u_\varepsilon \xrightarrow{\mathrm{s2}} U$ *in* $\mathrm{L}^p(\Omega\times\mathcal{Y})$*, then* $u_\varepsilon v_\varepsilon \xrightarrow{\mathrm{s2}} UV$ *in* $\mathrm{L}^r(\Omega\times\mathcal{Y})$*.*

Note that our definition of weak and strong convergences implies an equivalence in part (c). The corresponding result in [Gri05, Prop. 1.8] using convergence of $\mathcal{T}_\varepsilon u_\varepsilon$ in $\mathrm{L}^p(\Omega\times\mathcal{Y})$ (instead of convergence in $\mathrm{L}^p(\mathbb{R}^d\times\mathcal{Y})$) only allows for the implication "$\Rightarrow$"; see the counterexample [Gri05, p. 15].

Proof. Parts (a), (c), (d), and (g) are immediate consequences of the corresponding results of weak and strong convergences in $\mathrm{L}^p(\Omega\times\mathcal{Y})$.

Property (b) will be a consequence of Proposition 2.5 below.

Property (e) follows as the projector $\mathcal{P}_\varepsilon$ on $\mathrm{L}^p(\Omega\times\mathcal{Y})$ satisfies $\mathcal{P}_\varepsilon U \to U$ and the characteristic function χ_ε (cf. (2.8)) converges pointwise a.e. to $\chi_{\Omega\times\mathcal{Y}}$.

For property (f) we use that the unfolding operators $\mathcal{T}_\varepsilon$ have norm 1 and that for $w \in \mathrm{C}^1(\overline{\Omega})$ some calculation gives $\|\mathcal{T}_\varepsilon w - Ew\|_{\mathrm{L}^p(\Omega\times\mathcal{Y})} \le 2\,\mathrm{diam}Y\,\varepsilon\,|\Omega|^{1/p}\|\nabla w\|_{\mathrm{L}^\infty}$. Because of (2.4) the smooth functions are dense and the assertion follows. ∎

In fact, the difference between $\overset{2}{\rightharpoonup}$ and $\overset{\mathrm{w2}}{\rightharpoonup}$ disappears if we a priori impose boundedness of the sequence.

PROPOSITION 2.5. *Let $(u_\varepsilon)_\varepsilon$ be a bounded family in* $\mathrm{L}^p(\Omega)$ *with* $p \in (1,\infty)$. *Then the following statements are equivalent:*

$$\text{(i) } u_\varepsilon \overset{2}{\rightharpoonup} U \text{ in } \mathrm{L}^p(\Omega\times\mathcal{Y}), \qquad \text{(ii) } \mathcal{T}_\varepsilon u_\varepsilon\big|_{\Omega\times\mathcal{Y}} \rightharpoonup U \text{ in } \mathrm{L}^p(\Omega\times\mathcal{Y}),$$
$$\text{(iii) } u_\varepsilon \overset{\mathrm{w}2}{\rightharpoonup} U \text{ in } \mathrm{L}^p(\Omega\times\mathcal{Y}).$$

Proof. For the equivalence between (i) and (ii) see [CDG02]. The definition of $\overset{\mathrm{w}2}{\rightharpoonup}$ shows that (iii) implies (ii). Moreover, using (2.10) and the boundedness of $(u_\varepsilon)_\varepsilon$, it is sufficient to show $\int_\Omega u_\varepsilon \mathcal{F}_\varepsilon V \,\mathrm{d}x \to \int_\Omega \int_\mathcal{Y} UV \,\mathrm{d}y \,\mathrm{d}x$ on the dense subset $\Psi = \mathrm{C}^\infty_\mathrm{c}(\Omega\times\mathcal{Y})$. However, on Ψ we have $\|\mathcal{F}_\varepsilon\psi - F_\varepsilon\psi\|_{\mathrm{L}^p(\Omega)} = O(\varepsilon)$, and thus (i) implies (iii). ∎

The next result provides an improvement of part (g) in Proposition 2.4.

PROPOSITION 2.6. *Let* $p \in [1,\infty)$ *and let* $(u_\varepsilon)_\varepsilon \overset{\mathrm{s}2}{\longrightarrow} U$ *in* $\mathrm{L}^p(\Omega\times\mathcal{Y})$. *Moreover, consider a bounded sequence* $(m_\varepsilon)_\varepsilon$ *in* $\mathrm{L}^\infty(\Omega)$ *such that* $\mathcal{T}_\varepsilon m_\varepsilon(x,y) \to M(x,y)$ *for a.e.* $x \in \Omega\times\mathcal{Y}$. *Then* $m_\varepsilon u_\varepsilon \overset{\mathrm{s}2}{\longrightarrow} MU$ *in* $\mathrm{L}^p(\Omega\times\mathcal{Y})$.

Proof. By the assumption, $U_\varepsilon = \mathcal{T}_\varepsilon u_\varepsilon$ is bounded in $\mathrm{L}^p(\Omega\times\mathcal{Y})$, and hence there is a subsequence and a majorant $g \in \mathrm{L}^p(\Omega\times\mathcal{Y})$ such that $|U_{\varepsilon_k}(x,y)| \le g(x,y)$ and $U_{\varepsilon_k}(x,y) \to U(x,y)$ a.e. in $\Omega\times\mathcal{Y}$. Because of the assumptions on m_ε we find that $\mathcal{T}_{\varepsilon_k}(m_{\varepsilon_k}U_{\varepsilon_k}) = \mathcal{T}_{\varepsilon_k}m_{\varepsilon_k}\,\mathcal{T}_{\varepsilon_k}U_{\varepsilon_k}$ also has a joint majorant and converges pointwise a.e. From this we conclude $\mathcal{T}_{\varepsilon_k}m_{\varepsilon_k}U_{\varepsilon_k} \rightharpoonup MU$ in $\mathrm{L}^p(\Omega\times\mathcal{Y})$. Since the limit of all subsequences is the same, the usual contradiction argument provides the convergence of the whole family. ∎

The following example shows that the statement in Proposition 2.4(d) is not true if we do not insist on the convergence of $\mathcal{T}_\varepsilon u_\varepsilon$ and $\mathcal{T}_\varepsilon v_\varepsilon$ in $\mathrm{L}^p(\mathbb{R}^d\times\mathcal{Y})$. In [LNW02, Thm. 11] a related result to (c) is proved, namely, $\int_\Omega \tau u_\varepsilon v_\varepsilon \,\mathrm{d}x \to \int_\Omega \tau \int_\mathcal{Y} UV \,\mathrm{d}y \,\mathrm{d}x$ for all $\tau \in \mathrm{C}^\infty_\mathrm{c}(\Omega)$, where the cut-off function τ that is 0 near the boundary $\partial\Omega$ is needed to compensate for the usage of the weaker notion of two-scale convergence $\overset{2}{\rightharpoonup}$ defined in (2.9). In [LNW02, Thm. 11] strong two-scale convergence is implicitly defined by two-scale convergence $\overset{2}{\rightharpoonup}$ and additional norm convergence; see Proposition 2.4(c).

EXAMPLE 2.7. *We take* $\Omega = (0,1)$, $Y = [0,1)$, ε_k, *and* u_{ε_k} *as in Example* 2.3. *Moreover, we let* $a_k = k$ *and* $v_{\varepsilon_k} = u_{\varepsilon_k}$. *Obviously, we have* $\int u_{\varepsilon_k} v_{\varepsilon_k} \,\mathrm{d}x = \|u_{\varepsilon_k}\|^2_{\mathrm{L}^2(\Omega)} = 1$. *However, as shown above we have* $\mathcal{T}_{\varepsilon_k}u_{\varepsilon_k}\big|_{\Omega\times\mathcal{Y}} \to U_\Omega \equiv 0$ *in* $\mathrm{L}^2(\Omega\times\mathcal{Y})$. *Hence, Proposition* 2.4(d) *does not hold for the limits* U_Ω *and* V_Ω *defined in* $\mathrm{L}^p(\Omega\times\mathcal{Y})$ *only.*

2.4. Two-scale convergence of gradients. We now deal with bounded sequences in $\mathrm{W}^{1,p}(\Omega)$. The two-scale convergence for the associated gradients provides an additional structure. To formulate the result we define

$$\mathrm{W}^{1,p}_{\mathrm{av}}(\mathcal{Y}) = \left\{ w \in \mathrm{W}^{1,p}(\mathcal{Y}) \;\middle|\; \int_\mathcal{Y} w(y)\,\mathrm{d}y = 0 \right\}$$

and note that $\mathrm{L}^p(\Omega;\mathrm{W}^{1,p}_{\mathrm{av}}(\mathcal{Y}))$ is the set of functions V in $\mathrm{L}^p(\Omega\times\mathcal{Y}) = \mathrm{L}^p(\Omega;\mathrm{L}^p(\Omega))$ such that $\int_\mathcal{Y} V(x,y)\,\mathrm{d}y = 0$ for a.a. $x \in \Omega$ and that $\nabla_y V$ (in the sense of distributions) lies again in $\mathrm{L}^p(\Omega\times\mathcal{Y})$.

THEOREM 2.8. *Let $(v_\varepsilon)_\varepsilon$ be a sequence in* $\mathrm{W}^{1,p}(\Omega)$ *such that* $v_\varepsilon \rightharpoonup v_0$ *weakly in* $\mathrm{W}^{1,p}(\Omega)$, *where* $p \in (1,\infty)$. *Then* $v_\varepsilon \overset{\mathrm{s}2}{\longrightarrow} Ev_0$ *in* $\mathrm{L}^p(\Omega\times\mathcal{Y})$, *and there exist a subsequence* $(v_{\varepsilon'})_{\varepsilon'}$ *and a function* $V_1 \in \mathrm{L}^p(\Omega;\mathrm{W}^{1,p}_{\mathrm{av}}(\mathcal{Y}))$ *such that*

$$\nabla v_{\varepsilon'} \overset{\mathrm{w}2}{\rightharpoonup} E\nabla_x v_0 + \nabla_y V_1.$$

Proof. Using $v_\varepsilon \rightharpoonup v_0$ weakly in $\mathrm{W}^{1,p}(\Omega)$ gives by the compact embedding that $v_\varepsilon \to v_0$ (strongly) in $\mathrm{L}^p(\Omega)$. Employing by Propositions 2.1(a) and 2.4 we now have $\|\mathcal{T}_\varepsilon v_\varepsilon - Ev_0\|_p \le \|\mathcal{T}_\varepsilon(v_\varepsilon - v_0)\|_p + \|\mathcal{T}_\varepsilon v_0 - Ev_0\|_p \to 0$. Thus, $v_\varepsilon \xrightarrow{\mathrm{s2}} Ev_0$ is established.

The weak two-scale convergence of the gradients along a subsequence can be deduced by exploiting the corresponding result from the classical two-scale convergence; see [Ngu89, All92]. Since weak convergence in $\mathrm{W}^{1,p}(\Omega)$ implies boundedness of the gradients, the desired result follows using Proposition 2.5. ∎

As for the strong two-scale convergence for functions, we also need a density result for gradients converging in the two-scale sense. These results will be used to construct recovery sequences for the Γ-limits below. We first provide an explicit construction that is based on a smoothing procedure using the heat kernels for $\mathbb{R}^d$ and $\mathcal{Y}$. After that we provide a second construction based on ideas in [Vis04] involving the solutions of elliptic problems.

PROPOSITION 2.9. *Let $p \in (1,\infty)$ and $\Omega \subset \mathbb{R}^d$ as above. Then for every function $(u_0, U_1) \in \mathrm{W}^{1,p}(\Omega) \times \mathrm{L}^p(\Omega; \mathrm{W}^{1,p}_{\mathrm{av}}(\mathcal{Y}))$ there exists a family $(u_\varepsilon)_\varepsilon$ in $\mathrm{W}^{1,p}(\Omega)$ such that $u_\varepsilon \rightharpoonup u_0$ in $\mathrm{W}^{1,p}(\Omega)$ and that $\nabla u_\varepsilon \xrightarrow{\mathrm{s2}} E\nabla u_0 + \nabla_y U_1$.*

Proof. It is sufficient to prove the result for $u_0 \equiv 0$, since we may shift any sequence by u_0. Note that by Proposition 2.4(f) we have $\mathcal{T}_\varepsilon \nabla u_0 \xrightarrow{\mathrm{s2}} E\nabla u_0$.

Hence it suffices to find for each $V_1 \in \mathrm{L}^p(\Omega; \mathrm{W}^{1,p}_{\mathrm{av}}(\mathcal{Y}))$ a family $(v_\varepsilon)_\varepsilon$ such that

$$v_\varepsilon \rightharpoonup 0 \text{ in } \mathrm{W}^{1,p}(\Omega) \quad \text{and} \quad \nabla v_\varepsilon \xrightarrow{\mathrm{s2}} \nabla_y V_1 \text{ in } \mathrm{L}^p(\Omega \times \mathcal{Y}).$$

For this we use the heat kernels $H_{\mathbb{R}^d}$ and $H_{\mathcal{Y}}$ defined via

$$H_{\mathbb{R}^d}(t,\xi) = \frac{1}{(4\pi t)^{d/2}} \exp\big(|\xi|^2/(4t)\big) \quad \text{and} \quad H_{\mathcal{Y}}(t,\eta) = \sum_{\lambda \in \Lambda} H_{\mathbb{R}^d}(t, \eta + \lambda).$$

For $t > 0$ we now define the functions

$$V(t,x,y) = \int_{\mathbb{R}^d} \int_{\mathcal{Y}} H_{\mathbb{R}^d}(t, x-\xi) H_{\mathcal{Y}}(t, y-\eta) (V_1)_{\mathrm{ex}}(\xi,\eta) \,\mathrm{d}\eta \,\mathrm{d}\xi. \tag{2.11}$$

The classical semigroup theory for the parabolic equation $\partial_t V = \Delta_{\mathbb{R}^d} V + \Delta_{\mathcal{Y}} V$ implies $V(t,\cdot) \in \mathrm{C}^\infty(\mathbb{R}^d \times \mathcal{Y})$ for $t > 0$ and

$$\begin{aligned}
&\forall\, \alpha, \beta \in \mathbb{N}_0^d \;\exists\, C_{\alpha,\beta} > 0 \;\forall\, t > 0: \quad \|\mathrm{D}_x^\alpha \mathrm{D}_y^\beta V(t,\cdot)\|_{\mathrm{L}^p(\mathbb{R}^d \times \mathcal{Y})} \le C \,/\, t^{(|\alpha|+|\beta|)/2},\\
&\delta(t) = \|\nabla_y V(t,\cdot) - \nabla_y V_1\|_{\mathrm{L}^p(\mathbb{R}^d \times \mathcal{Y})} \to 0 \;\text{ for } t \searrow 0.
\end{aligned}$$

We define the two-scale function $v(\varepsilon,t,\cdot) \in \mathrm{W}^{1,p}(\Omega)$ via $v(\varepsilon,t,x) = \varepsilon\, V(t,x,\{\frac{x}{\varepsilon}\}_{\mathcal{Y}})$. We will choose $t = t_\varepsilon$ suitably to define $v_\varepsilon = v(\varepsilon, t_\varepsilon, \cdot)$. As a first result we obtain

$$\|v_\varepsilon\|_{\mathrm{L}^p(\Omega)} \le \varepsilon |\Omega|^{1/p} \|V(t_\varepsilon,\cdot)\|_{\mathrm{C}^0(\Omega \times \mathcal{Y})} \le \varepsilon C_{\mathrm{Sob}} \|V(t_\varepsilon,\cdot)\|_{\mathrm{W}^{k,p}(\Omega \times \mathcal{Y})} \le C \varepsilon t_\varepsilon^{-k/2},$$

where $k > (d+d)/p$ and C_{Sob} is the corresponding embedding constant for $\mathrm{W}^{k,p}(\Omega \times \mathcal{Y})$ into $\mathrm{C}^0(\Omega \times \mathcal{Y})$. Below we will choose t_ε such that $\varepsilon\, t_\varepsilon^{-k/2} \to 0$ for $\varepsilon \to 0$, and thus we conclude $v_\varepsilon \to 0$ in $\mathrm{L}^p(\Omega)$.

For the gradients we obtain $\nabla v_\varepsilon(\varepsilon,x) = \varepsilon \nabla_x V(t_\varepsilon, x, \{\frac{x}{\varepsilon}\}_{\mathcal{Y}}) + \nabla_y V(t_\varepsilon, x, \{\frac{x}{\varepsilon}\}_{\mathcal{Y}})$. Using $\|\mathcal{T}_\varepsilon \nabla v_\varepsilon - \nabla_y V_1\|_{\mathrm{L}^p(\Omega \times \mathcal{Y})} \le \|\mathcal{T}_\varepsilon v_\varepsilon - \nabla_y V(t_\varepsilon,\cdot)\|_{\mathrm{L}^p(\Omega \times \mathcal{Y})} + \delta(t_\varepsilon)$ with $\delta(t_\varepsilon) \to 0$

and recalling $\mathcal{T}_\varepsilon u(x,y) = (u\circ S_\varepsilon)(x,y) = u(\mathcal{N}_\varepsilon(x){+}\varepsilon y)$, it suffices to estimate

$$\begin{aligned}&|(\mathcal{T}_\varepsilon \nabla v_\varepsilon)(x,y) - V(t_\varepsilon,x,y)|\\ &\le \varepsilon|\nabla_x V(t_\varepsilon,\mathcal{N}_\varepsilon(x),y)| + |\nabla_y V(t_\varepsilon,\mathcal{N}_\varepsilon(x),y) - \nabla_y V(t_\varepsilon,x,y)|\\ &\le \varepsilon\|\nabla_x V(t_\varepsilon,\cdot)\|_{\mathrm{C}^0(\Omega\times\mathcal{Y})} + \varepsilon\mathrm{diam}(Y)\|\nabla_x\nabla_y V(t_\varepsilon,\cdot)\|_{\mathrm{C}^0(\Omega\times\mathcal{Y})}\\ &\le C_1\varepsilon C_{\mathrm{Sob}}\|V(t_\varepsilon,\cdot)\|_{\mathrm{W}^{k+2,p}(\Omega\times\mathcal{Y})} \le C_2\varepsilon\, t_\varepsilon^{-(k+2)/2}.\end{aligned}$$

Letting $t_\varepsilon = \varepsilon^\gamma$ with $\gamma \in (0, 2/(2{+}k))$ we obtain $\mathcal{T}_\varepsilon v_\varepsilon \rightharpoonup V_1$ in $\mathrm{L}^p(\Omega{\times}\mathcal{Y})$, and the result is proved. $\square$

The second construction is more direct and allows us to do unfolding and folding as well. It is based on [Vis04, Thm. 6.1], but we take care of the problems with the boundary $\partial\Omega$. For simplicity, we restrict ourselves to the case $p = 2$ and assume Dirichlet boundary conditions. We define the intermediate space $\mathcal{L} = \mathrm{L}^2(\Omega){\times}\mathrm{L}^2(\mathbb{R}^d{\times}\mathcal{Y})^d$, the two-scale Hilbert space $\mathcal{H} = \mathrm{H}^1_0(\Omega){\times}\mathrm{L}^2(\mathbb{R}^d, \mathrm{H}^1_{\mathrm{av}}(\mathcal{Y}))$, and the two norm-preserving linear operators

$$\mathbb{T}_\varepsilon : \begin{cases} \mathrm{H}^1_0(\Omega) & \to \quad \mathcal{L},\\ u & \mapsto \quad (u, \mathcal{T}_\varepsilon\nabla u),\end{cases} \qquad \mathbb{F}_\varepsilon : \begin{cases} \mathcal{H} & \to \quad \mathcal{L},\\ (u_0,U_1) & \mapsto \quad (u_0, (E\nabla_x u_0{+}\nabla_y U_1)_{\mathrm{ex}}).\end{cases}$$

For norm-preservation of $\mathbb{F}_\varepsilon$ we choose the norm $\|U_1\|^2_{\mathrm{H}^1_{\mathrm{av}}(\mathcal{Y})} = \|\nabla_y U_1\|_{\mathrm{L}^2(\mathcal{Y})}$.

In particular the images $\mathcal{X}^\varepsilon_{\mathbb{T}} := \mathbb{T}_\varepsilon \mathrm{H}^1(\Omega)$ and $\mathcal{X}^\varepsilon_{\mathbb{F}} = \mathbb{F}_\varepsilon\mathcal{H}$ are closed subspaces of $\mathrm{L}^2_{\mathrm{av}}(\mathcal{Y})$. We let $\mathbb{Q}^\varepsilon_{\mathbb{T}}$ and $\mathbb{Q}^\varepsilon_{\mathbb{F}}$ be the orthogonal projections onto $\mathcal{X}^\varepsilon_{\mathbb{T}}$ and $\mathcal{X}^\varepsilon_{\mathbb{F}}$, respectively. Then we are able to define a *gradient unfolding operator* $\mathcal{T}^{(1)}_\varepsilon = \mathbb{F}_\varepsilon^{-1}\mathbb{Q}^\varepsilon_{\mathbb{F}}\mathbb{T}_\varepsilon : \mathrm{H}^1_0(\Omega) \to \mathcal{H}$ and a *gradient folding operator* $\mathcal{G}_\varepsilon$ via

$$\mathcal{G}_\varepsilon : \begin{cases} \mathcal{H} & \to \quad \mathrm{H}^1_0(\Omega),\\ (u_0,U_1) & \mapsto \quad \mathbb{T}_\varepsilon^{-1}\Big(\mathbb{Q}^\varepsilon_{\mathbb{T}}\big(\mathbb{F}_\varepsilon(u_0,U_1)\big)\Big).\end{cases} \tag{2.12}$$

As the operators $\mathcal{T}^{(1)}_\varepsilon$ and $\mathcal{G}_\varepsilon$ are compositions of norm-preserving operators and orthogonal projections, they have a norm not exceeding 1. The following result shows that the definition of $\mathcal{G}_\varepsilon$ is such that it relates to solving an auxiliary elliptic problem and that it provides a recovery sequence with strongly two-scale convergent gradients.

PROPOSITION 2.10. *For given $(u_0,U_1) \in \mathcal{H}$ the function $\mathcal{G}_\varepsilon(u_0,U_1)$ is uniquely characterized as the solution $v \in \mathrm{H}^1_0(\Omega)$ of the weak elliptic problem*

$$\int_\Omega (v-u_0)\,w + \big(\nabla v - \mathcal{F}_\varepsilon(E\nabla_x u_0{+}\nabla_y U_1)\big)\cdot\nabla w\,\mathrm{d}x = 0 \quad \textit{for all } w \in \mathrm{H}^1_0(\Omega). \tag{2.13}$$

Moreover, for $\varepsilon \to 0$, we have the convergences

$$\begin{aligned} \mathcal{G}_\varepsilon(u_0,U_1) &\rightharpoonup u_0 &&\textit{in } \mathrm{H}^1_0(\Omega),\\ \nabla\mathcal{G}_\varepsilon(u_0,U_1) &\xrightarrow{\mathrm{s2}} E\nabla_x u_0 + \nabla_y U_1 &&\textit{in } \mathrm{L}^2(\Omega{\times}\mathcal{Y}).\end{aligned} \tag{2.14}$$

Proof. At first, we fix ε and let $v = \mathcal{G}_\varepsilon(u_0,U_1)$ such that $\mathbb{T}_\varepsilon v$ is the orthogonal projection of $\mathbb{F}_\varepsilon(u_0,U_1)$ onto $\mathcal{X}^\varepsilon_{\mathbb{T}} = \mathbb{T}_\varepsilon\mathrm{H}^1(\Omega)$. Denoting by $\langle\cdot,\cdot\rangle_{\mathcal{L}}$ the scalar product in $\mathcal{L}$, this means that for all $w \in \mathrm{H}^1_0(\Omega)$ we have

$$\begin{aligned}
0 &= \langle\, \mathbb{T}_\varepsilon v - \mathbb{F}_\varepsilon(u_0, U_1)\,, \mathbb{T}_\varepsilon w \rangle_{\mathcal{L}} \\
&= \int_\Omega (v-u_0)\, w \,\mathrm{d}x + \int_{\mathbb{R}^d\times\mathcal{Y}} (\mathcal{T}_\varepsilon(\nabla v) - \nabla_x u_0 - \nabla_y U_1)\cdot \mathcal{T}_\varepsilon(\nabla w)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_\Omega (v-u_0)\, w \,\mathrm{d}x + \int_\Omega (\nabla v)\cdot(\nabla w)\,\mathrm{d}x - \int_\Omega \mathcal{F}_\varepsilon(\nabla_x u_0 + \nabla_y U_1)\cdot \nabla w\,\mathrm{d}x.
\end{aligned}$$

Here we use the definitions of $\mathbb{T}_\varepsilon$ and $\mathbb{F}_\varepsilon$ as well as the properties of $\mathcal{T}_\varepsilon$ in Proposition 2.1(a) and (c). Clearly the last line gives (2.13).

To show the desired convergence we recall that the operators $\mathcal{G}_\varepsilon : \mathcal{H} \to \mathrm{H}^1(\Omega)$ have a norm bounded by 1. Hence, it suffices to prove the desired convergence on a dense subset, namely, $\mathcal{C} = \mathrm{C}^2_\mathrm{c}(\Omega)\times\mathrm{C}^2_\mathrm{c}(\Omega\times\mathcal{Y})$. For $(u_0, U_1) \in \mathcal{C}$ we write $u_\varepsilon = (\mathcal{G}_\varepsilon(u_0, U_1))$ in the form

$$u_\varepsilon(x) = v_\varepsilon(x) + g_\varepsilon(x) \quad \text{with } v_\varepsilon(x) = u_0(x) + \varepsilon\, U_1\left(x, \left\{\frac{x}{\varepsilon}\right\}_{\mathcal{Y}}\right),$$

where g_ε is the solution of the weak elliptic problem

$$\begin{aligned}
&\int_\Omega g_\varepsilon w + \nabla g_\varepsilon \cdot \nabla w \,\mathrm{d}x = \ell_\varepsilon(w) \quad \text{for all } w \in \mathrm{H}^1_0(\Omega), \\
&\text{where } \ell_\varepsilon(w) = \int_\Omega (u_0 - v_\varepsilon) w + \big(\mathcal{F}_\varepsilon(E\nabla_x u_0 + \nabla_y U_1) - \nabla v_\varepsilon\big)\cdot \nabla w\,\mathrm{d}x.
\end{aligned} \tag{2.15}$$

Clearly, the family $(v_\varepsilon)_{\varepsilon\in(0,1)}$ is bounded in $\mathrm{H}^1_0(\Omega)$. Moreover, we have $\|u_0 - v_\varepsilon\|_{\mathrm{L}^\infty} \le C_1\varepsilon$, which implies $v_\varepsilon \rightharpoonup u_0$ in $\mathrm{H}^1_0(\Omega)$. Using $\nabla v_\varepsilon(x) = \nabla u_0(x) + \nabla_y U_1(x, \{\frac{x}{\varepsilon}\}_{\mathcal{Y}}) + \varepsilon \nabla_x U_1(x, \{\frac{x}{\varepsilon}\}_{\mathcal{Y}})$ and $(u_0, U_1) \in \mathcal{C}$, we have $\|\mathcal{T}_\varepsilon \nabla v_\varepsilon - (E\nabla_x u_0 - \nabla_y U_1)_{\mathrm{ex}}\|_{\mathrm{L}^2(\mathbb{R}^d\times\mathcal{Y})} \le C_2\varepsilon$, i.e., $\nabla v_\varepsilon \xrightarrow{\mathrm{s2}} E\nabla_x u_0 - \nabla_y U_1$ in $\mathrm{L}^2(\Omega\times\mathcal{Y})$.

Hence, it suffices to show $\|g_\varepsilon\|_{\mathrm{H}^1(\Omega)} \to 0$, as this implies $\nabla g_\varepsilon \xrightarrow{\mathrm{s2}} 0$ in $\mathrm{L}^2(\Omega\times\mathcal{Y})$. From (2.15) we have

$$\begin{aligned}
\|g_\varepsilon\|^2_{\mathrm{H}^1(\Omega)} &\le \|(u_0 - v_\varepsilon, \mathcal{F}_\varepsilon(E\nabla_x u_0 + \nabla_y U_1) - \nabla v_\varepsilon)\|^2_{\mathcal{L}} \\
&= \|u_0 - v_\varepsilon\|^2_{\mathrm{L}^2(\Omega)} + \|E\nabla_x u_0 + \nabla_y U_1) - \mathcal{T}_\varepsilon \nabla\|^2_{\mathrm{L}^2(\mathbb{R}^d\times\mathcal{Y})} \;\le\; C_3\varepsilon^2.
\end{aligned}$$

This finishes the proof of the convergence result (2.14). ∎

Finally, let us note that we may extend the construction to functions $u, u_0 \in \mathrm{H}^1(\Omega)$, namely, without Dirichlet boundary conditions. In fact, for $u_0 \in \mathrm{H}^1(\Omega)$ we obtain a recovery sequence $u_\varepsilon = u_0 + \mathcal{G}_\varepsilon(0, U_1)$ by simply employing the above result and Proposition 2.4(f).

2.5. Two-scale Γ-limits. We now discuss the question of how functionals behave under two-scale convergence. This relates strongly to the question of homogenization. The two-scale convergence results we present here are well known in the literature, but often they are not easily accessible. Thus, we repeat here some simple versions which can be easily deduced by our theory and which are sufficient for our application in the next section. For more advanced results we refer to [All92, CD99, CDD06, Vis07].

Let $W : \mathcal{Y}\times\mathbb{R}^m \to \mathbb{R}_\infty := \mathbb{R}\cup\{\infty\}$ be a normal integrand, which means that for each $u \in \mathbb{R}^m$ the function $y \mapsto W(y, u)$ is measurable and that for a.e. $y \in \mathcal{Y}$ the function $u \mapsto W(y,u)$ is lower semicontinuous. Recalling our definitions of $\mathcal{T}_\varepsilon$, $\mathcal{F}_\varepsilon$, and $[\Omega\times\mathcal{Y}]_\varepsilon$ (cf. the line above (2.3)), we obtain the following central formulas:

$$\int_\Omega W\left(\left\{\frac{x}{\varepsilon}\right\}_{\mathcal{Y}}, u(x)\right)\mathrm{d}x = \int_{[\Omega\times\mathcal{Y}]_\varepsilon} W(y, \mathcal{T}_\varepsilon u\,(x,y))\,\mathrm{d}y\,\mathrm{d}x \quad \text{for all } u\in \mathrm{L}^p(\Omega). \tag{2.16}$$

This identity follows by a simple decomposition of Ω_ε^+ into small cells $\mathcal{N}_\varepsilon(\xi)+\varepsilon Y$ and using the definition of $\mathcal{T}_\varepsilon$.

The next two lemmas are the basis of the two-scale Γ-convergence for the functionals

$$\mathcal{W}_\varepsilon : \begin{cases} \mathrm{L}^p(\Omega) \ \to \ \mathbb{R}_\infty, \\ u \ \mapsto \ \int_\Omega W(\left\{\frac{x}{\varepsilon}\right\}_{\mathcal{Y}}, u(x))\,\mathrm{d}x \end{cases} \quad \text{and} \quad \boldsymbol{W} : \begin{cases} \mathrm{L}^p(\Omega\times\mathcal{Y}) \to \ \mathbb{R}_\infty, \\ U \ \mapsto \ \int_{\Omega\times\mathcal{Y}} W(y, U(x,y))\,\mathrm{d}y\,\mathrm{d}x. \end{cases}$$

LEMMA 2.11. *Assume that $p\in(1,\infty)$, that Ω is as above, and that $W : \mathcal{Y}\times\mathbb{R}^m \to \mathbb{R}_\infty$ is a convex normal integrand, i.e., $W(y,\cdot) : \mathbb{R}^m \to \mathbb{R}_\infty$ is convex for a.e. $y\in\mathcal{Y}$. Moreover, let W be bounded from below by $W(y,u) \geq -h(y)$ for a.e. $y\in\mathcal{Y}$ with $h\in\mathrm{L}^1(\Omega)$. Then*

$$u_\varepsilon \overset{\mathrm{w2}}{\rightharpoonup} U \text{ in } \mathrm{L}^p(\Omega\times\mathcal{Y}) \quad \Longrightarrow \quad \boldsymbol{W}(U) \leq \liminf_{\varepsilon\to 0} \mathcal{W}_\varepsilon(u_\varepsilon).$$

Proof. We choose an increasing sequence A_k, $k\in\mathbb{N}$, of open subsets of Ω such that $A_k \subset A_{k+1} \Subset \Omega$ and $|\Omega\backslash A_k| \to 0$ for $k\to\infty$. Then for each k there exists ε_0 such that $A_k\times\mathcal{Y} \subset \Omega_\varepsilon^-\times\mathcal{Y} \subset [\Omega\times\mathcal{Y}]_\varepsilon$ for $\varepsilon\in(0,\varepsilon_0)$.

Now consider a family with $u_\varepsilon \overset{\mathrm{w2}}{\rightharpoonup} U$. Using (2.16) and $W\geq 0$ we find

$$\mathcal{W}_\varepsilon(u_\varepsilon) = \int_{[\Omega\times\mathcal{Y}]_\varepsilon} W(y, \mathcal{T}_\varepsilon u_\varepsilon(x,y))\,\mathrm{d}y\,\mathrm{d}x \geq \int_{A_k\times\mathcal{Y}} W(y, \mathcal{T}_\varepsilon u_\varepsilon(x,y))\,\mathrm{d}y\,\mathrm{d}x - \int_{\Omega\backslash A_k} h(y)\,\mathrm{d}y.$$

On the right-hand side we may pass to the limit inferior for $\varepsilon\to 0$, as $\mathcal{T}_\varepsilon u_\varepsilon \rightharpoonup U$ in $\mathrm{L}^p(\Omega\times\mathcal{Y})$ and as W is a convex normal integrand. We obtain

$$\liminf_{\varepsilon\to 0} \mathcal{W}_\varepsilon(u_\varepsilon) \geq \int_{A_k\times\mathcal{Y}} W(y, U(x,y))\,\mathrm{d}y\,\mathrm{d}x \ - \int_{\Omega\backslash A_k} h(y)\,\mathrm{d}y.$$

Since k was arbitrary, we may consider now the limit $k\to\infty$. The second term tends to 0 as $|\Omega\backslash A_k| \to 0$, whereas the first term converges to $\boldsymbol{W}(U)$. ∎

LEMMA 2.12. *Assume that $p\in(1,\infty)$ and that Ω is as above.*

(a) *Let $W : \mathcal{Y}\times\mathbb{R}^m \to \mathbb{R}$ be a Carathéodory function, i.e., $W(y,\cdot)$ is continuous for a.e. $y\in\mathcal{Y}$ and $W(\cdot,u)$ is measurable for each $u\in\mathbb{R}^d$. Moreover, assume that there is a function $h\in\mathrm{L}^1(\mathcal{Y})$ and a constant $C>0$ such that $|W(y,u)| \leq h(y) + C(1+|u|)^p$ for all $u\in\mathbb{R}^m$ and a.e. $y\in\mathcal{Y}$. Then*

$$u_\varepsilon \xrightarrow{\mathrm{s2}} U \text{ in } \mathrm{L}^p(\Omega\times\mathcal{Y}) \quad \Longrightarrow \quad \boldsymbol{W}(U) = \lim_{\varepsilon\to 0} \mathcal{W}_\varepsilon(u_\varepsilon).$$

In particular, this implies that $\mathcal{W}_\varepsilon(\mathcal{F}_\varepsilon U_{\mathrm{ex}}) \to \boldsymbol{W}(U)$.

(b) *Let $W : \mathcal{Y}\times\mathbb{R}^m \to \mathbb{R}_\infty$ be a normal integrand such that for a.e. $y\in\mathcal{Y}$ the function $W(y,\cdot)$ is convex and that $|W(y,0)| \leq h(y)$ for some $h\in\mathrm{L}^1(\mathcal{Y})$. Then*

$$\boldsymbol{W}(U) = \lim_{\varepsilon\to 0} \mathcal{W}_\varepsilon(\mathcal{F}_\varepsilon U_{\mathrm{ex}}) \quad \text{for all } U\in\mathrm{L}^p(\Omega\times\mathcal{Y}).$$

Proof. (a) We let $U_\varepsilon = \mathcal{T}_\varepsilon u_\varepsilon$; then formula (2.16) gives

$$\mathcal{W}_\varepsilon(u_\varepsilon) = \int_{[\Omega\times\mathcal{Y}]_\varepsilon} W(y, U_\varepsilon(x,y))\,\mathrm{d}y\,\mathrm{d}x = \boldsymbol{W}(U) + I_1^\varepsilon + I_\varepsilon^2$$

with

$$I_1^\varepsilon = \int_{\Omega\times\mathcal{Y}} \big[W(y,U_\varepsilon(x,y)) - W(y,U(x,y))\big]\,\mathrm{d}y\,\mathrm{d}x = \boldsymbol{W}(U_\varepsilon) - \boldsymbol{W}(U),$$
$$I_2^\varepsilon = \int_{[\Omega\times\mathcal{Y}]_\varepsilon} W(y,U_\varepsilon(x,y))\,\mathrm{d}y\,\mathrm{d}x - \int_{\Omega\times\mathcal{Y}} W(y,U_\varepsilon(x,y))\,\mathrm{d}y\,\mathrm{d}x.$$

We have $I_1^\varepsilon \to 0$ because of $U_\varepsilon \to U_{\mathrm{ex}}$ in $\mathrm{L}^p(\mathbb{R}^d\times\mathcal{Y})$ and the strong continuity of the functional $\boldsymbol{R}$. For the later property we use the continuity of $W(y,\cdot)$ and the growth restrictions; cf. [Dac89, Val88].

For $I_2^\varepsilon \to 0$ we note that both integrals have the same integrand. Moreover, the difference of the domains $\Omega\times\mathcal{Y}$ and $[\Omega\times\mathcal{Y}]_\varepsilon$ is contained in $B_\varepsilon = \big(\overline{\Omega_\varepsilon^+\backslash\Omega_\varepsilon^-}\big)\times\mathcal{Y}$. By condition (2.4) the Lebesgue measure of this set tends to 0, whence $I_2^\varepsilon \to 0$ and we conclude

$$|I_2^\varepsilon| \le \int_{B_\varepsilon} h(y) + C(1+|U_\varepsilon(x,y)|)^p\,\mathrm{d}y\,\mathrm{d}x \to 0,$$

where again $U_\varepsilon \to U$ is used to obtain the equi-integrability of $|U_\varepsilon|^p$.

(b) We again use (2.16) for $u = \mathcal{F}_\varepsilon U_{\mathrm{ex}}$ and note that $\mathcal{T}_\varepsilon\mathcal{F}_\varepsilon U_{\mathrm{ex}} = \chi_\varepsilon \mathcal{P}_\varepsilon U_{\mathrm{ex}}$ by Proposition 2.1(d). With this we find

$$\begin{aligned}\mathcal{W}_\varepsilon(\mathcal{F}_\varepsilon U_{\mathrm{ex}}) &= \int_{[\Omega\times\mathcal{Y}]_\varepsilon} W(y,\mathcal{P}_\varepsilon U_{\mathrm{ex}}(x,y))\,\mathrm{d}y\,\mathrm{d}x = \int_{\mathbb{R}^d\times\mathcal{Y}} \chi_\varepsilon(x,y) W(y,\mathcal{P}_\varepsilon U_{\mathrm{ex}}(x,y))\,\mathrm{d}y\,\mathrm{d}x\\
&\le_{(1)} \int_{\mathbb{R}^d\times\mathcal{Y}} \chi_\varepsilon(x,y) \fint_{\mathcal{N}_\varepsilon(x)+\varepsilon Y} W(y,U_{\mathrm{ex}}(y,\xi))\,\mathrm{d}\xi\;\mathrm{d}y\,\mathrm{d}x\\
&=_{(2)} \int_{\mathbb{R}^d\times\mathcal{Y}} \chi_\varepsilon(\xi,y) W(y,U_{\mathrm{ex}}(y,\xi))\,\mathrm{d}y\,\mathrm{d}\xi \;\le_{(3)}\; \boldsymbol{W}(U) + \int_{(\Omega_\varepsilon^+\backslash\Omega)\times\mathcal{Y}} h(y)\,\mathrm{d}y\,\mathrm{d}x.\end{aligned}$$

For $\le_{(1)}$ we have used the convexity of $W(y,\cdot)$ and Jensen's inequality. The equality $=_{(2)}$ uses the fact that the integrand is piecewise constant in x on each $\mathcal{N}_\varepsilon(x)+\varepsilon Y$. For $\le_{(3)}$ we use $\chi_\varepsilon \le \chi_{\Omega\times\mathcal{Y}} + \chi_{(\Omega_\varepsilon^+\backslash\Omega)\times\mathcal{Y}}$ and $U_{\mathrm{ex}} = 0$ outside of $\Omega\times\mathcal{Y}$. Using $h \in \mathrm{L}^1(\mathcal{Y})$ and (2.4) we find $\limsup_\varepsilon \mathcal{W}_\varepsilon(\mathcal{F}_\varepsilon U_{\mathrm{ex}}) \le \boldsymbol{W}(U)$. The opposite inequality $\liminf_\varepsilon \mathcal{W}_\varepsilon(\mathcal{F}_\varepsilon U_{\mathrm{ex}}) \ge \boldsymbol{W}(U)$ was established in Lemma 2.11. ∎

The following result states that the two-scale functional $\boldsymbol{W}$ can be considered as the two-scale Γ-limit of the functionals $\mathcal{W}_\varepsilon$ in the sense of Mosco; i.e., it is the two-scale Γ-limit in the weak as well as in the strong topology.

COROLLARY 2.13. *Let $p \in (1,\infty)$ and let Ω be as above. Moreover, let $W : \mathcal{Y}\times\mathbb{R}^m \to \mathbb{R}$ be a convex, normal integrand satisfying the bounds $W(y,u) \ge -h(y)$ and $W(y,0) \le h(y)$ for all $u \in \mathbb{R}^m$ and a.a. $y \in \mathcal{Y}$ with $h \in \mathrm{L}^1(\mathcal{Y})$. Then we have the following:*

(i) *Lower estimate: $u_\varepsilon \xrightarrow{\mathrm{w2}} U$ in $\mathrm{L}^p(\Omega\times\mathcal{Y}) \implies \boldsymbol{W}(U) \le \liminf_{\varepsilon\to 0} \mathcal{W}_\varepsilon(u_\varepsilon)$.*

(ii) *Recovery sequence: $\forall U \in \mathrm{L}^p(\Omega\times\mathcal{Y})\ \exists (u_\varepsilon)_\varepsilon :\ u_\varepsilon \xrightarrow{\mathrm{s2}} U$ and $\boldsymbol{W}(U) = \lim_{\varepsilon\to 0} \mathcal{W}_\varepsilon(u_\varepsilon)$.*

REMARK 2.14. *It is possible to generalize the above results to the case that the density W also depends on the macroscopic variable $x \in \Omega$. The central identity* (2.16) *is easily generalized to*

$$\int_\Omega W_\varepsilon(x, u(x))\,\mathrm{d}x = \int_{[\Omega\times\mathcal{Y}]_\varepsilon} W_\varepsilon(\mathcal{S}_\varepsilon(x,y), \mathcal{T}_\varepsilon u(x,y))\,\mathrm{d}y\,\mathrm{d}x \quad \textit{for all } u \in \mathrm{L}^p(\Omega).$$

Thus, if we want to realize a general Carathéodory function $W : \Omega\times\mathcal{Y}\times\mathbb{R}^m \to \mathbb{R}_\infty$ in the two-scale limit functional $\boldsymbol{W}$, we define $\mathcal{W}_\varepsilon$ via the approximate energy density,

$$W_\varepsilon(x,u) = \widehat{W}_\varepsilon\left(x, \left\{\frac{x}{\varepsilon}\right\}_{\mathcal{Y}}, u\right) \quad \textit{with } \widehat{W}_\varepsilon(x,y,u) = \fint_{\mathcal{N}_\varepsilon(x)+\varepsilon Y} W(\xi, y, u)\,\mathrm{d}\xi,$$

instead of the traditionally used $W(x, \left\{\frac{x}{\varepsilon}\right\}_{\mathcal{Y}}, u)$. Note that W_ε satisfies $W_\varepsilon(\mathcal{S}_\varepsilon(x,y),u) = \widehat{W}_\varepsilon(x,y,u) \to W(x,y,u)$ a.e. for $\varepsilon \to 0$.

Under some mild additional conditions it is then possible to pass to the limit as in Lemmas 2.11 *and* 2.12*; see also Proposition* 2.6*. This also resolves the difficulties addressed in* [CDG02, Thm. 2]*. This will be the subject of future research.*

2.6. Two-scale cross-convergence. Finally we present a result concerning a functional involving gradients. For families $((u_\varepsilon, z_\varepsilon))_\varepsilon$ in $\mathrm{W}^{1,p}(\Omega)\times\mathrm{L}^p(\Omega)$ we define the notions of *weak* and *strong two-scale cross-convergences* as follows:

$$(u_\varepsilon, z_\varepsilon) \overset{\mathrm{w2c}}{\rightharpoonup} (u_0, U_1, Z) \text{ in } \boldsymbol{X}_p \quad \Longleftrightarrow \quad \begin{cases} u_\varepsilon \rightharpoonup u_0 & \text{in } \mathrm{W}^{1,p}(\Omega), \\ \nabla u_\varepsilon \overset{\mathrm{w2}}{\rightharpoonup} E\nabla u_0 + \nabla_y U_1 & \text{in } \mathrm{L}^p(\Omega\times\mathcal{Y}), \\ z_\varepsilon \overset{\mathrm{w2}}{\rightharpoonup} Z & \text{in } \mathrm{L}^p(\Omega\times\mathcal{Y}), \end{cases}$$

$$(u_\varepsilon, z_\varepsilon) \xrightarrow{\mathrm{s2c}} (u_0, U_1, Z) \text{ in } \boldsymbol{X}_p \quad \Longleftrightarrow \quad \begin{cases} u_\varepsilon \rightharpoonup u_0 & \text{in } \mathrm{W}^{1,p}(\Omega), \\ \nabla u_\varepsilon \xrightarrow{\mathrm{s2}} E\nabla u_0 + \nabla_y U_1 & \text{in } \mathrm{L}^p(\Omega\times\mathcal{Y}), \\ z_\varepsilon \xrightarrow{\mathrm{s2}} Z & \text{in } \mathrm{L}^p(\Omega\times\mathcal{Y}), \end{cases}$$

where $\boldsymbol{X}_p = \mathrm{W}^{1,p}(\Omega)\times\mathrm{L}^p(\Omega;\mathrm{W}^{1,p}_{\mathrm{av}}(\mathcal{Y}))\times\mathrm{L}^p(\Omega\times\mathcal{Y})$. The final result on two-scale Γ-convergence now provides relations between the functionals

$$\Phi_\varepsilon(u,z) = \int_\Omega \phi\left(\left\{\frac{x}{\varepsilon}\right\}_{\mathcal{Y}}, u(x), \nabla u(x), z(x)\right)\mathrm{d}x,$$

$$\boldsymbol{\Phi}_\varepsilon(u_0, U_1, Z) = \int_{\Omega\times\mathcal{Y}} \phi(y, u_0(x), \nabla u_0(x){+}\nabla_y U_1(x,y), Z(x,y))\,\mathrm{d}x.$$

PROPOSITION 2.15. *Let $p \in (1,\infty)$ and let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary. Assume that $\phi : \mathcal{Y}\times\mathbb{R}^k\times\mathbb{R}^{k\times d}\times\mathbb{R}^m \to \mathbb{R}$ is a Carathéodory function (measurable in $y \in \mathcal{Y}$ and continuous in $(u,F,z) \in \mathbb{R}^k\times\mathbb{R}^{k\times d}\times\mathbb{R}^m \to \mathbb{R}$) satisfying the bound $|\phi(y,u,A,z)| \le h(y) + C(1{+}|u|{+}|A|{+}|z|)^p$ for $h \in \mathrm{L}^1(\mathcal{Y})$. Then we have*

$$(u_\varepsilon, z_\varepsilon) \xrightarrow{\mathrm{s2c}} (u_0, U_1, Z) \textit{ in } \boldsymbol{X}_p \quad \Longrightarrow \quad \Phi_\varepsilon(u_\varepsilon, z_\varepsilon) \to \boldsymbol{\Phi}(u_0, U_1, Z).$$

Moreover, if $\phi(y,\cdot)$ is convex for a.a. $y \in \mathcal{Y}$, we also have

$$(u_\varepsilon, z_\varepsilon) \overset{\mathrm{w2c}}{\rightharpoonup} (u_0, U_1, Z) \textit{ in } \boldsymbol{X}_p \quad \Longrightarrow \quad \boldsymbol{\Phi}(u_0, U_1, Z) \le \liminf_{\varepsilon\to 0} \Phi_\varepsilon(u_\varepsilon, z_\varepsilon).$$

The proof is a direct consequence of combining Lemmas 2.11 and 2.12(a).

3. Existence and uniqueness of solution.

3.1. Abstract result. For the convenience of the reader we recall the standard existence and uniqueness results for evolutionary variational inequalities; see, e.g., [BS96, Vis94, Mie05]. We start with a Hilbert space $\mathcal{Q}$ with dual $\mathcal{Q}^*$ and dual pairing $\langle\cdot,\cdot\rangle : \mathcal{Q}^*\times\mathcal{Q} \to \mathbb{R}$ and a positive semidefinite operator $\mathcal{A} \in \mathrm{Lin}(\mathcal{Q},\mathcal{Q}^*)$, i.e., $\mathcal{A} = \mathcal{A}^*$ and $\langle \mathcal{A}q, q\rangle \geq 0$ for all $q \in \mathcal{Q}$. For a function $\ell \in \mathrm{C}^1([0,T],\mathcal{Q}^*)$ we define the energy functional

$$\mathcal{E}(t,q) = \frac{1}{2}\langle \mathcal{A}q, q\rangle - \langle \ell(t), q\rangle.$$

Moreover, let a dissipation functional $\mathcal{R} : \mathcal{Q} \to [0,\infty]$ be given that is convex, lower semicontinuous, and positively homogeneous of degree 1, viz.,

$$\mathcal{R}(\gamma q) = \gamma\mathcal{R}(q) \quad \text{ for all } \gamma \geq 0 \text{ and } q \in \mathcal{Q}.$$

The *energetic formulation* (S) and (E) of the rate-independent hysteresis problem associated with $\mathcal{E}$ and $\mathcal{R}$ is based on the *global stability condition* (S) and the *energy balance* (E):

$$\text{(S)} \qquad \mathcal{E}(t,q(t)) \leq \mathcal{E}(t,\widetilde{q}) + \mathcal{R}(\widetilde{q}-q(t)) \quad \text{for every } \widetilde{q} \in \mathcal{Q},$$

$$\text{(E)} \qquad \mathcal{E}(t,q(t)) + \mathrm{Diss}_{\mathcal{R}}(q;[0,t]) = \mathcal{E}(0,q(0)) + \int_0^t \partial_s\mathcal{E}(s,q(s))\,\mathrm{d}s,$$

where $\mathrm{Diss}_{\mathcal{R}}(q;[r,s]) = \int_r^s \mathcal{R}(\dot{q}(t))\,\mathrm{d}t$ and $\partial_s\mathcal{E}(s,q(s)) = -\langle\dot{\ell}(s), q(s)\rangle$. We call $q : [0,T] \to \mathcal{Q}$ satisfying (S) and (E) for all $t \in [0,T]$ an *energetic solution* associated with $(\mathcal{E},\mathcal{R})$.

The stability condition can be formulated in terms of the *sets of stable states*

$$\mathcal{S}(t) = \{\, q \in \mathcal{Q} \mid \mathcal{E}(t,q) \leq \mathcal{E}(t,\widehat{q}) + \mathcal{R}(\widehat{q}-q) \text{ for every } \widehat{q} \in \mathcal{Q}\,\}.$$

Now, (S) just means $q(t) \in \mathcal{S}(t)$.

There are several equivalent formulation for (S) and (E), for instance, the subdifferential inclusion $0 \in \partial\mathcal{R}(\dot{q}(t)) + \mathrm{D}_q\mathcal{E}(t,q(t))$ or the variational inequality

$$\text{(3.1)} \qquad \langle \mathcal{A}q(t)-\ell(t), v-\dot{q}(t)\rangle + \mathcal{R}(v) - \mathcal{R}(\dot{q}(t)) \geq 0 \quad \text{for every } v \in \mathcal{Q}.$$

For these equivalences, we refer to [MT04, Mie05], where also a proof of the following existence and uniqueness result can be found.

THEOREM 3.1. *Let $\ell \in \mathrm{C}^1([0,T],\mathcal{Q}^*)$ and $q_0 \in \mathcal{S}(0)$. Moreover, assume that the following coercivity condition holds:*

$$\text{(3.2)} \qquad \exists\, \alpha > 0\ \forall\, v \in \mathcal{Q} \text{ with } \mathcal{R}(v) < \infty : \quad \langle Av, v\rangle \geq \alpha\|v\|^2.$$

Then the energetic problem (S) *and* (E) *has a unique solution $q \in \mathrm{C}^{\mathrm{Lip}}([0,T],\mathcal{Q})$ with*

$$\|q(t) - q(s)\|_{\mathcal{Q}} \leq \frac{\mathrm{Lip}_{\mathcal{Q}^*}(\ell)}{\alpha}|t-s| \quad \textit{for all } s,t \in [0,T].$$

For the reader's convenience we repeat the main argument for the a priori estimate. Assume that for t the derivative $\dot{q}(t)$ exists. Using (3.1) with $v = 0$ we find $\langle \mathcal{A}q(t) - \ell(t), -\dot{q}(t)\rangle - \mathcal{R}(\dot{q}(t)) \leq 0$. For a sequence $t_n \to t$ where (3.1) holds we test with $v = \mu\dot{q}(t)$, divide by μ, and consider the limit $\mu \to \infty$. Using 1-homogeneity of

$\mathcal{R}$ we obtain $\langle \mathcal{A}q(t_n) - \ell(t_n), \dot{q}(t)\rangle + \mathcal{R}(\dot{q}(t)) \leq 0$. Adding this to the above estimate gives

$$\langle \big(\mathcal{A}q(t_n) - \ell(t_n)\big) - \big(\mathcal{A}q(t) - \ell(t)\big), \dot{q}(t)\rangle \leq 0.$$

Assuming $t_n > t$ we may divide the above inequality and pass to the limit to find $\langle \mathcal{A}\dot{q}(t) - \dot{\ell}(t), \dot{q}(t)\rangle \leq 0$. For $t_n < t$ we find the opposite inequality. Since we may approach t by sequences from both sides, this implies $\langle \mathcal{A}\dot{q}(t), \dot{q}(t)\rangle = \langle \dot{\ell}(t), \dot{q}(t)\rangle$. Now, (3.2) leads to the desired result $\alpha\|\dot{q}(t)\| \leq \|\dot{\ell}(t)\|_*$.

3.2. Elastoplasticity with periodic coefficients. In this section we formulate the continuum mechanics that describes the rate-independent evolution of an elastoplastic body under prescribed loading. This model is the classical one introduced by Moreau and is still used in many engineering applications; cf. [Mor76, HR99].

The body occupies a domain $\Omega \subset \mathbb{R}^d$, which is assumed to be a nonempty connected bounded open set with Lipschitz boundary $\partial\Omega$. As above we have a length scale parameter ε and a periodicity lattice Λ with unit cell $Y \subset \mathbb{R}^d$. With $u : \Omega \to \mathbb{R}^d$ we denote the displacement of the body, and $z : \Omega \to \mathbb{R}^m$ denotes a vector of internal variables which will account for inelastic effects due to plastic strains and plastic hardening.

The material properties are assumed to be periodic with respect to the microscopic lattice $\varepsilon\Lambda$, which leads to the dependence on $\left\{\frac{x}{\varepsilon}\right\}_{\mathcal{Y}}$. The energy functional $\mathcal{E}_\varepsilon$ is based on a stored-energy density $W : \mathcal{Y}\times\mathbb{R}^{d\times d}_{\mathrm{sym}}\times\mathbb{R}^m \to \mathbb{R}; (y, \boldsymbol{e}, z) \mapsto W(y, \boldsymbol{e}, z)$, where $\mathbb{R}^{d\times d}_{\mathrm{sym}} = \{\, A \in \mathbb{R}^{d\times d} \mid A = A^{\mathsf{T}} \,\}$ and $\boldsymbol{e} = \boldsymbol{e}(u) = \frac{1}{2}(\nabla u + \nabla u^{\mathsf{T}}) \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$ is the linearized strain tensor. With this, $\mathcal{E}_\varepsilon$ takes the form

$$\mathcal{E}_\varepsilon(t, u, z) = \int_\Omega W\left(\left\{\frac{x}{\varepsilon}\right\}_{\mathcal{Y}}, \boldsymbol{e}(u)(x), z(x)\right) \mathrm{d}x - \langle \ell(t), u\rangle$$

$$\text{with } \langle \ell(t), u\rangle = \int_\Omega u(x)\cdot f_{\mathrm{ap}}(t,x)\,\mathrm{d}x + \int_{\partial\Omega} u(\xi)\cdot g_{\mathrm{ap}}(t,\xi)\,\mathrm{d}\xi,$$

where f_{ap} and g_{ap} are the applied, time-dependent loading in the volume and on the surface, respectively. We assume that they satisfy $f_{\mathrm{ap}} \in \mathrm{C}^1([0,T], \mathrm{L}^2(\Omega;\mathbb{R}^d))$ and $g_{\mathrm{ap}} \in \mathrm{C}^1([0,T], \mathrm{L}^2(\partial\Omega;\mathbb{R}^d))$, such that $\ell \in \mathrm{C}^1([0,T], \mathrm{H}^1(\Omega;\mathbb{R}^d)^*)$.

For the stored energy W we assume that it is a quadratic form in $(\boldsymbol{e}, z)$, namely,

$$W(y, \boldsymbol{e}, z) = \frac{1}{2}\left\langle\!\!\left\langle \mathbb{A}(y)\begin{pmatrix}\boldsymbol{e}\\ z\end{pmatrix}, \begin{pmatrix}\boldsymbol{e}\\ z\end{pmatrix}\right\rangle\!\!\right\rangle,$$

where $\mathbb{A}(y) : \mathbb{R}^{d\times d}_{\mathrm{sym}}\times\mathbb{R}^m \to \mathbb{R}^{d\times d}_{\mathrm{sym}}\times\mathbb{R}^m$ is a positive semidefinite linear operator and $\langle\!\langle \binom{\boldsymbol{e}}{z}, \binom{\widetilde{\boldsymbol{e}}}{\widetilde{z}} \rangle\!\rangle = \sum_{i,j=1}^d \boldsymbol{e}_{ij}\widetilde{\boldsymbol{e}}_{ij} + \sum_{k=1}^m z_k\widetilde{z}_k$ is the scalar product on $\mathbb{R}^{d\times d}_{\mathrm{sym}}\times\mathbb{R}^m$.

The dissipation potential $\mathcal{R}_\varepsilon$ is defined via a dissipation density $\rho : \mathcal{Y}\times\mathbb{R}^m \to [0,\infty]$, i.e., $\mathcal{R}_\varepsilon(\dot{z}) = \int_\Omega \rho(\left(\left\{\frac{x}{\varepsilon}\right\}_{\mathcal{Y}}\right), \dot{z}(x))\,\mathrm{d}x$. Rate-independence is imposed by assuming that $\rho(y,\cdot)$ is positively homogeneous of degree 1 (for short, 1-homogeneous). Note that ρ is not assumed to be symmetric (i.e., $\rho(y,-\dot{z}) \neq \rho(y,\dot{z})$ is allowed), since this freedom is necessary to model hardening.

Our precise assumptions on the material data $\mathbb{A}$ and ρ are

(3.3a) $\mathbb{A} \in \mathrm{L}^\infty(\mathcal{Y}, \mathrm{Lin}(\mathbb{R}^{d\times d}_{\mathrm{sym}}\times\mathbb{R}^m))$ with $\mathbb{A}(y) = \mathbb{A}(y)^\mathsf{T} \geq 0$,

(3.3b) $\rho : \mathcal{Y} \to [0,\infty]$ is a convex, normal integrand and $\rho(y,\cdot)$ is 1-homogeneous,

(3.3c)
$$\exists\, \widehat{\alpha} > 0\ \forall_{a.a.} y \in \mathcal{Y}\ \forall \begin{pmatrix} \boldsymbol{e} \\ z \end{pmatrix} \in \mathbb{R}^{d\times d}_{\mathrm{sym}}\times\mathbb{R}^m \text{ with } \rho(y,z) < \infty\ :$$
$$\left\langle\!\!\left\langle \mathbb{A}(y)\begin{pmatrix} \boldsymbol{e} \\ z \end{pmatrix}, \begin{pmatrix} \boldsymbol{e} \\ z \end{pmatrix} \right\rangle\!\!\right\rangle \geq \widehat{\alpha} \left| \begin{pmatrix} \boldsymbol{e} \\ z \end{pmatrix} \right|^2.$$

Remark 3.2. *Here we describe the exact setting for the linearized theory of elastoplasticity which is the motivation of this work. However, in the remainder of the paper we do not rely on the further specifications given here.*

The basis of linearized elastoplasticity is the additive split of the strain into an elastic part $\boldsymbol{e}_{\mathrm{el}} = \boldsymbol{e}(u) - \boldsymbol{p}$ *and a plastic part* $\boldsymbol{p} = \mathbb{B}(y)z$, *where* $\mathbb{B}(y) : \mathbb{R}^m \to \mathbb{R}^{d\times d}_{\mathrm{sym}}$ *is a linear mapping. Then* W *is taken in the form*

$$W(y,\boldsymbol{e},z) = \big\langle \mathbb{C}(y)(\boldsymbol{e}-\mathbb{B}(y)z), \boldsymbol{e}-\mathbb{B}(y)z \big\rangle_{d\times d} + \langle \mathbb{H}(y)z, z\rangle_m,$$

where $\mathbb{C}(y) : \mathbb{R}^{d\times d}_{\mathrm{sym}} \to \mathbb{R}^{d\times d}_{\mathrm{sym}}$ *is the symmetric (fourth order) elasticity tensor and* $\mathbb{H}(y)$ *denotes the hardening tensor. This means* $\mathbb{A}$ *has the block structure* $\big(\begin{smallmatrix} \mathbb{C} & \mathbb{C}\mathbb{B} \\ -\mathbb{B}^*\mathbb{C} & \mathbb{H}+\mathbb{B}^*\mathbb{C}\mathbb{B} \end{smallmatrix}\big)$.

The typical case of isotropic hardening may be written in the way that $z = (\boldsymbol{p}, h)$, *where* $\boldsymbol{p} \in (\mathbb{R}^{d\times d}_{\mathrm{sym}})_0 = \{\, A \in \mathbb{R}^{d\times d}_{\mathrm{sym}} \mid \operatorname{tr} A = 0 \,\}$ *is the (deviatoric) plastic strain (i.e.,* $\mathbb{B}(y)(\boldsymbol{p},h) = \boldsymbol{p}$*),* $h \in \mathbb{R}$ *is the isotropic hardening parameter, and* $\mathbb{H}(y)$ *is taken as* $\kappa(y) > 0$. *Moreover,* ρ *is assumed to have the form*

$$\rho(y,(\dot{\boldsymbol{p}},\dot{h})) = \begin{cases} r(y)\dot{h} & \text{for } \dot{h} \geq 0 \text{ and } \dot{\boldsymbol{p}} \in \dot{h}\,\Sigma(y), \\ \infty & \text{otherwise}, \end{cases}$$

where $r(y) > 0$ *and* $\Sigma(y) \subset (\mathbb{R}^{d\times d}_{\mathrm{sym}})^*_0$ *is the compact and convex elastic domain (with* $\partial\Sigma(y)$ *being the yield surface) at the point* $y \in \mathcal{Y}$ *for the the initial hardening state* $h = 1$.

The coercivity assumption (3.3c)*, which is essential for our analysis, then follows if we assume that there exist positive constants* c *and* C *such that for a.a.* $y \in \mathcal{Y}$ *we have the estimates*

$$\kappa(y) \geq c, \quad \langle \mathbb{C}(y)\boldsymbol{e},\boldsymbol{e}\rangle \geq c|\boldsymbol{e}|^2 \quad \text{for all } \boldsymbol{e}, \qquad |\boldsymbol{\sigma}| \leq C \quad \text{for all } \boldsymbol{\sigma} \in \Sigma(y).$$

Note that the restriction $\rho(y,(\boldsymbol{p},h)) < \infty$ *implies* $|\boldsymbol{p}| \leq Ch$. *Without the coercivity postulated in* (3.3c) *the analysis would be significantly more difficult due to strain localization; see, e.g.,* [DDM06].

Finally, we fix the function spaces by prescribing Dirichlet boundary conditions $u = 0$ along the part Γ_{Dir} of $\partial\Gamma$. This defines the underlying Hilbert space

$$\mathcal{Q} = \mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d\times\mathrm{L}^2(\Omega)^m \quad \text{with } \mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega) = \{\, u \in \mathrm{H}^1(\Omega) \mid u_{\Gamma_{\mathrm{Dir}}} = 0 \,\}.$$

The domain Ω and the Dirichlet boundary part Γ_{Dir} are specified further in the next result to guarantee coercivity of the energy $\mathcal{E}_\varepsilon$.

Proposition 3.3 (Korn's inequality). *Let* $\Omega \subset \mathbb{R}^d$ *be a connected, open, bounded set with Lipschitz boundary* Γ. *Moreover, let* Γ_{Dir} *be a measurable subset of* Γ, *such that* $\int_{\Gamma_{\mathrm{Dir}}} 1\,\mathrm{d}a > 0$. *Then there exists a constant* $C_{\mathrm{Korn}} > 0$, *such that*

$$\int_\Omega |\boldsymbol{e}(u)|^2\,\mathrm{d}x \geq C_{\mathrm{Korn}}\|u\|^2_{\mathrm{H}^1(\Omega)} \quad \textit{for all } u \in \mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d. \tag{3.4}$$

Clearly, we may write $\mathcal{E}_\varepsilon(t,\boldsymbol{e},z) = \frac{1}{2}\langle\!\langle \mathcal{A}_\varepsilon\binom{u}{z},\binom{u}{z}\rangle\!\rangle - \langle\widetilde{\ell}(t),\binom{\boldsymbol{e}}{z}\rangle$, where $\mathcal{A}_\varepsilon : \mathcal{Q}\to\mathcal{Q}^*$ is symmetric and positive semidefinite. Moreover, combining assumption (3.3c) and Korn's inequality, we find for all $\binom{\boldsymbol{e}}{z}\in\mathcal{Q}$ with $\mathcal{R}_\varepsilon(z)<\infty$ the coercivity estimate

(3.5)
$$\left\langle\!\!\left\langle \mathcal{A}_\varepsilon\begin{pmatrix}u\\ z\end{pmatrix},\begin{pmatrix}u\\ z\end{pmatrix}\right\rangle\!\!\right\rangle \ge \widehat{\alpha}\left\|\begin{pmatrix}\boldsymbol{e}(u)\\ z\end{pmatrix}\right\|^2_{\mathrm{L}^2(\Omega)} \ge \alpha\left\|\begin{pmatrix}u\\ z\end{pmatrix}\right\|^2_{\mathcal{Q}} \quad \text{with } \alpha=\widehat{\alpha}\min\{1,C_{\mathrm{Korn}}\}.$$

We call $q_\varepsilon=(u_\varepsilon,z_\varepsilon):[0,T]\to\mathcal{Q}$ an *energetic solution* associated with $(\mathcal{E}_\varepsilon,\mathcal{R}_\varepsilon)$ if for all $t\in[0,T]$ the *stability condition* (S^ε) and the *energy balance* (E^ε) hold:

(3.6)
$$(\mathrm{S}^\varepsilon)\quad \mathcal{E}_\varepsilon(t,u_\varepsilon(t),z_\varepsilon(t)) \le \mathcal{E}_\varepsilon(t,\widetilde{u},\widetilde{z}) + \mathcal{R}_\varepsilon(\widetilde{z}-z_\varepsilon(t)) \quad \text{for every } (\widetilde{u},\widetilde{z})\in\mathcal{Q},$$
$$(\mathrm{E}^\varepsilon)\quad \mathcal{E}_\varepsilon(t,u_\varepsilon(t),z_\varepsilon(t)) + \int_0^t \mathcal{R}_\varepsilon(\dot{z}_\varepsilon(s))\,\mathrm{d}s = \mathcal{E}_\varepsilon(0,u_\varepsilon(0),z_\varepsilon(0)) - \int_0^t \langle\dot{\ell}(s),u(s)\rangle\,\mathrm{d}s.$$

Applying the abstract Theorem 3.1 we immediately obtain the following existence and uniqueness result, which contains an a priori Lipschitz bound that is independent of $\varepsilon>0$.

PROPOSITION 3.4. *Let $\ell\in\mathrm{C}^{\mathrm{Lip}}([0,T],(\mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d)^*)$. Then for all $\varepsilon>0$ and all stable $(u^0_\varepsilon,z^0_\varepsilon)\in\mathcal{Q}$ there exists a unique solution $(u_\varepsilon,z_\varepsilon)\in\mathrm{C}^{\mathrm{Lip}}([0,T],\mathcal{Q})$ of* (S^ε) *and* (E^ε) *with $(u_\varepsilon(0),z_\varepsilon(0))=(u^0_\varepsilon,z^0_\varepsilon)$. Moreover, all these solutions satisfy*

$$(3.7)\quad \|(u_\varepsilon(t),z_\varepsilon(t))-(u_\varepsilon(s),z_\varepsilon(s))\|_{\mathcal{Q}} \le \frac{\mathrm{Lip}_{\mathcal{Q}^*}((\ell,0))}{\alpha}\,|t-s| \quad \textit{for all } t,s\in[0,T],$$

where α is defined in (3.5) *and is independent of ε.*

3.3. The two-scale homogenized problem. Instead of the functionals $\mathcal{E}_\varepsilon$ and $\mathcal{R}_\varepsilon$ we may consider their two-scale limits. As the energy storage functional depends on the gradient of u, we use the notion of two-scale cross-convergence introduced in section 2.6 on the space

$$\boldsymbol{Q}=\boldsymbol{H}\times\boldsymbol{Z} \quad \text{with } \boldsymbol{H}=\mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d\times\mathrm{L}^2(\Omega,\mathrm{H}^1_{\mathrm{av}}(\mathcal{Y}))^d \text{ and } \boldsymbol{Z}=\mathrm{L}^2(\Omega\times\mathcal{Y})^m.$$

We use $U=(u_0,U_1)$ for the elements in $\boldsymbol{H}$ and Z for the internal elements lying in $\boldsymbol{Z}$. The functionals $\boldsymbol{E}$ and $\boldsymbol{R}$ are defined via

$$\boldsymbol{E}(t,U,Z) = \int_{\Omega\times\mathcal{Y}} \frac{1}{2}\left\langle\!\!\left\langle \mathbb{A}(y)\begin{pmatrix}\widehat{\boldsymbol{e}}(U)\\ Z\end{pmatrix},\begin{pmatrix}\widehat{\boldsymbol{e}}(U)\\ Z\end{pmatrix}\right\rangle\!\!\right\rangle - \langle\ell(t),u_0\rangle,$$
$$\boldsymbol{R}(Z) = \int_{\Omega\times\mathcal{Y}} \rho(y,Z(x,y))\,\mathrm{d}y\,\mathrm{d}x,$$

where

$$\widehat{\boldsymbol{e}}(U) = \boldsymbol{e}_x(u_0)+\boldsymbol{e}_y(U_1) = \frac{1}{2}(\nabla_x u_0+(\nabla_x u_0)^{\mathsf{T}}) + \frac{1}{2}(\nabla_y U_1+(\nabla_y U_1)^{\mathsf{T}}).$$

Again we define the energetic formulation for $\boldsymbol{E}$ and $\boldsymbol{R}$ on $\boldsymbol{Q}$ via the global stability condition (**S**) and the energy balance (**E**). As above, a mapping $(U,Z):[0,T]\to\boldsymbol{H}\times\boldsymbol{Z}=\boldsymbol{Q}$ is called an energetic solution associated with $\boldsymbol{E}$ and $\boldsymbol{R}$ if for all $t\in[0,T]$ we have

(3.8)
$$(\mathbf{S})\quad \boldsymbol{E}(t,U(t),Z(t)) \le \boldsymbol{E}(t,\widetilde{U},\widetilde{Z}) + \boldsymbol{R}(\widetilde{Z}-Z(t)) \quad \text{for all } (\widetilde{U},\widetilde{Z})\in\boldsymbol{H}\times\boldsymbol{Z},$$
$$(\mathbf{E})\quad \boldsymbol{E}(t,U(t),Z(t)) + \int_0^t \boldsymbol{R}(\dot{Z}(s))\,\mathrm{d}s = \boldsymbol{E}(0,U(0),Z(0)) - \int_0^t \langle\dot{\ell}(s),u_0(s)\rangle\,\mathrm{d}s.$$

Using abstract existence Theorem 3.1 we again obtain the following result as soon as we have established the coercivity assumption (3.2) for the energy $\boldsymbol{E}$.

PROPOSITION 3.5. *Let* $\ell \in \mathrm{C}^{\mathrm{Lip}}([0,T],(\mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d)^*)$. *Then for all stable* $Q^0 = (U^0,Z^0) \in \boldsymbol{Q}$, *problem* (**S**) *and* (**E**) *has a unique solution* $Q=(U,Z) \in \mathrm{C}^{\mathrm{Lip}}([0,T],\boldsymbol{Q})$ *with* $Q(0)=Q^0$.

Proof. It remains to prove that $\boldsymbol{A}:\boldsymbol{Q}\to\boldsymbol{Q}^*$, which is defined via $\boldsymbol{E}(t,U,Z) = \frac{1}{2}\langle \boldsymbol{A}\binom{U}{Z},\binom{U}{Z}\rangle_{\boldsymbol{Q}} - \langle \ell(t),u_0\rangle_{\mathrm{H}^1}$, satisfies (3.2):

$$\exists\,\alpha>0\ \forall\,(U,Z)\in\boldsymbol{Q} \text{ with } \boldsymbol{R}(Z)<\infty: \quad \left\langle \boldsymbol{A}\begin{pmatrix}U\\Z\end{pmatrix},\begin{pmatrix}U\\Z\end{pmatrix}\right\rangle_{\boldsymbol{Q}} \geq \alpha\|(U,Z)\|^2_{\boldsymbol{Q}}. \tag{3.9}$$

By our assumption (3.3c), we immediately obtain the lower estimate

$$\left\langle \boldsymbol{A}\begin{pmatrix}U\\Z\end{pmatrix},\begin{pmatrix}U\\Z\end{pmatrix}\right\rangle_{\boldsymbol{Q}} \geq \widehat{\alpha}\|(\widehat{\boldsymbol{e}}(U),Z)\|^2_{\mathrm{L}^2(\Omega\times\mathcal{Y})} \quad \text{for all } (U,Z)\in\boldsymbol{Q}. \tag{3.10}$$

Next, we use an orthogonality condition for the two-scale limit of gradients. If $\nabla u_\varepsilon \stackrel{\mathrm{w2}}{\rightharpoonup} E\nabla_x u_0 + \nabla_y U_1$ in $\mathrm{L}^2(\Omega\times\mathcal{Y})$, then

$$\int_{\Omega\times\mathcal{Y}} |\nabla_x u_0(x)+\nabla_y U_1(x,y)|^2\,\mathrm{d}y\,\mathrm{d}x = \int_\Omega |\nabla u_0(x)|^2\,\mathrm{d}x + \int_{\Omega\times\mathcal{Y}} |\nabla_y U_1(x,y)|^2\,\mathrm{d}y\,\mathrm{d}x.$$

The mixed terms drop out, since $E\nabla u_0(x,\cdot)$ is constant on $\mathcal{Y}$, while $\nabla_y U_1(x,\cdot)$ has average 0 as it is a derivative of a periodic function. For the symmetric strains we similarly obtain

$$\|\widehat{\boldsymbol{e}}((u_0,U_1))\|^2_{\mathrm{L}^2(\Omega\times\mathcal{Y})} = \|\boldsymbol{e}(u_0)\|^2_{\mathrm{L}^2(\Omega)} + \|\boldsymbol{e}_y(U_1)\|^2_{\mathrm{L}^2(\Omega\times\mathcal{Y})}.$$

With $K_{\mathcal{Y}} = 2\pi^2 \min\{\,|\lambda|^2 \mid 0\neq\lambda\in\Lambda\,\}$ we have the Korn–Poincaré-type inequalities:

$$\forall\,V\in\mathrm{H}^1_{\mathrm{av}}(\mathcal{Y}): \quad \|\boldsymbol{e}_y(V)\|^2_{\mathrm{L}^2(\mathcal{Y})} \geq K_{\mathcal{Y}}\|V\|^2_{\mathrm{L}^2(\mathcal{Y})} \ \text{ and } \ \|\boldsymbol{e}_y(V)\|^2_{\mathrm{L}^2(\mathcal{Y})} \geq \frac{1}{2}\|\nabla_y V\|^2_{\mathrm{L}^2(\mathcal{Y})}.$$

This follows easily by writing $V(y) = \sum_\Lambda V_\lambda \mathrm{e}^{2\mathrm{i}\pi\lambda\cdot y}$ and using Plancherel's identity. Inserting these estimates into (3.10) and employing Korn's inequality for u_0 we obtain

$$\left\langle \boldsymbol{A}\begin{pmatrix}U\\Z\end{pmatrix},\begin{pmatrix}U\\Z\end{pmatrix}\right\rangle_{\boldsymbol{Q}} \geq \widehat{\alpha}\left(C_{\mathrm{Korn}}\|u_0\|^2_{\mathrm{H}^1(\Omega)} + \frac{K_{\mathcal{Y}}}{1+2K_{\mathcal{Y}}}\int_\Omega \|U_1(x,\cdot)\|^2_{\mathrm{H}^1(\mathcal{Y})}\,\mathrm{d}x + \|Z\|^2_{\mathrm{L}^2(\Omega\times\mathcal{Y})}\right),$$

which provides the desired estimate (3.9). ∎

4. Convergence results. This final section addresses the question of under which conditions the solutions $(u_\varepsilon,z_\varepsilon)$ of (S^ε) and (E^ε) have a two-scale limit (U,Z) which is a solution of (**S**) and (**E**). The convergence is taken in the sense of two-scale cross-convergence, and we can build on our theory in section 4.3. In particular, the results of section 2.5 state that $\boldsymbol{E}$ and $\boldsymbol{R}$ are the Γ-limits of the families $(\mathcal{E}_\varepsilon)_\varepsilon$ and $(\mathcal{R}_\varepsilon)_\varepsilon$, respectively, in the Mosco sense.

PROPOSITION 4.1. *Let* $\Omega\subset\mathbb{R}^d$ *be bounded with Lipschitz boundary. Moreover, let* $\mathcal{E}_\varepsilon$, $\mathcal{R}_\varepsilon$, $\boldsymbol{E}$, *and* $\boldsymbol{R}$ *be defined as above such that* (3.3) *and* $\ell\in\mathrm{C}^0([0,T],(\mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d)^*)$

hold. Then for each $t \in [0,T]$ we have the convergences

$$\text{(4.1a)} \quad (u_\varepsilon, z_\varepsilon) \overset{\text{w2c}}{\rightharpoonup} (u_0, U_1, Z) \in \boldsymbol{Q} \implies \left\{ \begin{array}{c} \boldsymbol{E}(t, u_0, U_1, Z) \leq \liminf_{\varepsilon\to 0} \mathcal{E}_\varepsilon(t, u_\varepsilon, z_\varepsilon), \\ \boldsymbol{R}(Z) \leq \liminf_{\varepsilon\to 0} \mathcal{R}_\varepsilon(z_\varepsilon); \end{array} \right.$$

$$\text{(4.1b)} \quad \begin{array}{l} \forall\, (u_0, U_1, Z) \in \boldsymbol{Q}\ \exists\, ((u_\varepsilon, z_\varepsilon))_\varepsilon : \\ (u_\varepsilon, z_\varepsilon) \xrightarrow{\text{s2c}} (u_0, U_1, Z) \textit{ in } \boldsymbol{Q} \quad \textit{and} \end{array} \left\{ \begin{array}{c} \mathcal{E}_\varepsilon(t, u_\varepsilon, z_\varepsilon) \to \boldsymbol{E}(t, u_0, U_1, Z), \\ \mathcal{R}_\varepsilon(z_\varepsilon) \to \boldsymbol{R}(Z), \end{array} \right.$$

where for the recovery sequence in (4.1b) *we may take* $(u_\varepsilon, z_\varepsilon) = (u_0 + \mathcal{G}_\varepsilon(0, U_1), \mathcal{F}_\varepsilon Z)$ *with* $\mathcal{F}_\varepsilon$ *and* $\mathcal{G}_\varepsilon$ *as defined in* (2.8) *and* (2.12)*, respectively.*

Here it is important that $\mathcal{G}_\varepsilon$ maps into $\mathrm{H}^1_0(\Omega)$, such that $u_0 + \mathcal{G}_\varepsilon(0, U_1) \in \mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d$.

Our convergence result for the solutions $(u_\varepsilon, z_\varepsilon) \in \mathrm{C}^{\mathrm{Lip}}([0,T], \mathcal{Q})$ of (S^ε) and (E^ε) to the solution $(U, Z) \in \mathrm{C}^{\mathrm{Lip}}([0,T], \boldsymbol{Q})$ will be an adapted and simplified variant of the two abstract Theorems 3.1 and 3.3 in [MRS07]. The abstract theory is formulated on one single space $\widehat{\mathcal{Q}}$, but in fact the results there are easily generalized to the setting needed here. The following remark gives the alternative way of embedding everything into one big function space $\widehat{\mathcal{Q}}$.

REMARK 4.2. *To show that our situation is included exactly in this setting, we choose*

$$\widehat{\mathcal{Q}} = \widehat{\mathcal{H}} \times \widehat{\mathcal{Z}} \quad \textit{with } \widehat{\mathcal{H}} = \mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d \times \mathrm{L}^2(\mathbb{R}^d; \mathrm{H}^1_{\mathrm{av}}(\mathcal{Y})) \textit{ and } \widehat{\mathcal{Z}} = \mathrm{L}^2(\mathbb{R}^d \times \mathcal{Y})$$

and define an ε-dependent embedding $(u, z) \mapsto (\mathcal{Q}_\varepsilon u, \mathcal{U}_\varepsilon u, \mathcal{T}_\varepsilon z)$, where the $\mathcal{Q}_\varepsilon : \mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d \to \mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d$ and $\mathcal{U} : \mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d \to \mathrm{L}^2(\mathbb{R}^d; \mathrm{H}^1_{\mathrm{av}}(\mathcal{Y}))$ can be defined as indicated in [CDG02]. *Define H_ε as the subspace of $\mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d$ containing the functions u such that $\fint_{\varepsilon(\lambda+Y)} u(x)\, \mathrm{d}x = 0$ for all $\lambda \in \Lambda^-_\varepsilon$; see section* 2.1. *Then let $\mathcal{Q}_\varepsilon$ be the orthogonal projection to the orthogonal complement of H_ε and set $\mathcal{U}_\varepsilon u = \frac{1}{\varepsilon}(\mathrm{id} - \mathcal{Q}_\varepsilon)u$. Finally, we define the functionals in $\widehat{\mathcal{Q}}$ via*

$$\widehat{\mathcal{E}}_\varepsilon(t, u_0, \widehat{U}_1, \widehat{Z}) = \left\{ \begin{array}{cl} \mathcal{E}_\varepsilon(t, u, z) & \textit{if } (u_0, \widehat{U}_1, \widehat{Z}) = (u, \mathcal{Q}_\varepsilon u, \mathcal{T}_\varepsilon z), \\ \infty & \textit{else,} \end{array} \right.$$

$$\widehat{\mathcal{E}}_0(t, u_0, \widehat{U}_1, \widehat{Z}) = \left\{ \begin{array}{cl} \boldsymbol{E}(t, u_0, U_1, Z) & \textit{if } \operatorname{sppt}(\widehat{U}_1, \widehat{Z}) \subset \overline{\Omega} \times \mathcal{Y}, \\ \infty & \textit{else,} \end{array} \right.$$

$$\widehat{\mathcal{R}}_\varepsilon(\widehat{Z}) = \left\{ \begin{array}{cl} \mathcal{R}_\varepsilon(z) & \textit{if } \widehat{Z} = \mathcal{T}_\varepsilon z, \\ \infty & \textit{else,} \end{array} \right. \qquad \widehat{\mathcal{R}}_0(Z) = \left\{ \begin{array}{cl} \boldsymbol{R}(Z) & \textit{if } \operatorname{sppt}(Z) \subset \overline{\Omega} \times \mathcal{Y}, \\ \infty & \textit{else.} \end{array} \right.$$

Hence, under the additional assumption that for all considered functions the corresponding functionals have finite values, we have concluded that weak and strong convergences in $\widehat{\mathcal{Q}}$ are equivalent to weak or strong two-scale convergence of families $(u_\varepsilon, z_\varepsilon)_\varepsilon$ in $\mathcal{Q}$ towards a limit $(u_0, U_1, Z) \in \boldsymbol{Q}$.

Now we are able to formulate the main result of this paper. It states that the solutions $(u_\varepsilon, z_\varepsilon)_\varepsilon$ of the ε-periodic problem (S^ε) and (E^ε) strongly two-scale cross-converge to a solution (U, Z) of the two-scale homogenized problem (**S**) and (**E**) under the sole assumption that the initial conditions strongly two-scale cross-converge.

THEOREM 4.3. *Let $(u_\varepsilon, z_\varepsilon) : [0,T] \to \mathcal{Q}$ be the solution for (S^ε) and (E^ε) as obtained in Proposition* 3.4. *Assume that the initial data satisfy*

$$(u_\varepsilon(0), z_\varepsilon(0)) \xrightarrow{\text{s2c}} Q^0 = (u^0, U^0, Z^0) \textit{ in } \boldsymbol{Q}.$$

Then Q^0 is stable (i.e., $Q^0 \in \boldsymbol{S}(0)$) and

$$\forall\, t \in [0,T]: \quad (u_\varepsilon(t), z_\varepsilon(t)) \xrightarrow{\text{s2c}} Q(t) = (u_0(t), U_1(t), Z(t)) \text{ in } \boldsymbol{Q},$$

where $Q : [0,T] \to \boldsymbol{Q}$ is the unique solution of (**S**) *and* (**E**) *with initial condition $Q(0) = Q^0$ as provided in Proposition* 3.5.

Recall the definition of the stable sets

$$\begin{aligned}
&\mathcal{S}_\varepsilon(t) = \{\, (u,z) \in \mathcal{Q} \mid \text{for all } (\widetilde{u},\widetilde{z}) \in \mathcal{Q}\text{: } \mathcal{E}_\varepsilon(t,\widetilde{u},\widetilde{z}) \le \mathcal{E}_\varepsilon(0,\widetilde{u},\widetilde{z}) - \mathcal{R}_\varepsilon(\widetilde{z}-z) \,\},\\
&\boldsymbol{S}(t) = \{\, (U,Z) \in \boldsymbol{Q} \mid \text{for all } (\widetilde{U},\widetilde{Z}) \in \boldsymbol{Q}\text{: } \boldsymbol{E}(t,\widetilde{U},\widetilde{Z}) \le \boldsymbol{E}(0,\widetilde{U},\widetilde{Z}) - \boldsymbol{R}(\widetilde{Z}-Z) \,\}.
\end{aligned}$$

REMARK 4.4. *In* [MRS07] *the convergence of the initial condition and of the solutions is formulated in terms of the underlying topology, which in the present setting means weak two-scale cross-convergence. However, the abstract theory assumes convergence of the initial energies and proves convergence of the energies $\mathcal{E}_\varepsilon(t, u_\varepsilon(t), z_\varepsilon(t)) \to \boldsymbol{E}(t, U(t), Z(t))$. Because of uniform convexity (cf.* (3.9)*) we see that weak convergence and energy convergence imply strong convergence. The details of this argument are worked out at the end of the proof of Theorem* 4.3. *See also* [Vis84] *for general arguments of this type.*

The main difficulty in the proof of the desired result is proving that the weak limit of stable states is again stable. In [MRS07] this property is reduced to a property which postulates the existence of suitable *joint recovery sequences* for a combination of $\mathcal{E}_\varepsilon$ and $\mathcal{R}_\varepsilon$. In our setting this reads as follows.

PROPOSITION 4.5. *For $t \in [0,T]$ assume that $(u_\varepsilon, z_\varepsilon) \in \mathcal{S}_\varepsilon(t)$ and $(u_\varepsilon, z_\varepsilon) \overset{\text{w2c}}{\rightharpoonup} (u_0, U_1, Z)$ in $\boldsymbol{Q}$.*

(a) *Then for each $(\widetilde{u}_0, \widetilde{U}_1, \widetilde{Z}) \in \boldsymbol{Q}$ there exists a joint recovery family $(\widetilde{u}_\varepsilon, \widetilde{z}_\varepsilon)_\varepsilon$ with $(\widetilde{u}_\varepsilon, \widetilde{z}_\varepsilon) \overset{\text{w2c}}{\rightharpoonup} (\widetilde{u}_0, \widetilde{U}_1, \widetilde{Z})$ in $\boldsymbol{Q}$ such that*

$$\begin{aligned}
&\limsup_{\varepsilon\to 0} \big[\mathcal{E}_\varepsilon(t,\widetilde{u}_\varepsilon,\widetilde{z}_\varepsilon) + \mathcal{R}_\varepsilon(\widetilde{z}_\varepsilon - z_\varepsilon) - \mathcal{E}_\varepsilon(t,u_\varepsilon,z_\varepsilon)\big]\\
&\le \boldsymbol{E}(t,\widetilde{U},\widetilde{Z}) + \boldsymbol{R}(\widetilde{Z}-Z) - \boldsymbol{E}(t,U,Z).
\end{aligned} \tag{4.2}$$

(b) *As a consequence $(u_0, U_1, Z) \in \boldsymbol{S}(t)$.*

Proof. (a) We give the joint recovery sequence explicitly in the form

$$(\widetilde{u}_\varepsilon, \widetilde{z}_\varepsilon) = (u_\varepsilon, z_\varepsilon) + (\widetilde{u}_0 - u_0 + \mathcal{G}_\varepsilon(0, \widetilde{U}_1 - U_1)\,,\ \mathcal{F}_\varepsilon(\widetilde{Z} - Z)).$$

Note that the arguments for $\mathcal{G}_\varepsilon$ and $\mathcal{F}_\varepsilon$ do not depend on ε. Hence, by Propositions 2.10 and 2.4 we obtain the important relation

$$(\widetilde{u}_\varepsilon, \widetilde{z}_\varepsilon) - (u_\varepsilon, z_\varepsilon) = (\widetilde{u}_0 - u_0 + \mathcal{G}_\varepsilon(0, \widetilde{U}_1 - U_1)\,,\ \mathcal{F}_\varepsilon(\widetilde{Z} - Z)) \xrightarrow{\text{s2c}} (\widetilde{u}_0 - u_0, \widetilde{U}_1 - U_1, \widetilde{Z} - Z). \tag{4.3}$$

In turn, this implies the obvious convergence $(\widetilde{u}_\varepsilon, \widetilde{z}_\varepsilon) \overset{\text{w2c}}{\rightharpoonup} (\widetilde{u}_0, \widetilde{U}_1, \widetilde{Z})$.

From (4.3) and Lemma 2.12(b) we obtain $\mathcal{R}_\varepsilon(\widetilde{z}_\varepsilon - z_\varepsilon) \to \boldsymbol{R}(\widetilde{Z} - Z)$.

For the energies we use the quadratic nature and obtain

$$\begin{aligned}
&\mathcal{E}_\varepsilon(t,\widetilde{u}_\varepsilon,\widetilde{z}_\varepsilon) - \mathcal{E}_\varepsilon(t,u_\varepsilon,z_\varepsilon)\\
&= \frac{1}{2}\int_\Omega \left\langle\!\!\left\langle \mathbb{A}\left(\left\{\frac{x}{\varepsilon}\right\}_{\mathcal{Y}}\right) \begin{pmatrix} \boldsymbol{e}(\widetilde{u}_\varepsilon - u_\varepsilon) \\ \widetilde{z}_\varepsilon + z_\varepsilon \end{pmatrix}, \begin{pmatrix} \boldsymbol{e}(\widetilde{u}_\varepsilon - u_\varepsilon) \\ \widetilde{z}_\varepsilon + z_\varepsilon \end{pmatrix} \right\rangle\!\!\right\rangle \mathrm{d}x - \langle \ell(t), \widetilde{u}_\varepsilon - u_\varepsilon \rangle.
\end{aligned}$$

The last term obviously converges to $\langle \ell(t), \widetilde{u}_0{-}u_0\rangle$ by the usual weak convergence in $\mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega)^d$. Under the integral we have a quadratic form, where the right factor weakly two-scale converges to $\binom{\widetilde{\mathbf{e}}(\widetilde{U}+U)}{\widetilde{Z}+Z}$ in $\mathrm{L}^2(\Omega{\times}\mathcal{Y})$. The left-hand factor is a product of the matrix $m_\varepsilon = \mathbb{A}(\{\frac{\cdot}{\varepsilon}\}_{\mathcal{Y}})$ and a strongly two-scale convergent sequence with limit $\binom{\widetilde{\mathbf{e}}(\widetilde{U}-U)}{\widetilde{Z}-Z}$ in $\mathrm{L}^2(\Omega{\times}\mathcal{Y})$. As $\mathcal{T}_\varepsilon m_\varepsilon(x,y) = \mathbb{A}(y)$ Proposition 2.6 implies

$$\mathbb{A}\left(\left\{\frac{\cdot}{\varepsilon}\right\}_{\mathcal{Y}}\right)\begin{pmatrix}\boldsymbol{e}(\widetilde{u}_\varepsilon{-}u_\varepsilon)\\ \widetilde{z}_\varepsilon{+}z_\varepsilon\end{pmatrix} \xrightarrow{\mathrm{s2}} \mathbb{A}\begin{pmatrix}\widetilde{\boldsymbol{e}}(\widetilde{U}{-}U)\\ \widetilde{Z}{-}Z\end{pmatrix} \text{ in } \mathrm{L}^2(\Omega{\times}\mathcal{Y}).$$

Since a scalar product of a weakly and a strongly converging sequence converges (see Proposition 2.4(d)), we conclude

$$\mathcal{E}_\varepsilon(t,\widetilde{u}_\varepsilon,\widetilde{z}_\varepsilon){-}\mathcal{E}_\varepsilon(t,u_\varepsilon,z_\varepsilon) \to \boldsymbol{E}(t,\widetilde{U},\widetilde{Z}){-}\boldsymbol{E}(t,U,Z).$$

Thus, we have established (4.2) in the stronger version that the limsup is a limit and the "$\leq$" is "$=$".

(b) This is a direct consequence of part (a). Let (U,Z) be the limit of stable states and take any test state $(\widetilde{U},\widetilde{Z}) \in \boldsymbol{Q}$. Now take the joint recovery sequence obtained in part (a) and insert $(\widetilde{u}_\varepsilon,\widetilde{z}_\varepsilon)$ into the stability condition for $(u_\varepsilon,z_\varepsilon)$, namely,

$$0 \leq \mathcal{E}_\varepsilon(t,\widetilde{u}_\varepsilon,\widetilde{z}_\varepsilon) + \mathcal{R}_\varepsilon(\widetilde{z}_\varepsilon{-}z_\varepsilon) - \mathcal{E}_\varepsilon(t,u_\varepsilon,z_\varepsilon).$$

As the right-hand side converges we conclude $0 \leq \boldsymbol{E}(t,\widetilde{U},\widetilde{Z}){+}\boldsymbol{R}(\widetilde{Z}{-}Z){-}\boldsymbol{E}(t,U,Z)$ and stability is established as $(\widetilde{U},\widetilde{Z})$ was arbitrary. $\square$

Proof of Theorem 4.3. By Proposition 3.4 we know that the family $(u_\varepsilon,z_\varepsilon)_\varepsilon$ is uniformly bounded in $\mathrm{C}^{\mathrm{Lip}}([0,T],\mathcal{Q})$. As closed balls in $\mathcal{Q}$ are weakly compact and have a metrizable topology, the Arzelà–Ascoli theorem can be applied in $\mathrm{C}^0([0,T],\mathcal{Q}_{\mathrm{weak}})$, and we find a subsequence $(\varepsilon_k)_{k\in\mathbb{N}}$ with $0<\varepsilon_k\to 0$ such that

$$\forall\, t\in[0,T]: \quad (u_{\varepsilon_k}(t),z_{\varepsilon_k}(t)) \stackrel{\mathrm{w2c}}{\rightharpoonup} (U(t),Z(t)) \text{ in } \boldsymbol{Q}.$$

By the lower semicontinuity of the norm, we have $(U,Z) \in \mathrm{C}^{\mathrm{Lip}}([0,T],\boldsymbol{Q})$ and it remains to show that (U,Z) is a solution of (**S**) and (**E**). As the initial condition (U^0,Z^0) is known, the solution is unique, and we even conclude that the whole family converges (by the standard argument via contradiction).

By Proposition 4.5 we know that $(U(t),Z(t))$ is stable for all $t\in[0,T]$, and hence (**S**) is satisfied and we have to establish the energy balance (**E**) in (3.8). For this, we pass to the limit $\varepsilon\to 0$ in (E^ε); cf. (3.6). The first term on the right-hand side converges, as the energy $\mathcal{E}_\varepsilon(0,u_\varepsilon(0),z_\varepsilon(0))$ converges applying the strong two-scale cross-convergence and Proposition 2.15. The second term converges by Lebesgue's dominated convergence theorem as the integrands are uniformly bounded and converge pointwise.

To treat the left-hand side of (E^ε) we let $e_\varepsilon(t) = \mathcal{E}_\varepsilon(t,u_\varepsilon(t),z_\varepsilon(t))$ and $d_\varepsilon(t) = \int_0^t \mathcal{R}_\varepsilon(z_\varepsilon(s))\,\mathrm{d}s$. By the above, we know that $r_\varepsilon(t) = e_\varepsilon(t)+d_\varepsilon(t)$ converges to $r_0(t)$, which is the limit of the right-hand side. We let $e^*(t) = \limsup_{\varepsilon\to 0} e_\varepsilon(t)$ and $d_*(t) = \liminf_{\varepsilon\to 0} d_\varepsilon(t)$ and conclude $e^*(t)+d_*(t) = r_0(t)$. Now we use the lower estimates for the functionals. For the stored energy we use (4.1a) to obtain

$$\boldsymbol{E}(t,U(t),Z(t)) \leq \liminf_{\varepsilon\to 0} e_\varepsilon(t) \leq \limsup_{\varepsilon\to 0} e_\varepsilon(t) = e^*(t).$$

For the dissipation integral we use $\int_0^t \boldsymbol{R}(\dot Z(s))\,\mathrm{d}s = \sup \sum_{j=1}^N \boldsymbol{R}(Z(t_j){-}Z(t_{j-1}))$, where the supremum is taken over all finite partitions of $[0,t]$. Again by (4.1a) we find

$$\begin{aligned}\sum_{j=1}^N \boldsymbol{R}(Z(t_j){-}Z(t_{j-1})) &\le \liminf_{\varepsilon\to 0} \sum_{j=1}^N \mathcal{R}_\varepsilon(z_\varepsilon(t_j){-}z_\varepsilon(t_{j-1}))\\ &\le \liminf_{\varepsilon\to 0}\int_0^t \mathcal{R}_\varepsilon(\dot z_\varepsilon(s))\,\mathrm{d}s = d_*(t).\end{aligned}\tag{4.4}$$

Thus, recalling $e^*+d_*=r_0$ we proved the lower energy estimate

$$\boldsymbol{E}(t,U(t),Z(t)){+}\int_0^T \boldsymbol{R}(\dot Z(s))\,\mathrm{d}s \le e^*(t){+}d_*(t) = \boldsymbol{E}(0,U(0),Z(0)){-}\int_0^t \langle \ell(s),u_0(s)\rangle\,\mathrm{d}s.$$

The upper energy estimate (just replace "$\le$" by "$\ge$") follows from the already established stability of (U,Z); see [MTL02, Thm. 2.5] or [MM05, Thm. 4.4]. Thus, (**E**) holds and, moreover, we also conclude that the inequality in (4.4) must be an equality. This in turn implies that $\boldsymbol{E}(t,U(t),Z(t)) = e^*(t) = \lim_{\varepsilon\to 0}\mathcal{E}_\varepsilon(t,u_\varepsilon(t),z_\varepsilon(t))$.

As the value of $t\in[0,T]$ is kept from now on, we omit it in the rest of the proof. From the above and using the weak two-scale convergence $q_\varepsilon=(u_\varepsilon,z_\varepsilon)\overset{\mathrm{w2c}}{\rightharpoonup} Q = (u_0,U_1,Z)$, we want to conclude $q_\varepsilon \xrightarrow{\mathrm{s2c}} Q$.

For this, we define $\widehat q_\varepsilon = (u_0+\mathcal{G}_\varepsilon(0,U_1),\mathcal{F}_\varepsilon Z)\in \mathcal{Q}$, which satisfies $\widehat q_\varepsilon \xrightarrow{\mathrm{s2c}} Q$. Moreover, we have

$$\begin{aligned}&\frac{\alpha}{2}\|\widehat q_\varepsilon{-}q_\varepsilon\|^2_{\mathcal{Q}} \le \frac12 \big\langle\!\big\langle \mathcal{A}_\varepsilon(\widehat q_\varepsilon{-}q_\varepsilon),(\widehat q_\varepsilon{-}q_\varepsilon)\big\rangle\!\big\rangle\\ &= \mathcal{E}_\varepsilon(t,q_\varepsilon)-\mathcal{E}_\varepsilon(t,\widehat q_\varepsilon) + \big\langle\!\big\langle \mathcal{A}_\varepsilon\widehat q_\varepsilon{-}\ell, q_\varepsilon{-}\widehat q_\varepsilon\big\rangle\!\big\rangle\\ &\to e^*(t)-\boldsymbol{E}(t,Q)+0=0.\end{aligned}$$

For the convergence note that the first term was treated above, that the second term converges because of "$\xrightarrow{\mathrm{s2c}}$" and Proposition 2.15, and that the third term converges as a scalar product, since the left-hand term is strongly convergent and while the right-hand term weakly converges to 0; see Proposition 2.4(d). Finally, we conclude by noting that

$$\begin{aligned}&\|(\mathcal{T}_\varepsilon(\nabla u_\varepsilon),z_\varepsilon)-(E\nabla_x u_0{+}\nabla_y U_1,Z)\|_{\mathrm{L}^2(\mathbb{R}^d\times\mathcal{Y})}\\ &\le \|(\mathcal{T}_\varepsilon(\nabla u_\varepsilon{-}\nabla\widehat u_\varepsilon),z_\varepsilon{-}\widehat z_\varepsilon)\|_{\mathrm{L}^2(\mathbb{R}^d\times\mathcal{Y})}+\delta_\varepsilon \;\le\; \|(u_\varepsilon,z_\varepsilon){-}(\widehat u_\varepsilon,\widehat z_\varepsilon)\|_{\mathcal{Q}}+\delta_\varepsilon\to 0\end{aligned}$$

with $\delta_\varepsilon = \|(\mathcal{T}_\varepsilon(\nabla\widehat u_\varepsilon),\widehat z_\varepsilon)-(E\nabla_x u_0{+}\nabla_y U_1,Z)\|_{\mathrm{L}^2(\mathbb{R}^d\times\mathcal{Y})}\to 0$ because of $\widehat q_\varepsilon\xrightarrow{\mathrm{s2c}} Q$. This establishes $q_\varepsilon\xrightarrow{\mathrm{s2c}} Q$ and we are done. ∎

Acknowledgments. The authors are grateful to two unknown referees and to Doina Cioranescu and Alain Damlamian for helpful comments and for pointing us to relevant references.

REFERENCES

[ADH90] T. Arbogast, J. Douglas, Jr., and U. Hornung, *Derivation of the double porosity model of single phase flow via homogenization theory*, SIAM J. Math. Anal., 21 (1990), pp. 823–836.

[Alb00] H.-D. Alber, *Evolving microstructure and homogenization*, Contin. Mech. Thermodyn., 12 (2000), pp. 235–286.

[Alb03] H.-D. Alber, *Justification of homogenized models for viscoplastic material behavior*, in Deformation and Failure in Metallic Materials, K. Hutter and H. Baaser, eds., Lecture Notes in Appl. Mech. 10, Springer-Verlag, New York, 2003, pp. 295–319.

[All92] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[BLM96] A. Bourgeat, S. Luckhaus, and A. Mikelić, *Convergence of the homogenization process for a double-porosity model of immiscible two-phase flow*, SIAM J. Math. Anal., 27 (1996), pp. 1520–1543.

[BM93] A. Bourgeat and A. Mikelić, *A note on homogenization of Bingham flow through a porous media*, J. Math. Pure Appl., 72 (1993), pp. 405–414.

[BS96] M. Brokate and J. Sprekels, *Hysteresis and Phase Transitions*, Springer-Verlag, New York, 1996.

[CD99] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford Lecture Ser. Math. Appl. 17, Clarendon Press, Oxford University Press, New York, 1999.

[CDD04] D. Cioranescu, A. Damlamian, and R. De Arcangelis, *Homogenization of nonlinear integrals via the periodic unfolding method*, C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 77–82.

[CDD06] D. Cioranescu, A. Damlamian, and R. De Arcangelis, *Homogenization of quasiconvex integrals via the periodic unfolding method*, SIAM J. Math. Anal., 37 (2006), pp. 1435–1453.

[CDG02] D. Cioranescu, A. Damlamian, and G. Griso, *Periodic unfolding and homogenization*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 99–104.

[CS04] V. Chiadò Piat and G. V. Sandrakov, *Homogenization of some variational inequalities for elasto-plastic torsion problems*, Asymptot. Anal., 40 (2004), pp. 1–23.

[Dac89] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, 1989.

[DDM06] G. Dal Maso, A. DeSimone, and M. G. Mora, *Quasistatic evolution problems for linearly elastic-perfectly plastic materials*, Arch. Ration. Mech. Anal., 180 (2006), pp. 237–291.

[EKK02] C. Eck, P. Knabner, and S. Korotov, *A two-scale method for the computation of solid-liquid phase transitions with dendritic microstructure*, J. Comput. Phys., 178 (2002), pp. 58–80.

[FM06] G. Francfort and A. Mielke, *Existence results for a class of rate-independent material models with nonconvex elastic energies*, J. Reine Angew. Math., 595 (2006), pp. 55–91.

[Gri05] G. Griso, *Les méthodes d'éclatement en homogénéisation périodique et en élasticité linéarisée*, Thèse d'Habilitation, Université Pierre et Marie Curie, Paris, 2005.

[HJM94] U. Hornung, W. Jäger, and A. Mikelic, *Reactive transport through an array of cells with semi-permeable membranes*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 59–94.

[HR99] W. Han and B. D. Reddy, *Plasticity (Mathematical Theory and Numerical Analysis)*, Interdiscip. Appl. Math. 9, Springer-Verlag, New York, 1999.

[JKO94] V. V. Jikov, S. M. Kozlov, and O. A. Oleĭnik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994. Translated from the Russian by G. A. Yosifian.

[LNW02] D. Lukkassen, G. Nguetseng, and P. Wall, *Two-scale convergence*, Int. J. Pure Appl. Math., 2 (2002), pp. 35–86.

[LSB99] M. Lenczner and G. Senouci-Bereksi, *Homogenization of electrical networks including voltage-to-voltage amplifiers*, Math. Models Methods Appl. Sci., 9 (1999), pp. 899–932.

[Mie03] A. Mielke, *Energetic formulation of multiplicative elasto-plasticity using dissipation distances*, Contin. Mech. Thermodyn., 15 (2003), pp. 351–382.

[Mie04] A. Mielke, *Existence of minimizers in incremental elasto-plasticity with finite strains*, SIAM J. Math. Anal., 36 (2004), pp. 384–404.

[Mie05] A. Mielke, *Evolution in rate-independent systems*, in Handbook of Differential Equations, Evolutionary Equations, Vol. 2, C. Dafermos and E. Feireisl, eds., Elsevier B.V., Amsterdam, 2005, Chap. 6, pp. 461–559.

[MM05] A. Mainik and A. Mielke, *Existence results for energetic models for rate-independent systems*, Calc. Var. Partial Differential Equations, 22 (2005), pp. 73–99.

[Mor76] J.-J. Moreau, *Application of convex analysis to the treatment of elastoplastic systems*, in Applications of Methods of Functional Analysis to Problems in Mechanics, P. Germain and B. Nayroles, eds., Lecture Notes in Math. 503, Springer-Verlag, Berlin, 1976, pp. 56–89.

[MRS07] A. Mielke, T. Roubíček, and U. Stefanelli, *Γ-limits and relaxations for rate-independent evolutionary problems*, Calc. Var. Partial Differential Equations, to appear. WIAS Preprint 1156.

[MS02] A. Matache and C. Schwab, *Two-scale FEM for homogenization problems*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 537–572.

[MT99] A. Mielke and F. Theil, *A mathematical model for rate-independent phase transformations with hysteresis*, in Proceedings of the Workshop on Models of Continuum Mechanics in Analysis and Engineering, H.-D. Alber, R. Balean, and R. Farwig, eds., Shaker-Verlag, Aachen, Germany, 1999, pp. 117–129.

[MT04] A. Mielke and F. Theil, *On rate-independent hysteresis models*, NoDEA Nonlinear Differential Equations Appl., 11 (2004), pp. 151–189.

[MT06] A. Mielke and A. Timofte, *An energetic material model for time-dependent ferroelectric behavior: Existence and uniqueness*, Math. Methods Appl. Sci., 29 (2006), pp. 1393–1410.

[MTL02] A. Mielke, F. Theil, and V. I. Levitas, *A variational formulation of rate-independent phase transformations using an extremum principle*, Arch. Ration. Mech. Anal., 162 (2002), pp. 137–177.

[Nes06] S. Nesenenko, *Homogenization and Regularity in Viscoplasticity*, Ph.D. thesis, Technische Universtät Darmstadt, Germnay 2006.

[Nes07] S. Nesenenko, *Homogenization in viscoplasticity*, SIAM J. Math. Anal., 39 (2007), 236–262.

[Ngu89] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.

[Val88] T. Valent, *Boundary Value Problems of Finite Elasticity*, Springer Tracts Nat. Philos. 31, Springer-Verlag, New York, 1988.

[Vis84] A. Visintin, *Strong convergence results related to strict convexity*, Comm. Partial Differential Equations, 9 (1984), pp. 439–466.

[Vis94] A. Visintin, *Differential Models of Hysteresis*, Springer-Verlag, Berlin, 1994.

[Vis96] A. Visintin, *Two-scale Stefan problem*, in Nonlinear Analysis and Applications (Warsaw, 1994), Vol. 7, Gakkōtosho, Tokyo, 1996, pp. 405–424.

[Vis97] A. Visintin, *Two-scale model of phase transitions*, Phys. D, 106 (1997), pp. 66–80.

[Vis04] A. Visintin, *Some properties of two-scale convergence*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl., 15 (2004), pp. 93–107.

[Vis06a] A. Visintin, *Homogenization of the nonlinear Kelvin-Voigt model of viscoelasticity and of the Prager model of plasticity*, Contin. Mech. Thermodyn., 18 (2006), pp. 223–252.

[Vis06b] A. Visintin, *Towards a two-scale calculus*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 371–397.

[Vis07] A. Visintin, *Two-scale convergence of some integral functionals*, Calc. Var. Partial Differential Equations, 29 (2007), pp. 239–265.

[Yos01] G. A. Yosifian, *Homogenization of some contact problems for the system of elasticity in perforated domains*, Rend. Sem. Mat. Univ. Padova, 105 (2001), pp. 37–64.

SIAM J. MATH. ANAL.
Vol. 39, No. 2, pp. 669–685

A SUFFICIENT CONDITION OF REGULARITY FOR AXIALLY SYMMETRIC SOLUTIONS TO THE NAVIER–STOKES EQUATIONS*

G. SEREGIN† AND W. ZAJACZKOWSKI‡

Abstract. In the present paper, we prove a sufficient condition of local regularity for suitable weak solutions to the Navier–Stokes equations having axial symmetry. Our condition is an axially symmetric analogue of the so-called $L_{3,\infty}$-case in the general local regularity theory.

Key words. Navier–Stokes equations, axial symmetry, suitable weak solutions, backward uniqueness

AMS subject classifications. 35K, 76D

DOI. 10.1137/060669838

1. Introduction. In the present paper, we address the problem of regularity for axisymmetric solutions to the Navier–Stokes equations. In contrast to many others (see, for example, [7], [21], [9], [11], [12], [22], [23], and [24]), we study this problem in a local setting.

Our work is motivated by the results of two different papers, [2] and [4]. To explain that, we need the following simple notation. Let e_1, e_2, e_3 be an orthogonal basis of the Cartesian coordinates x_1, x_2, x_3, and let e_ϱ, e_φ, e_3 be an orthogonal basis of the cylindrical coordinates ϱ, φ, x_3 chosen so that

$$e_\varrho = \cos\varphi e_1 + \sin\varphi e_2, \quad e_\varphi = -\sin\varphi e_1 + \cos\varphi e_2, \quad e_3 = e_3.$$

Then, for any vector-valued field v, we have the representations

$$v = v_i e_i = v_1 e_1 + v_2 e_2 + v_3 e_3 = v_\varrho e_\varrho + v_\varphi e_\varphi + v_3 e_3.$$

The classical Navier–Stokes equations, which are invariant with respect to transformation of coordinates, have the form

$$\partial_t v + v\cdot\nabla v - \Delta v + \nabla p = 0, \qquad \mathrm{div} v = 0 \tag{1.1}$$

and are satisfied in some space-time domain. Here, as usual, v and p stand for the velocity field and the pressure field, respectively.

In our considerations, we always assume that v_ϱ, v_φ, v_3, and p are independent of the polar angle φ. In [2], Chae and Lee consider the Cauchy problem for the Navier–Stokes equations under the above assumption on axial symmetry. In addition to usual conditions on the initial data, the authors of [2] assume that velocity field v obeys

$$\int_0^T dt\Big(\int_{\mathbb{R}^3} |v|^\gamma d\varrho d\varphi dx_3\Big)^{\frac{\alpha}{\gamma}} < +\infty, \tag{1.2}$$

*Received by the editors September 14, 2006; accepted for publication (in revised form) March 22, 2007; published electronically August 15, 2007. The work was supported by the agreement on cooperation between the Polish and Russian Academies of Sciences, signed in Warsaw, December 27, 2002.

http://www.siam.org/journals/sima/39-2/66983.html

†Steklov Institute of Mathematics at St. Petersburg, Fontanka 27, 191023, St. Petersburg, Russia (seregin@pdmi.ras.ru). The work of this author was supported by the Alexander von Humboldt Foundation and by RFFI grant 05-01-00941-a.

‡Institute of Mathematics, Polish Academy of Sciences, Sniadeckich 8, 00-956 Warsaw, Poland (wz@impan.gov.pl). The work of this author was partially supported by MNiSW grant 1 P03A 021 30.

with $1/\alpha + 1/\gamma \leq 1/2$, $2 < \gamma < +\infty$, $2 < \alpha \leq +\infty$, and prove the regularity of solutions to the Cauchy problem for (1.1) on time interval $]0, T[$. In fact, they prove even more: their statement is still true if $|v|$ is replaced with $\sqrt{v_\varrho^2 + v_\varphi^2}$. However, it remains unclear whether or not the regularity takes place in the marginal case $\gamma = 2$ and $\alpha = +\infty$. In our opinion, the case cannot be treated by methods developed in [2] because, in a sense, it is an analogue of the so-called $L_{3,\infty}$-case studied in [4]. In turn, the $L_{3,\infty}$-case is marginal to the so-called Ladyzhenskaya–Prodi–Serrin condition; see [13], [19], [6], [20], [5], [16], and [17]. It seems quite reasonable to interpret the result of [2] (see [2, Theorem 3]) as the Ladyzhenskaya–Prodi–Serrin condition for axially symmetric problems. To treat $L_{3,\infty}$-solutions in a generic setting, one needs a new technique based on backward uniqueness for the heat operator with variable lower order terms. In this paper, we wish to extend this method to the axially symmetric case.

To formulate our main result, we introduce the additional notation:

$$\mathcal{C}(x_0, R) = \{x \in \mathbb{R}^3 \parallel x = (x', x_3),\ x' = (x_1, x_2),$$

$$|x' - x_0'| < R,\ |x_3 - x_{03}| < R\}, \quad \mathcal{C}(R) = \mathcal{C}(0, R), \quad \mathcal{C} = \mathcal{C}(1);$$

$$z = (x, t), \quad z_0 = (x_0, t_0), \quad Q(z_0, R) = \mathcal{C}(x_0, R) \times]t_0 - R^2, t_0[,$$

$$Q(R) = Q(0, R), \quad Q = Q(1).$$

In local analysis, the most reasonable object to study is so-called suitable weak solutions, introduced by Caffarelli, Kohn, and Nirenberg in their celebrated paper [1]. We are going to use the slightly simpler definition of Lin in [10].

Definition 1.1. *The pair v and p is called a suitable weak solution to the Navier–Stokes equations in Q if the following conditions are satisfied:*

$$v \in L_{2,\infty}(Q) \cap W_2^{1,0}(Q), \qquad p \in L_{\frac{3}{2}}(Q),$$

where $W_2^{1,0}(Q) = \{v \in L_2(Q) \parallel \nabla v \in L_2(Q)\}$;

v and p satisfy the Navier–Stokes equations in the sense of distributions;

for a.a. $t \in]-1, 0[$, the local energy inequality

$$\int_{\mathcal{C}} \varphi(x,t)|v(x,t)|^2 dx + 2\int_{-1}^{t}\int_{\mathcal{C}} \varphi|\nabla v|^2 dxdt'$$
$$\leq \int_{-1}^{t}\int_{\mathcal{C}} \Big\{|v|^2(\Delta\varphi + \partial_t\varphi) + v\cdot\nabla\varphi(|v|^2 + 2p)\Big\} dxdt'$$

holds for all nonnegative cut-off functions $\varphi \in C_0^\infty(\mathbb{R}^3 \times \mathbb{R})$ vanishing in a neighborhood of the parabolic boundary of Q.

For discussions of the above definition, we refer the reader to [8] and [17].

The following is our main result.

Theorem 1.2. *Let v and p be an axially symmetric suitable weak solution to the Navier–Stokes equations in Q. Assume that*

$$\mathcal{A}_0 = \operatorname{ess}\sup_{-1 \leq t \leq 0} \int_{\mathcal{C}} \frac{1}{\varrho}|v(x,t)|^2 dx < +\infty. \tag{1.3}$$

Then the point $(x,t)=(0,0)$ is a regular point of v; i.e., there exists $r\in\,]0,1]$ such that v is Hölder continuous in the closure of the cylinder $Q(r)$.

By c, we shall denote all generic constants that may vary from one bound to others.

Our paper is organized as follows. In the second section, we discuss known inequalities of the local regularity theory and prove some useful facts about suitable weak solutions. The proof of the main result is started in the third section with scaling and blowup of our solution at a singular point. We also discuss properties of the blowup velocity and the blowup pressure in this section. In the fourth section, we prove some additional differential properties of axially symmetric suitable weak solutions. They are needed to establish a decay of the blowup velocity at infinity. Finally, we end up with the proof of the main theorem in the fifth section. Here, with the help of backward uniqueness results for the heat operator with variable lower order terms, we show that in fact our blowup velocity is trivial.

2. Preliminaries. In what follows, we are going to make use of the following scaling invariant functionals:

$$A(z_0,r;v)=\mathrm{ess}\sup_{t_0-r^2<t<t_0}\frac{1}{r}\int_{\mathcal{C}(x_0,r)}|v(x,t)|^2dx,\quad C(z_0,r;v)=\frac{1}{r^2}\int_{Q(z_0,r)}|v|^3dz,$$

$$E(z_0,r;v)=\frac{1}{r}\int_{Q(z_0,r)}|\nabla v|^2dz,\qquad D(z_0,r;p)=\frac{1}{r^2}\int_{Q(z_0,r)}|p|^{\frac{3}{2}}dz.$$

First, let us recall that, by the Navier–Stokes equations scaling,

$$v^\lambda(x,t)=\lambda v(\lambda x,\lambda^2t),\qquad p^\lambda(x,t)=\lambda^2p(\lambda x,\lambda^2t),$$

we may define suitable weak solutions to the Navier–Stokes equations in $Q(z_0,R)$. So, if v and p form a suitable weak solution to the Navier–Stokes equations in $Q(z_0,R)$, then, for appropriate choice of the cut-off function in the local energy inequality, we can reduce it to the following invariant form:

$$\begin{aligned}A(z_0,R/2;v)+E(z_0,R/2;v)\le c\Big(&C^{\frac{2}{3}}(z_0,R;v)\\&+C(z_0,R;v)+D(z_0,R;p)\Big).\end{aligned}\tag{2.1}$$

We also need the so-called decay estimate for pressure,

$$D(z_0,r;p)\le c\Big[\frac{r}{r_1}D(z_0,r_1;p)+\Big(\frac{r_1}{r}\Big)^2C(z_0,r_1;v)\Big],\tag{2.2}$$

which is valid for all $0<r\le r_1\le R$. The proof of the latter estimate is given in [14]. Repeating the arguments of Lemma 1.8 in [18], we can prove the following.

LEMMA 2.1. *Let v and p be a suitable weak solution to the Navier–Stokes equations in Q and let*

$$A_0=\sup_{0<r<1}A(0,r;v)<+\infty.\tag{2.3}$$

Then, for any $r\in]0,1/2[$, we have

$$\begin{aligned}C^{\frac{4}{3}}(0,r;v)+D(0,r;p)+E(0,r;v)\le c\Big(&(A_0+1)r^{\frac{1}{2}}(D(0,1;p)\\&+E(0,1;v))+A_0^4+A_0^2+A_0\Big).\end{aligned}\tag{2.4}$$

Lemma 2.1, together with the invariance of our functionals under the Navier–Stokes equations scaling and under the shift in the direction of x_3, gives us the following.

Lemma 2.2. *Under the conditions of Theorem* 1.2*, we have*

$$A(z_0,r;v)+C(z_0,r;v)+D(z_0,r;p)+E(z_0,r;v)\leq \mathcal{A}<+\infty \tag{2.5}$$

for all $z_0=(x_0,0)$, $x_0=(0,b)$, $|b|\leq 1/4$, and for all $0<r\leq 1/4$, where a bound $\mathcal{A}$ depends only on quantities $D(0,1;p)$ and $E(0,1;v)$ and the number $\mathcal{A}_0$, defined in Theorem 1.2.

We say that the pair v and p is a suitable weak solution to the Navier–Stokes equations in the space-time cylinder $\Omega\times]T_1,T_2[$ if, for any $z_0=(x_0,t_0)$ with $x_0\in\Omega$ and $T_1<t\leq T_2$, the pair v and p is a suitable weak solution to the Navier–Stokes equations in $Q(z_0,R)$ for some $R>0$.

Next, let us introduce the family of sets

$$\mathcal{P}(R_1,R_2;a)=\{x\in\mathbb{R}^3\parallel R_1<|x'|<R_2,\,|x_3|<a\}.$$

Now, we would like to formulate and prove the following statement.

Lemma 2.3. *Let v and p be a suitable weak solution to the Navier–Stokes equations in the set $\widehat{Q}=\mathcal{P}(3/4,9/4;3/2)\times]-(3/2)^2,0[$. Assume that*

$$\int_{\widehat{Q}}|v(z)|^6dz\leq m<+\infty. \tag{2.6}$$

Then there exists a function $\Phi_0:\mathbb{R}_+\times\mathbb{R}_+\to\mathbb{R}_+$, nondecreasing in each variables, such that

$$|v(z)|+|\nabla v(z)|\leq \Phi_0(m,\mathcal{A}_*)<+\infty \tag{2.7}$$

for any $z\in\mathcal{P}(1,2;1)\times]-1,0[$. Here,

$$\mathcal{A}_*=\int_{\widehat{Q}}|p(z)|^{\frac32}dz=\int_{-(3/2)^2}^0dt\int_{\mathcal{P}(3/4,9/4;3/2)}|p|^{\frac32}dx.$$

Proof. First, we remark that $Q(z_0,1/4)\subset\widehat{Q}$ for any $z_0\in\mathcal{P}(1,2;1)\times]-1,0[$. It follows from (2.2), Hölder's inequality, and (2.6) that

$$D(z_0,r;p)\leq c\Big[\frac{r}{r_1}D(z_0,r_1;p)+\Big(\frac{r_1}{r}\Big)^2m^{\frac12}r_1^{\frac12}\Big], \tag{2.8}$$

which is valid for all $0<r\leq r_1\leq 1/4$. For $\tau\in]0,1[$, let us take $r=\tau^{k+1}/4$ and $r_1=\tau^k/4$ in (2.8) and find

$$D(z_0,\tau^{k+1}/4;p)\leq c\tau\Big[D(z_0,\tau^k/4;p)+m^{\frac12}\tau^{-3}\tau^{\frac{k}{2}}\Big]$$

for all nonnegative integer numbers k. We can choose $\tau\in]0,1[$ so small to provide

$$c\tau^{\frac34}\leq 1$$

and conclude

$$D(z_0,\tau^{k+1}/4;p)\leq \tau^{\frac14}\Big[D(z_0,\tau^k/4;p)+m^{\frac12}\tau^{-3}\tau^{\frac{k}{2}}\Big]$$

for all nonnegative integer numbers k. The latter inequality may be easily iterated. As a result, we have

$$D(z_0,\tau^{k+1}/4;p)\le \tau^{\frac{k+1}{4}}\Big[D(z_0,1/4;p)+m^{\frac12}\tau^{-3}\sum_{i=0}^{k}\tau^{\frac{i}{4}}\Big]$$

for all nonnegative integer numbers k. So,

$$\begin{aligned}C(z_0,\tau^{k+1}/4;v)+D(z_0,\tau^{k+1}/4;p)&\le cm^{\frac12}\tau^{\frac{k+1}{2}}\\ &\quad+\tau^{\frac{k+1}{4}}\Big[D(z_0,1/4;p)+m^{\frac12}\tau^{-3}(1-\tau^{\frac14})^{-1}\Big]\\ &\le c\Big[m^{\frac12}\tau^{\frac{k+1}{2}}+\tau^{\frac{k+1}{4}}\Big(\mathcal{A}_*+m^{\frac12}\tau^{-3}(1-\tau^{\frac14})^{-1}\Big)\Big]\end{aligned}$$

for all nonnegative integer numbers k. Given $\varepsilon>0$, we can find an integer number k_0 so that

$$c\Big[m^{\frac12}\tau^{\frac{k_0+1}{2}}+\tau^{\frac{k_0+1}{4}}\Big(\mathcal{A}_*+m^{\frac12}\tau^{-3}(1-\tau^{\frac14})^{-1}\Big)\Big]\le\varepsilon.$$

But according to the so-called ε-regularity theory (see, for example, [8], [4], and [17]), the latter implies two bounds,

$$|v(z_0)|\le\frac{c}{r_0}\quad\text{and}\quad|\nabla v(z_0)|\le\frac{c}{r_0^2},$$

where $r_0=\tau^{(k_0+1)}/4$. Lemma 2.3 is proved. $\square$

The last preliminary statement is as follows.

LEMMA 2.4. *Assume that all conditions of Theorem* 1.2 *hold. Then*

$$\int_{\mathcal{C}}\frac{1}{\varrho}|v(x,t)|^2dx\le\mathcal{A}_0 \tag{2.9}$$

for all $t\in\,]-1,0[$.

Proof. It is easy to derive the estimate

$$\int_Q\partial_t v\cdot w dz\le\mathcal{A}_1\Big(\int_Q|\nabla w|^3dz\Big)^{\frac13}$$

for any $C_0^\infty(Q)$. Here, a constant $\mathcal{A}_1$ depends on $C(0,1;v)$, $E(0,1;v)$, and $D(0,1;p)$ only. So, v has the first derivative in t in the space

$$L_{\frac32}(-1,0;W^{-1}_{\frac32}(\mathcal{C})).$$

In turn, the latter, together with boundedness of the energy, implies weak continuity in time in the following sense: the function

$$t\to\int_{\mathcal{C}}v(x,t)\cdot w(x)dx$$

is continuous on $[-1,0]$ for any $w\in L_2(\mathcal{C})$. Now, the statement of the lemma follows from the weak lower semicontinuity of the functional

$$w\in L_2(\mathcal{C})\to\int_{\mathcal{C}}\frac{1}{\varrho}|w(x)|^2dx.$$

Lemma 2.4 is proved. $\square$

3. Scaling and blowup. Here, we begin the proof of Theorem 1.2. Assume that the statement of this theorem is false. Then, according to the local regularity theory for the Navier–Stokes equations, there exist an absolute positive constant ε and a sequence $\{R_k\}_{k=1}^{\infty}$ such that $R_k \to 0$ as $k \to +\infty$ and

$$\frac{1}{R_k^2}\int_{Q(R_k)} |v|^3 dz \geq \varepsilon > 0 \tag{3.1}$$

for all $k \in \mathbb{N}$; see [15, Lemma 3.3] for a similar situation or [17, Proposition 5.7].

Next, we scale v and p in the following way:

$$u^k(y,s) = R_k v(R_k y, R_k^2 s), \quad q^k(y,s) = R_k^2 p(R_k y, R_k^2 s),$$

where $e = (y,s) \in Q(1/R_k)$. Functions u^k and q^k are extended by zero to the whole space-time $\mathbb{R}^3 \times \mathbb{R}$.

Now let us fix numbers a and b in $\mathbb{R}$ so that $a > 0$. Let

$$x_k^b = (0, bR_k), \quad y^b = (0,b), \quad z_k^b = (x_k^b, 0), \quad e^b = (y^b, 0).$$

Obviously, for sufficiently large k,

$$|b|R_k < 1/4, \qquad aR_k < 1/4;$$

by Lemma 2.2, the following estimates are valid:

$$C(z_k^b, aR_k; v) = C(e^b, a; u^k) \leq \mathcal{A},$$

$$E(z_k^b, aR_k; v) = E(e^b, a; u^k) \leq \mathcal{A},$$

$$A(z_k^b, aR_k; v) = A(e^b, a; u^k) \leq \mathcal{A}, \tag{3.2}$$

$$D(z_k^b, aR_k; p) = D(e^b, a; q^k) \leq \mathcal{A}$$

for all $k \geq k_0(a,b)$.

First, let b be equal to zero. In this particular case, we can produce three estimates. The first of them is well known in the Navier–Stokes theory and is but a consequence of multiplicative inequalities:

$$\frac{1}{a^{\frac{5}{2}}}\int_{Q(a)} |u^k|^{\frac{10}{3}} de \leq c(\mathcal{A}). \tag{3.3}$$

The second estimate follows from the Navier–Stokes equations, written for u^k and q^k in the weak form, and from (3.2):

$$\int_{Q(a)} \partial_t u^k \cdot w de \leq c(a,\mathcal{A})\Big(\int_{Q(a)} |\nabla w|^3 de\Big)^{\frac{1}{3}}$$

for all $w \in C_0^\infty(Q(a))$. Hence,

$$\partial_t u^k \text{ is bounded in } L_{\frac{3}{2}}(-a^2, 0; W^{-1}_{\frac{3}{2}}(\mathcal{C}(a))). \tag{3.4}$$

The third estimate comes from our main condition (1.3) and has the form

$$\text{ess}\sup_{-(aR_k)^2\le t\le 0}\int_{\mathcal{C}(aR_k)}\frac{|v(x,t)|^2}{|x'|}dx=\text{ess}\sup_{-a^2\le s\le 0}\int_{\mathcal{C}(a)}\frac{|u^k(y,t)|^2}{|y'|}dy\le\mathcal{A}_0. \tag{3.5}$$

Now, making use of the diagonal process for extending space-time cylinders $Q(a)$ and known compactness arguments, we can select subsequences (still denoted by u^k and q^k) such that, for each $a>0$,

$$\begin{aligned} u^k &\rightharpoonup u \qquad \text{in } W^{1,0}_2(Q(a)),\\ u^k &\overset{\star}{\rightharpoonup} u \qquad \text{in } L_{2,\infty}(Q(a)),\\ u^k &\to u \qquad \text{in } L_3(Q(a)),\\ q^k &\rightharpoonup q \qquad \text{in } L_{\frac{3}{2}}(Q(a)). \end{aligned} \tag{3.6}$$

The aim of our further considerations is to describe properties of limit functions u and q called the blowup velocity and blowup pressure, respectively. They are defined on $\mathbb{R}^3\times\mathbb{R}_-$, where $\mathbb{R}_-=\{s\in\mathbb{R}\,\|\,s\le 0\}$. For each $a>0$, the pair u and q is a suitable weak solution to the Navier–Stokes equations in $Q(a)$. From (3.2) and (3.6), it follows that the limit functions obey the inequalities

$$\begin{aligned} C(e^b,a;u)&\le\mathcal{A},\\ A(e^b,a;u)&\le\mathcal{A},\\ E(e^b,a;u)&\le\mathcal{A},\\ D(e^b,a;q)&\le\mathcal{A} \end{aligned} \tag{3.7}$$

for all $b\in\mathbb{R}$ and for all $0<a\in\mathbb{R}$. Moreover, we can derive from (3.6), (3.5), and (3.1) two additional estimates:

$$\text{ess}\sup_{-\infty<s\le 0}\int_{\mathbb{R}^3}\frac{|u(y,t)|^2}{|y'|}dy\le\mathcal{A}_0 \tag{3.8}$$

and

$$\frac{1}{R_k^2}\int_{Q(R_k)}|v|^3dz=\int_Q|u^k|^3de\to\int_Q|u|^3de\ge\varepsilon. \tag{3.9}$$

According to (3.9), the blowup velocity u is a nontrivial solution to the Navier–Stokes equations in $\mathbb{R}^3\times\mathbb{R}_-$. But we are going to show that in fact $u\equiv 0$. This would contradict (3.9) and prove Theorem 1.2.

Obviously, the blowup velocity field u is axially symmetric and, by Caffarelli–Kohn–Nirenberg-type results, all points $y'\ne 0$ are regular, which make it possible to conclude that all spatial derivatives of u are Hölder continuous in a vicinity of each point with $y'\neq 0$.

We can also make use of the local regularity theory for the Stokes system; see [16] and [17]. According to it and by known multiplicative inequality, we have

$$\|\partial_t u^k\|_{L_{\frac98,\frac32}(Q(a/2))}+\|\nabla^2 u^k\|_{L_{\frac98,\frac32}(Q(a/2))}+\|\nabla q^k\|_{L_{\frac98,\frac32}(Q(a/2))}$$

$$\leq c(a)\Big[\|u^k\cdot\nabla u^k\|_{L_{\frac98,\frac32}(Q(a))}+\|u^k\|_{W^{1,0}_2(Q(a))}+\|q^k\|_{L_{\frac32}(Q(a))}\Big]$$

$$\leq c(a)\Big[\|u^k\|^{\frac23}_{L_{2,\infty}(Q(a))}\|u^k\|^{\frac13}_{W^{1,0}_2(Q(a))}+\cdots\Big]\leq c(a,\mathcal{A}).$$

The latter estimate shows that we can select a subsequence (still denoted by u^k) such that, for any $a>1$,

$$u^k\to u \qquad \text{in } C([-1,0];L_{\frac98}(\mathcal{C}(a))). \tag{3.10}$$

We can exploit (3.10) in the following way. For any fixed positive numbers r_1, r_2, and h, we have

$$\Big(\int_{\mathcal{P}(r_1,r_2;h)}|u(y,0)|^{\frac98}dy\Big)^{\frac89}\leq\Big(\int_{\mathcal{P}(r_1,r_2;h)}|u^k(y,0)-u(y,0)|^{\frac98}dy\Big)^{\frac89}$$
$$+\Big(\int_{\mathcal{P}(r_1,r_2;h)}|u^k(y,0)|^{\frac98}dy\Big)^{\frac89}=\alpha_k+\beta_k.$$

By (3.10),

$$\alpha_k\to 0$$

as $k\to+\infty$. To evaluate β_k, we make use of inverse scaling and Hölder's inequality:

$$\beta_k=\Big(R_k^{-\frac{15}{8}}\int_{\mathcal{P}(R_kr_1,R_kr_2;R_kh)}|v(x,0)|^{\frac98}dx\Big)^{\frac89}$$
$$\leq c(r_1,r_2,h)\Big(\frac{1}{R_k}\int_{\mathcal{P}(R_kr_1,R_kr_2;R_kh)}|v(x,0)|^2dx\Big)^{\frac12}$$
$$\leq c(r_1,r_2,h)\Big(\int_{\mathcal{P}(R_kr_1,R_kr_2;R_kh)}\frac{|v(x,0)|^2}{|x'|}dx\Big)^{\frac12}.$$

Now, it remains to apply Lemma 2.4 at $t=0$ and absolute continuity of Lebesgue's integral and conclude that

$$\beta_k\to 0$$

as $k\to+\infty$. This implies the identity

$$\int_{\mathcal{P}(r_1,r_2;h)}|u(y,0)|^{\frac98}dy=0$$

for all positive numbers r_1, r_2, and h. So, we can state that

$$u(\cdot,0)=0 \qquad \text{in}\quad \mathbb{R}^3. \tag{3.11}$$

4. Estimates of axially symmetric solutions. The main result of this section is going to be as follows.

PROPOSITION 4.1. *Let V and P be a sufficiently smooth axially symmetric solution to the Navier–Stokes equations in $\widetilde{Q} = \widetilde{\mathcal{P}} \times \,]-2^2, 0[$, where $\widetilde{\mathcal{P}} = \mathcal{P}(1/4, 3; 2)$. Then there exists a nondecreasing function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ such that*

$$\sup_{z\in\mathcal{P}(1,2;1)\times]-1,0[} \Big(|V(z)| + |\nabla V(z)|\Big) \leq \Phi(\mathcal{A}_2), \tag{4.1}$$

where

$$\mathcal{A}_2 = \sup_{-2^2<t<0} \int_{\widetilde{\mathcal{P}}} |V(x,t)|^2 dx + \int_{\widetilde{Q}} \Big(|\nabla V|^2 + |V|^3 + |P|^{\frac{3}{2}}\Big) dz.$$

To prove the above proposition, we need the following.

LEMMA 4.2. *Under assumptions of Proposition 4.1, there exists a function $\Phi_1 : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$, nondecreasing in each variable, such that*

$$\sup_{-(7/4)^2<t<0} \int_{\widetilde{\mathcal{P}}_1} |V^a(x,t)|^q dx \leq \Phi_1(q, \mathcal{A}_2), \qquad 1 \leq q < +\infty. \tag{4.2}$$

Here, $V^a = (V_\varrho, V_3)$, $|V^a| = \sqrt{|V_\varrho|^2 + |V_3|^2}$, $\widetilde{\mathcal{P}}_1 = \mathcal{P}(5/16, 11/4; 7/4)$, and $\widetilde{Q}_1 = \widetilde{\mathcal{P}}_1 \times \,]-(7/4)^2, 0[$.

Proof. Let us denote by ω the vorticity of V, i.e., $\omega = \nabla \wedge V$. For $\chi = \omega_\varphi$, V_ϱ, and V_3, we have the following identities:

$$V_{\varrho,\varrho} + V_{3,3} = -\frac{1}{\varrho} V_\varrho, \tag{4.3}$$

$$V_{\varrho,3} - V_{3,\varrho} = \chi, \tag{4.4}$$

$$\partial_t \chi + V_\varrho \chi_{,\varrho} + V_3 \chi_{,3} - \frac{1}{\varrho}\chi V_\varrho - \Big(\chi_{,\varrho\varrho} + \chi_{,33} + \frac{1}{\varrho}\chi_{,\varrho} - \frac{1}{\varrho^2}\chi\Big) = \frac{2}{\varrho} V_\varphi V_{\varphi,3}, \tag{4.5}$$

where we have used the notation

$$f_{,\varrho} = \frac{\partial f}{\partial \varrho}, \qquad f_{,3} = \frac{\partial f}{\partial x_3}.$$

Next, we let $\widetilde{\chi} = \chi\psi$, $\widetilde{V} = V^a\psi$, $\widetilde{V}_\varrho = V_\varrho\psi$, and $\widetilde{V}_3 = V_3\psi$, where a nonnegative smooth and axially symmetric cut-off function ψ vanishes in a neighborhood of the parabolic boundary of $\widetilde{Q}$ and is equal to 1 in $\widetilde{Q}_1$. For $\widetilde{\chi}$, $\widetilde{V}_\varrho$, and $\widetilde{V}_3$, we have

$$\widetilde{V}_{\varrho,\varrho} + \widetilde{V}_{3,3} = -\frac{1}{\varrho}\widetilde{V}_\varrho + V_\varrho \psi_{,\varrho} + V_3 \psi_{,3}, \tag{4.6}$$

$$\widetilde{V}_{\varrho,3} - \widetilde{V}_{3,\varrho} = \widetilde{\chi} + V_\varrho \psi_{,3} - V_3 \psi_{,\varrho}, \tag{4.7}$$

$$\partial_t \widetilde{\chi} + V_\varrho \widetilde{\chi}_{,\varrho} + V_3 \widetilde{\chi}_{,3} - \frac{1}{\varrho} V_\varrho \widetilde{\chi} - \Big(\widetilde{\chi}_{,\varrho\varrho} + \widetilde{\chi}_{,33} + \frac{1}{\varrho}\widetilde{\chi}_{,\varrho} - \frac{1}{\varrho^2}\widetilde{\chi}\Big) = J_1 + J_2 + J_3, \tag{4.8}$$

where

$$J_1=\frac{2}{\varrho}V_\varphi V_{\varphi,3}\psi,$$

$$J_2=\chi\Big(\partial_t\psi-\psi_{,\varrho\varrho}-\psi_{,33}-\frac{1}{\varrho}\psi_{,\varrho}\Big)-2\Big(\chi_{,\varrho}\psi_{,\varrho}+\chi_{,3}\psi_{,3}\Big),$$

$$J_3=\chi\Big(V_\varrho\psi_{,\varrho}+V_3\psi_{,3}\Big).$$

Now, we multiply (4.8) by $\widetilde{\chi}\varrho^{-2}$ and integrate the product by parts over $\widetilde{\mathcal{P}}$:

$$\begin{aligned}&\frac{1}{2}\partial_t\int\limits_{\widetilde{\mathcal{P}}}\Big|\frac{\widetilde{\chi}}{\varrho}\Big|^2dx+\int\limits_{\widetilde{\mathcal{P}}}\Big(\Big|\Big(\frac{\widetilde{\chi}}{\varrho}\Big)_{,\varrho}\Big|^2+\Big|\Big(\frac{\widetilde{\chi}}{\varrho}\Big)_{,3}\Big|^2\Big)dx\\ &=\int\limits_{\widetilde{\mathcal{P}}}J_1\frac{\widetilde{\chi}}{\varrho^2}dx+\int\limits_{\widetilde{\mathcal{P}}}J_2\frac{\widetilde{\chi}}{\varrho^2}dx+\int\limits_{\widetilde{\mathcal{P}}}J_3\frac{\widetilde{\chi}}{\varrho^2}dx.\end{aligned}\tag{4.9}$$

Our aim is to evaluate the right-hand side of (4.9). We start with the first term there:

$$\begin{aligned}\int\limits_{\widetilde{\mathcal{P}}}J_1\frac{\widetilde{\chi}}{\varrho^2}dx&=-\int\limits_{\widetilde{\mathcal{P}}}\frac{V_\varphi^2}{\varrho^2}\Big(\frac{\widetilde{\chi}}{\varrho}\Big)_{,3}\psi dx-\int\limits_{\widetilde{\mathcal{P}}}\frac{V_\varphi^2}{\varrho^2}\frac{\widetilde{\chi}}{\varrho}\psi_{,3}dx\\ &\le c\Big(\int\limits_{\widetilde{\mathcal{P}}}\frac{|V_\varphi|^4}{\varrho^4}dx\Big)^{\frac12}\Big(\int\limits_{\widetilde{\mathcal{P}}}\Big|\Big(\frac{\widetilde{\chi}}{\varrho}\Big)_{,3}\Big|^2dx+\int\limits_{\widetilde{\mathcal{P}}}\Big|\frac{\widetilde{\chi}}{\varrho}\Big|^2dx\Big)^{\frac12}.\end{aligned}$$

To estimate the first multiplier of the right-hand side of the latter inequality, we are going to exploit a two-dimensional feature of our axially symmetric problem in the following way. So, by Ladyzhenskaya's inequality,

$$\begin{aligned}\int\limits_{\widetilde{\mathcal{P}}}|V_\varphi|^4dx&\le c\int\limits_{-2}^{2}\int\limits_{1/4}^{3}|V_\varphi|^4d\varrho dx_3\\ &\le c\int\limits_{-2}^{2}\int\limits_{1/4}^{3}|V_\varphi|^2d\varrho dx_3\int\limits_{-2}^{2}\int\limits_{1/4}^{3}\Big(|V_\varphi|^2+|\nabla_aV_\varphi|^2\Big)d\varrho dx_3\\ &\le c\int\limits_{\widetilde{\mathcal{P}}}|V|^2dx\int\limits_{\widetilde{\mathcal{P}}}\Big(|V|^2+|\nabla V|^2\Big)dx\le c\mathcal{A}_2\int\limits_{\widetilde{\mathcal{P}}}\Big(|V|^2+|\nabla V|^2\Big)dx,\end{aligned}$$

where the notation $\nabla_af=(f_{,\varrho},f_{,3})$ has been used. Thus, we find the first estimate:

$$\begin{aligned}\int\limits_{\widetilde{\mathcal{P}}}J_1\frac{\widetilde{\chi}}{\varrho^2}dx&\le c\mathcal{A}_2^{\frac12}\Big(\int\limits_{\widetilde{\mathcal{P}}}\Big(|V|^2+|\nabla V|^2\Big)dx\Big)^{\frac12}\\ &\times\Big(\int\limits_{\widetilde{\mathcal{P}}}\Big|\nabla_a\Big(\frac{\widetilde{\chi}}{\varrho}\Big)\Big|^2dx+\int\limits_{\widetilde{\mathcal{P}}}\Big|\frac{\widetilde{\chi}}{\varrho}\Big|^2dx\Big)^{\frac12}.\end{aligned}\tag{4.10}$$

For the second term, we have

$$\int\limits_{\widetilde{\mathcal{P}}}J_2\frac{\widetilde{\chi}}{\varrho^2}dx\le c\int\limits_{\widetilde{\mathcal{P}}}|\chi|^2dx.\tag{4.11}$$

The third term is estimated in a slightly different way:

$$\begin{aligned}\int_{\widetilde{\mathcal{P}}} J_3\frac{\widetilde{\chi}}{\varrho^2}dx &= \int_{\widetilde{\mathcal{P}}}\frac{\chi\widetilde{\chi}}{\varrho^2}\Big(V_\varrho\psi_{,\varrho}+V_3\psi_{,3}\Big)dx\\ &\le c\Big(\int_{\widetilde{\mathcal{P}}}|\chi|^2dx\Big)^{\frac12}\Big(\int_{\widetilde{\mathcal{P}}}|V^a\cdot\nabla_a\psi|^2\Big|\frac{\widetilde{\chi}}{\varrho}\Big|^2dx\Big)^{\frac12}\\ &\le c\int_{\widetilde{\mathcal{P}}}|\chi|^2dx+c\Big(\int_{\widetilde{\mathcal{P}}}|V^a\cdot\nabla_a\psi|^4dx\Big)^{\frac12}\Big(\int_{\widetilde{\mathcal{P}}}\Big|\frac{\widetilde{\chi}}{\varrho}\Big|^4dx\Big)^{\frac12},\end{aligned}$$

where we let $V^a\cdot\nabla_a\psi=V_\varrho\psi_{,\varrho}+V_3\psi_{,3}$. To estimate the last term on the right-hand side of the latter relation, we exploit Ladyzhenskaya's inequality once more. So, we have

$$\begin{aligned}\int_{\widetilde{\mathcal{P}}}\Big|\frac{\widetilde{\chi}}{\varrho}\Big|^4dx &\le c\int_{-2}^{2}\int_{1/4}^{3}\Big|\frac{\widetilde{\chi}}{\varrho}\Big|^4d\varrho dx_3\\ &\le c\int_{-2}^{2}\int_{1/4}^{3}\Big|\nabla_a\Big(\frac{\widetilde{\chi}}{\varrho}\Big)\Big|^2d\varrho dx_3\int_{-2}^{2}\int_{1/4}^{3}\Big|\frac{\widetilde{\chi}}{\varrho}\Big|^2d\varrho dx_3\\ &\le c\int_{\widetilde{\mathcal{P}}}\Big|\nabla_a\Big(\frac{\widetilde{\chi}}{\varrho}\Big)\Big|^2dx\int_{\widetilde{\mathcal{P}}}\Big|\frac{\widetilde{\chi}}{\varrho}\Big|^2dx\end{aligned}$$

and, in the same way,

$$\int_{\widetilde{\mathcal{P}}}|V^a\cdot\nabla_a\psi|^4dx\le c\int_{\widetilde{\mathcal{P}}}\Big|\nabla_a\Big(V^a\cdot\nabla_a\psi\Big)\Big|^2dx\int_{\widetilde{\mathcal{P}}}|V^a\cdot\nabla_a\psi|^2dx.$$

As a result, we find

$$\begin{aligned}\int_{\widetilde{\mathcal{P}}} J_3\frac{\widetilde{\chi}}{\varrho^2}dx &\le c\int_{\widetilde{\mathcal{P}}}|\chi|^2dx+c\Big(\int_{\widetilde{\mathcal{P}}}|V_a|^2dx+\int_{\widetilde{\mathcal{P}}}|\nabla_aV^a|^2dx\Big)^{\frac12}\\ &\quad\times\Big(\int_{\widetilde{\mathcal{P}}}|V_a|^2dx\Big)^{\frac12}\Big(\int_{\widetilde{\mathcal{P}}}\Big|\nabla_a\Big(\frac{\widetilde{\chi}}{\varrho}\Big)\Big|^2dx\Big)^{\frac12}\Big(\int_{\widetilde{\mathcal{P}}}\Big|\frac{\widetilde{\chi}}{\varrho}\Big|^2dx\Big)^{\frac12}\\ &\le c\int_{\widetilde{\mathcal{P}}}|\nabla V|^2dx+c\Big(\mathcal{A}_2+\mathcal{A}_2^{\frac12}\Big(\int_{\widetilde{\mathcal{P}}}|\nabla V|^2dx\Big)^{\frac12}\Big)\\ &\quad\times\Big(\int_{\widetilde{\mathcal{P}}}\Big|\nabla_a\Big(\frac{\widetilde{\chi}}{\varrho}\Big)\Big|^2dx\Big)^{\frac12}\Big(\int_{\widetilde{\mathcal{P}}}\Big|\frac{\widetilde{\chi}}{\varrho}\Big|^2dx\Big)^{\frac12}.\end{aligned}\tag{4.12}$$

Combining estimates (4.9)–(4.12) and applying Young's inequality, we arrive at the final inequality:

$$\begin{aligned}\partial_t\int_{\widetilde{\mathcal{P}}}\Big|\frac{\widetilde{\chi}}{\varrho}\Big|^2dx+\int_{\widetilde{\mathcal{P}}}\Big|\nabla_a\Big(\frac{\widetilde{\chi}}{\varrho}\Big)\Big|^2dx &\le c\int_{\widetilde{\mathcal{P}}}|\nabla V|^2dx\\ &\quad+\Big(\mathcal{A}_2^2+\mathcal{A}_2\int_{\widetilde{\mathcal{P}}}|\nabla V|^2dx\Big)\Big(\int_{\widetilde{\mathcal{P}}}\Big|\frac{\widetilde{\chi}}{\varrho}\Big|^2dx+1\Big).\end{aligned}\tag{4.13}$$

Estimate (4.13) implies

$$\|\widetilde{\chi}\|_{L_{2,\infty}(\widetilde{Q})}\le\Phi_3(\mathcal{A}_2).$$

According to (4.6) and (4.7), one may conclude

$$\int_{-2}^{2}\int_{1/4}^{3}\Big|\nabla_a\widetilde{V}\Big|^2d\varrho dx_3\le c\int_{-2}^{2}\int_{1/4}^{3}\Big(|\widetilde{\chi}|^2+|V^a|^2\Big)d\varrho dx_3\le\Phi_3(\mathcal{A}_2)$$

and thus

$$\int_{-2}^{2}\int_{1/4}^{3}\left|\widetilde{V}(x,t)\right|^{q}d\varrho dx_3 \leq \Phi_4(q,\mathcal{A}_2)$$

for all $t\in\,]-2^2,0[$. Now, (4.2) immediately follows from the latter inequality. Lemma 4.2 is proved. $\square$

The second counterpart of the proof of Proposition 4.1 is the following statement.

LEMMA 4.3. *Under the assumptions of Proposition* 4.1*, there exists a nondecreasing function* $\Phi_5:\mathbb{R}_+\to\mathbb{R}_+$ *such that*

$$\int_{\widetilde{Q}_2}|V_\varphi|^6dz \leq \Phi_5(\mathcal{A}_2), \tag{4.14}$$

where $\widetilde{Q}_2=\widetilde{\mathcal{P}}_2\times\,]-(3/2)^2,0[$ *and* $\widetilde{\mathcal{P}}_2=\mathcal{P}(3/8,5/2;3/2)$.

Proof. We know that V_φ satisfies the equation

$$\partial_t V_\varphi + V_\varrho V_{\varphi,\varrho} + V_3V_{\varphi,3} + \frac{1}{\varrho}V_\varrho V_\varphi - \Big(V_{\varphi,\varrho\varrho} + V_{\varphi,33} + \frac{1}{\varrho}V_{\varphi,\varrho} - \frac{1}{\varrho^2}V_\varphi\Big) = 0. \tag{4.15}$$

We fix a nonnegative smooth and axially symmetric cut-off function ψ vanishing in a neighborhood of the parabolic boundary of $\widetilde{Q}_1$ and being equal to 1 in $\widetilde{Q}_2$. Then, for $\widetilde{\alpha}=V_\varphi\psi\varrho$, we have the following identity:

$$\begin{aligned}&\partial_t\widetilde{\alpha} + V_\varrho\widetilde{\alpha}_{,\varrho} + V_3\widetilde{\alpha}_{,3} - \Big(\widetilde{\alpha}_{,\varrho\varrho} + \widetilde{\alpha}_{,33} + \frac{1}{\varrho}\widetilde{\alpha}_{,\varrho}\Big) + \frac{2}{\varrho}\widetilde{\alpha}_{,\varrho}\\ &= \alpha\Big(\partial_t\psi + V_\varrho\psi_{,\varrho} + V_3\psi_{,3}\Big) - \Big(2\alpha_{,\varrho}\psi_{,\varrho} + 2\alpha_{,3}\psi_{,3} + \alpha\psi_{,\varrho\varrho} + \alpha\psi_{,33}\Big) + \frac{1}{\varrho}\alpha\psi_{,\varrho},\end{aligned}$$

where $\alpha=V_\varphi\varrho$.

Then we multiply the latter identity by $\widetilde{\alpha}|\widetilde{\alpha}|^2$ and integrate the product by parts over $\widetilde{\mathcal{P}}_1$,

$$\frac{1}{4}\partial_t\int_{\widetilde{\mathcal{P}}_1}|\widetilde{\alpha}|^4dx + \frac{3}{4}\int_{\widetilde{\mathcal{P}}_1}|\nabla_a(|\widetilde{\alpha}|^2)|^2dx = J_1 + J_2, \tag{4.16}$$

where

$$J_1 = \int_{\widetilde{\mathcal{P}}_1}\alpha\widetilde{\alpha}|\widetilde{\alpha}|^2\Big(V_\varrho\psi_{,\varrho} + V_3\psi_{,3}\Big)dx$$

and

$$J_2 = \int_{\widetilde{\mathcal{P}}_1}\widetilde{\alpha}|\widetilde{\alpha}|^2\Big(\alpha\partial_t\psi - 2\alpha_{,\varrho}\psi_{,\varrho} - 2\alpha_{,3}\psi_{,3} - \alpha\psi_{,\varrho\varrho} - \alpha\psi_{,33} + \frac{1}{\varrho}\alpha\psi_{,\varrho}\Big)dx.$$

We let $\beta=|\widetilde{\alpha}|^2$, then $|\beta|^{\frac{10}{3}}=|\widetilde{\alpha}|^{\frac{20}{3}}$ and

$$\int_{\widetilde{\mathcal{P}}_1}|\beta|^{\frac{10}{3}}dx \leq c\Big(\int_{\widetilde{\mathcal{P}}_1}|\beta|^2dx\Big)^{\frac{2}{3}}\int_{\widetilde{\mathcal{P}}_1}|\nabla\beta|^2dx. \tag{4.17}$$

We start with J_1, setting $\mathcal{A}_3 = \|V^a\|_{L_{4,\infty}(\widetilde{Q}_1)}$. By Hölder's inequality and by multiplicative inequality (4.17),

$$\begin{aligned}
J_1 &\le c\Big(\int_{\widetilde{\mathcal{P}}_1} |\widetilde{\alpha}|^{\frac{20}{3}}\,dx\Big)^{\frac{9}{20}}\Big(\int_{\widetilde{\mathcal{P}}_1} |\alpha|^{\frac{20}{11}}|V^a|^{\frac{20}{11}}\,dx\Big)^{\frac{11}{20}}\\
&\le c\Big(\int_{\widetilde{\mathcal{P}}_1} |\beta|^{\frac{10}{3}}\,dx\Big)^{\frac{9}{20}}\Big(\int_{\widetilde{\mathcal{P}}_1} |\alpha|^{\frac{10}{3}}\,dx\Big)^{\frac{3}{10}}\Big(\int_{\widetilde{\mathcal{P}}_1} |V^a|^4 dx\Big)^{\frac{1}{4}}\\
&\le c\Big(\int_{\widetilde{\mathcal{P}}_1} |\beta|^{\frac{10}{3}}\,dx\Big)^{\frac{9}{20}}\Big(\int_{\widetilde{\mathcal{P}}_1} |V|^2 dx\Big)^{\frac{1}{5}}\Big(\int_{\widetilde{\mathcal{P}}_1} |V|^2 dx+\int_{\widetilde{\mathcal{P}}_1} |\nabla V|^2 dx\Big)^{\frac{3}{10}}\mathcal{A}_3\\
&\le c\Big(\int_{\widetilde{\mathcal{P}}_1} |\beta|^{\frac{10}{3}}\,dx\Big)^{\frac{9}{20}}\mathcal{A}_2^{\frac{1}{5}}\Big(\int_{\widetilde{\mathcal{P}}_1} |\nabla V|^2 dx+\mathcal{A}_2\Big)^{\frac{3}{10}}\mathcal{A}_3\\
&\le c\Big(\int_{\widetilde{\mathcal{P}}_1} |\beta|^2 dx\Big)^{\frac{3}{10}}\Big(\int_{\widetilde{\mathcal{P}}_1} |\nabla\beta|^2 dx\Big)^{\frac{9}{20}}\mathcal{A}_2^{\frac{1}{5}}\Big(\int_{\widetilde{\mathcal{P}}_1} |\nabla V|^2 dx+\mathcal{A}_2\Big)^{\frac{3}{10}}\mathcal{A}_3.
\end{aligned}$$

Term J_2 is estimated in the same way:

$$\begin{aligned}
J_2 &\le c\Big(\int_{\widetilde{\mathcal{P}}_1} |\widetilde{\alpha}|^{\frac{20}{3}}\,dx\Big)^{\frac{9}{20}}\Big(\int_{\widetilde{\mathcal{P}}_1} (|\alpha|+|\alpha_{,\varrho}|+|\alpha_{,3}|)^{\frac{20}{11}}\,dx\Big)^{\frac{11}{20}}\\
&\le c\Big(\int_{\widetilde{\mathcal{P}}_1} |\beta|^2 dx\Big)^{\frac{3}{10}}\Big(\int_{\widetilde{\mathcal{P}}_1} |\nabla\beta|^2 dx\Big)^{\frac{9}{20}}\Big(\int_{\widetilde{\mathcal{P}}_1} |\nabla V|^2 dx+\mathcal{A}_2\Big)^{\frac{1}{2}}.
\end{aligned}$$

Now, making use of Young's inequality, we derive from (4.17) and from the two latter estimates the main inequality:

$$\begin{aligned}
&\partial_t\int_{\widetilde{\mathcal{P}}_1} |\widetilde{\alpha}|^4 dx+\int_{\widetilde{\mathcal{P}}_1} |\nabla_a(|\widetilde{\alpha}|^2)|^2 dx\\
&\le c\Big(\int_{\widetilde{\mathcal{P}}_1} |\beta|^2 dx\Big)^{\frac{6}{11}}\mathcal{A}_2^{\frac{4}{11}}\Big(\int_{\widetilde{\mathcal{P}}_1} |\nabla V|^2 dx+\mathcal{A}_2\Big)^{\frac{6}{11}}\mathcal{A}_3^{\frac{20}{11}}\\
&\quad+c\Big(\int_{\widetilde{\mathcal{P}}_1} |\beta|^2 dx\Big)^{\frac{6}{11}}\Big(\int_{\widetilde{\mathcal{P}}_1} |\nabla V|^2 dx+\mathcal{A}_2\Big)^{\frac{10}{11}}\\
&\le c\int_{\widetilde{\mathcal{P}}_1} |\beta|^2 dx\Big(\int_{\widetilde{\mathcal{P}}_1} |\nabla V|^2 dx+\mathcal{A}_2\Big)+c(\mathcal{A}_2\mathcal{A}_3^5)^{\frac{4}{5}}+c\Big(\int_{\widetilde{\mathcal{P}}_1} |\nabla V|^2 dx+\mathcal{A}_2\Big)^{\frac{4}{5}}.
\end{aligned}$$

This, together with the statement of Lemma 4.2 at $q = 4$, implies

$$\sup_{-(7/4)^2\le t\le 0}\int_{\widetilde{\mathcal{P}}_1} |\beta(x,t)|^2 dx+\int_{\widetilde{Q}_1} |\nabla\beta|^2 dz\le\Phi_5(\mathcal{A}_2). \tag{4.18}$$

So, (4.14) follows from (4.17) and (4.18). Lemma 4.3 is proved. □

From Lemmas 4.2 and 4.3, we find the following.

COROLLARY 4.4. *Under the assumptions of Proposition* 4.1*, there exists a non-decreasing function* $\Phi_6 : \mathbb{R}_+ \to \mathbb{R}_+$ *such that*

$$\int_{\widetilde{Q}_2} |V|^6 dz\le\Phi_6(\mathcal{A}_2). \tag{4.19}$$

Proof of Proposition 4.1. Applying Corollary 4.4 and Lemma 2.3, we end up with the proof of Proposition 4.1. □

5. Proof of Theorem 1.2. Given $R > 1$, let us consider the space-time cylinder

$$\widetilde{Q}^b_R = \widetilde{\mathcal{P}}^b_R \times]-(2R)^2, 0[,$$

where $b \in \mathbb{R}$ and

$$\widetilde{\mathcal{P}}^b_R = \widetilde{\mathcal{P}}_R + be_3, \qquad \widetilde{\mathcal{P}}_R = \mathcal{P}(R/4, 3R; 2R).$$

Now, we scale our blowup functions u and q in the following way:

$$u^R(x,t) = Ru(Rx + be_3, R^2t), \qquad q^R(x,t) = R^2 q(Rx + be_3, R^2t)$$

for $z = (x,t) \in \widetilde{Q}$.

Functions u^R and q^R are axially symmetric and, as was explained before, sufficiently smooth to apply Proposition 4.1. According to that, we have

$$\sup_{z \in \widetilde{Q}_0} \Big\{ |u^R(z)| + |\nabla u^R(z)| \Big\} \le \Phi(\mathcal{A}_2),$$

where $\widetilde{Q}_0 = \mathcal{P}(1,2;1) \times]-1,0[$ and

$$\mathcal{A}_2 = \sup_{-2^2 \le t \le 0} \int_{\widetilde{\mathcal{P}}} |u^R(x,t)|^2 dx + \int_{\widetilde{Q}} \Big(|\nabla u^R|^2 + |u^R|^3 + |q^R|^{\frac{3}{2}} \Big) dz.$$

Then we make the inverse change of variables. As a result, we find

$$\sup_{(y,s) \in Q^b_R} \Big\{ R|u(y,s)| + R^2 |\nabla u(y,s)| \Big\} \le \Phi(\widetilde{\mathcal{A}}_{2R}),$$

where $Q^b_R = \widetilde{\mathcal{P}}^b_{0R} \times]-R^2, 0[$, $\widetilde{\mathcal{P}}^b_{0R} = be_3 + \mathcal{P}_{0R}$, $\mathcal{P}_{0R} = \mathcal{P}(R, 2R; R)$, and

$$\begin{aligned}
\widetilde{\mathcal{A}}_{2R} = & \sup_{-(2R)^2 \le s \le 0} \frac{1}{R} \int_{-2R+b}^{2R+b} dy_3 \int_{R/4 < |y'| < 3R} |u(y,s)|^2 dy' \\
& + \frac{1}{R} \int_{-(2R)^2}^0 ds \int_{-2R+b}^{2R+b} dy_3 \int_{R/4 < |y'| < 3R} |\nabla u(y,s)|^2 dy' \\
& + \frac{1}{R^2} \int_{-(2R)^2}^0 ds \int_{-2R+b}^{2R+b} dy_3 \int_{R/4 < |y'| < 3R} \Big(|u(y,s)|^3 + |q(y,s)|^{\frac{3}{2}} \Big) dy' \\
\le & \; c\Big(A(e^b, 3R; u) + E(e^b, 3R; u) + C(e^b, 3R; u) + D(e^b, 3R; q) \Big) \le c\mathcal{A},
\end{aligned}$$

$e^b = (y^b, 0)$ and $y^b = (0, b)$. So, assuming that $|y'| > 20$, we can derive from the latter estimates

$$|y'||u(y', b, s)| + |y'|^2 |\nabla u(y', b, s)| \le \Phi(c\mathcal{A}) \tag{5.1}$$

for any $b \in \mathbb{R}$, for any $|y'| > 20$, and for any $s \in [-20, 0]$. It follows directly from (5.1) that

$$|u(y,s)| + |\nabla u(y,s)| \le c\Phi(c\mathcal{A}) = c(\mathcal{A}) \tag{5.2}$$

for any y such that $|y'| > 20$ and for any $s \in [-20, 0]$.

Now, we consider the vorticity $\omega(u) = \nabla \wedge u$. It satisfies the vorticity equation

$$\partial_t \omega - \Delta\omega = \omega \cdot \nabla u - u \cdot \nabla \omega,$$

which, together with (5.2), implies

$$|\partial_t \omega - \Delta \omega| \leq c(\mathcal{A})(|\omega| + |\nabla \omega|) \tag{5.3}$$

for any y such that $|y'| > 20$ and for any $s \in [-20, 0]$. Moreover, by (3.11),

$$\omega(\cdot, 0) = 0 \qquad \text{in} \quad \mathbb{R}^3. \tag{5.4}$$

By the backward uniqueness results for the heat operator with variable lower order terms in a half-space (see [3], [4], and [15]) and by (5.3) and (5.4), we state

$$\omega(y, s) = 0 \tag{5.5}$$

for any y such that $|y'| > 20$ and for any $s \in [-20, 0]$.

Since our solution is sufficiently smooth in $\mathbb{R}^3 \setminus \{y' = 0\} \times [-10, 0]$, one can make use of the unique continuation through spatial boundaries and conclude that

$$\nabla \wedge u \equiv 0 \qquad \text{in} \quad \mathbb{R}^3 \setminus \{y' = 0\} \times [-8, 0]. \tag{5.6}$$

On the other hand, from (3.8), it follows that

$$\mathcal{A}_0 \geq \operatorname{ess} \sup_{-20 \leq s \leq 0} \int_{|y'| \leq 40} \frac{|u(y,s)|^2}{|y'|} dy.$$

So, we observe that, for any $s \in S$,

$$\int_{-\infty}^{+\infty} dy_3 \int_{|y'| \leq 40} \frac{|u(y,s)|^2}{|y'|} dy' \leq \mathcal{A}_0 < +\infty, \tag{5.7}$$

where $S \subset [-20, 0]$ and $|S| = 20$.

Now, we wish to show

$$\nabla \wedge u(\cdot, s) \equiv 0 \qquad \text{in} \quad \mathbb{R}^3 \tag{5.8}$$

for any $s \in S$. To this end, we proceed as follows. Let $\varphi \in C_0^\infty(B')$ be a nonnegative cut-off function being equal to 1 in $B'(1/2)$. Here, B' and $B'(1/2)$ are two-dimensional balls centered at the origin with radii 1 and 1/2, respectively. Next, let ψ be an arbitrary smooth, compactly supported in $\mathbb{R}^3$, vector-valued function. Then, by (5.6), for any $s \in [-8, 0]$,

$$\begin{aligned}
&\int_{\mathbb{R}^3} u(y,s) \cdot \nabla \wedge \Big(\psi(y)(1 - \varphi(y'/R))\Big) dy = 0 \\
&= \int_{\mathbb{R}^3} u(y,s) \cdot \nabla \wedge \psi(y) dy - \int_{\mathbb{R}^3} u(y,s) \cdot \nabla \wedge \Big(\psi(y)\varphi(y'/R)\Big) dy = J_1(s) + J_2(s).
\end{aligned}$$

For J_2, we have the estimate

$$\begin{aligned}
|J_2(s)| &\leq c\Big(1 + \frac{1}{R}\Big) \int_{\operatorname{spt}\psi \cap \{|y'| < R\}} |u(y,s)| dy \\
&= c\Big(1 + \frac{1}{R}\Big) \int_{\operatorname{spt}\psi \cap \{|y'| < R\}} \frac{|u(y,s)|}{|y'|^{\frac12}} |y'|^{\frac12} dy \\
&\leq c\Big(1 + \frac{1}{R}\Big) \Big(\int_{\operatorname{spt}\psi \cap \{|y'| < R\}} \frac{|u(y,s)|^2}{|y'|} dy\Big)^{\frac12} \Big(\int_{\operatorname{spt}\psi \cap \{|y'| < R\}} |y'| dy\Big)^{\frac12} \\
&\leq c(\psi)\Big(1 + \frac{1}{R}\Big) R^{\frac32} \Big(\int_{-\infty}^{+\infty} dy_3 \int_{|y'| < 40} \frac{|u(y,s)|^2}{|y'|} dy'\Big)^{\frac12}.
\end{aligned}$$

By (5.7), the right-hand side of the latter inequality goes to zero as $R \to 0$ for any $s \in S$. Hence, $J_1(s) = 0$ for any $s \in S \cap [-8, 0]$, which is but a weak form of (5.8). By the fact that u is divergence-free, we then show

$$\Delta u(\cdot, s) = 0 \qquad \text{in} \quad \mathbb{R}^3$$

for any $s \in S \cap [-8, 0]$.

Now, let $B(y_0, R)$ be a ball of radius R with the center at point y_0. For any $y_0 \in \{|y'| \leq 30,\, y_3 \in \mathbb{R}\}$,

$$B(y_0, 1) \subset \{|y'| \leq 40,\, y_3 \in \mathbb{R}\}$$

and, since u is harmonic,

$$|u(y_0, s)| \leq c\Big(\int_{B(y_0,1)} |u(y,s)|^2 dy\Big)^{\frac{1}{2}} \leq c\Big(\int_{|y'|\leq 40} |u(y,s)|^2 dy\Big)^{\frac{1}{2}} \leq c\sqrt{40\mathcal{A}_0}$$

for any $s \in S \cap [-8, 0]$. So, according to (5.2), the function $u(\cdot, s)$ is bounded in $\mathbb{R}^3$ for any $s \in S \cap [-8, 0]$. However, by (5.1), in fact, $u(\cdot, s) = 0$ in $\mathbb{R}^3$ for any $s \in S \cap [-8, 0]$. This contradicts (3.9). Theorem 1.2 is proved. $\square$

REFERENCES

[1] L. Caffarelli, R.-V. Kohn, and L. Nirenberg, *Partial regularity of suitable weak solutions of the Navier–Stokes equations*, Comm. Pure Appl. Math., 35 (1982), pp. 771–831.

[2] D. Chae and J. Lee, *On the regularity of the axisymmetric solutions of the Navier–Stokes equations*, Math. Z., 239 (2002), pp. 645–671.

[3] L. Escauriaza, G. Seregin, and V. Šverák, *Backward uniqueness for the heat operator in half space*, Algebra and Analysis, 15 (2003), pp. 201–214.

[4] L. Escauriaza, G. Seregin, and V. Šverák, *$L_{3,\infty}$-Solutions to the Navier–Stokes equations and backward uniqueness*, Uspekhi Mat. Nauk, 58 (350), pp. 3–44 (in Russian). English translation in Russian Math. Surveys, 58 (2003), pp. 211–250.

[5] Y. Giga, *Solutions for semilinear parabolic equations in L^p and regularity of weak solutions of the Navier–Stokes equations*, J. Differential Equations, 62 (1986), pp. 186–212.

[6] O. A. Ladyzhenskaya, *On uniqueness and smoothness of generalized solutions to the Navier–Stokes equations*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI), 5 (1967), pp. 169–185.

[7] O. A. Ladyzhenskaya, *On unique solvability of the three-dimensional Cauchy problem for the Navier–Stokes equations under the axial symmetry*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI), 7 (1968), pp. 155–177.

[8] O. A. Ladyzhenskaya and G. A. Seregin, *On partial regularity of suitable weak solutions to the three-dimensional Navier–Stokes equations*, J. Math. Fluid Mech., 1 (1999), pp. 356–387.

[9] S. Leonardi, J. Malek, J. Necas, and M. Pokorny, *On axially symmetric flows in $\mathbb{R}^3$*, Z. Anal. Anwendungen, 18 (1999), pp. 639–649.

[10] F.-H. Lin, *A new proof of the Caffarelli–Kohn–Nirenberg theorem*, Comm. Pure Appl. Math., 51 (1998), pp. 241–257.

[11] J. Neustupa and M. Pokorny, *Axisymmetric flow of Navier–Stokes fluid in the whole space with non-zero angular velocity components*, Math. Bohem., 126 (2001), pp. 469–481.

[12] M. Pokorny, *A regularity criterion for the angular velocity component in the case of axisymmetric Navier–Stokes equations*, in Elliptic and Parabolic Problems (Rolduc/Gaeta, 2001), World Scientific, River Edge, NJ, 2002, pp. 233–242.

[13] G. Prodi, *Un teorema di unicità per el equazioni di Navier–Stokes*, Ann. Mat. Pura Appl., 48 (1959), pp. 173–182.

[14] G. A. Seregin, *On the number of singular points of weak solutions to the Navier–Stokes equations*, Comm. Pure Appl. Math., 54 (2001), pp. 1019–1028.

[15] G. A. Seregin, *On smoothness of $L_{3,\infty}$-solutions to the Navier–Stokes equations up to boundary*, Math. Ann., 332 (2005), pp. 219–238.

[16] G. Seregin, *New version of the Ladyzhenskaya–Prodi–Serrin condition*, Algebra i Analiz, 18 (2006), pp. 124–143.

[17] G. Seregin, *Local regularity theory of the Navier–Stokes equations*, in Handbook of Mathematical Fluid Mechanics, Vol. 4, S. Friedlander and D. Serre, eds., North-Holland, Amsterdam, 2007, pp. 156–200.

[18] G. Seregin, *Estimates of suitable weak solutions to the Navier–Stokes equations in critical Morrey spaces*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI), 336 (2006), pp. 199–210.

[19] J. Serrin, *On the interior regularity of weak solutions of the Navier–Stokes equations*, Arch. Ration. Mech. Anal., 9 (1962), pp. 187–195.

[20] M. Struwe, *On partial regularity results for the Navier–Stokes equations*, Comm. Pure Appl. Math., 41 (1988), pp. 437–458.

[21] M. R. Ukhovskij and V. L. Yudovich, *Axially symmetric motions of ideal and viscous fluids filling all space*, Prikl. Mat. Mech., 32 (1968), pp. 59–69.

[22] M. Wiegner and W. M. Zajaczkowski, *On stability of axially symmetric solutions to Navier–Stokes equations in a cylindrical domain and with boundary slip conditions*, in Regularity and Other Aspects of the Navier-Stokes Equations, Banach Center Publ. 70, Polish Acad. Sci., Warsaw, 2005, pp. 251–278.

[23] W. M. Zajaczkowski, *Global special regular solutions to the Navier–Stokes equations in axially symmetric domains under boundary slip conditions*, Dissertationes Math. (Rozprawy Mat.), 432 (2005).

[24] W. M. Zajaczkowski, *Global Special Regular Solutions to the Navier–Stokes Equations in a Cylindrical Domain under Boundary Slip Conditions*, GAKUTO Internat. Ser. Math. Sci. Appl. 21, Gakkōtosho, Tokyo, 2004.

SIAM J. MATH. ANAL.
Vol. 39, No. 3, pp. 687–720

EFFECTIVE TRANSMISSION CONDITIONS FOR REACTION-DIFFUSION PROCESSES IN DOMAINS SEPARATED BY AN INTERFACE*

MARIA NEUSS-RADU† AND WILLI JÄGER‡

Abstract. In this paper, we develop multiscale methods appropriate for the homogenization of processes in domains containing thin heterogeneous layers. Our model problem consists of a nonlinear reaction-diffusion system defined in such a domain, and properly scaled in the layer region. Both the period of the heterogeneities and the thickness of the layer are of order ε. By performing an asymptotic analysis with respect to the scale parameter ε we derive an effective model which consists of the reaction-diffusion equations on two domains separated by an interface together with appropriate transmission conditions across this interface. These conditions are determined by solving local problems on the standard periodicity cell in the layer. Our asymptotic analysis is based on weak and strong two-scale convergence results for sequences of functions defined on thin heterogeneous layers. For the derivation of the transmission conditions, we develop a new method based on test functions of boundary layer type.

Key words. nonlinear reaction-diffusion systems, thin heterogeneous layer, homogenization, two-scale convergence, transmission conditions

AMS subject classifications. 35K57, 35B27, 80M35

DOI. 10.1137/060665452

1. Introduction.

In this paper, we will be concerned with a nonlinear system of reaction-diffusion equations in a domain containing a thin heterogeneous layer. Such problems often occur in applications like, e.g., transdermal diffusion of drugs, diffusion of substances through the epithelial monolayer, and transport of ions through membranes.

We start from a microscopic model defined on a domain containing a thin layer of thickness ε. The processes are modeled by a system of reaction-diffusion equations properly scaled inside the layer. Our aim is to study the behavior of the solutions of the microscopic equations when the thickness ε tends to zero.

In the limit $\varepsilon \to 0$ the thin layer reduces to an interface between the two bulk regions. We derive an effective model which consists of a system of reaction-diffusion equations on both sides of this interface together with appropriate transmission conditions for the limit concentrations across the interface.

For the derivation of the limit equations in the bulk regions we use standard compactness results based on classical a priori estimates. However, the thin heterogeneous layer poses additional problems. We have to adapt the concepts of weak and strong two-scale convergences to functions on thin domains with oscillatory and (for simplicity) periodic structure. Due to the arising nonlinearities, it is necessary to prove strong two-scale convergence of the solutions to the ε-problems in the thin layer. To this end, we introduce macroscopic and microscopic coordinates and analyze the regularity of the solutions with respect to this pair of coordinates. Whereas the gradients

*Received by the editors July 19, 2006; accepted for publication (in revised from) March 12, 2007; published electronically August 17, 2007.
http://www.siam.org/journals/sima/39-3/66545.html
†IAM, University of Heidelberg, INF 294, 69120 Heidelberg, Germany (maria.neuss-radu@iwr.uni-heidelberg.de).
‡IWR, University of Heidelberg, INF 368, 69120 Heidelberg, Germany (jaeger@iwr.uni-heidelberg.de).

with respect to the microvariable and the time can be controlled in L^2 in a standard way, the control of the dependence on the macrovariable demands a new approach. Using the partial differential equation in the layer and the strong compactness of the traces of the microscopic concentrations on the interfaces between the bulk regions and the thin layer, we are able to derive L^2-equicontinuity also with respect to the macrovariable. When these results are combined, a Kolmogorov criterion for strong two-scale convergence in the thin layer is satisfied.

For the derivation of the transmission conditions we introduce a new method which is based on testing the microscopic equations with test functions of boundary layer type. The effective transmission conditions consist of relations for the jump of the concentrations and the normal fluxes across the interface. They are calculated by solving local problems on the standard periodicity cell in the layer. These local problems are coupled with the effective equations in the bulk regions through the boundary values on the upper and lower boundaries of the standard cell.

Our paper is organized as follows: We start with the precise description of the geometry and of the microscopic equations and the formulation of the main results (section 2). Then we show a priori estimates (section 3). In section 4, we present the convergence concepts needed to pass to the limit in the thin layer. Section 5 gives the proofs for the convergence (up to subsequences) of the microscopic solutions in the bulk and in the layer. Here, the most challenging part is to show strong two-scale convergence of the concentration inside the layer. In section 6, we derive the effective model including the macroscopic equations in the bulk regions, the local problem in the layer, and the transmission conditions across the interface separating the bulk regions. In the last section, we show uniqueness for the homogenized solution and thus the convergence of the whole sequence.

Reducing the computational work is one of the main purposes of the homogenization limit. The algorithms for solving the derived transmission problem numerically will be considered in a forthcoming paper.

The limiting methods developed in this paper are crucial in treating the ion transport through membrane channels modeled by the Nernst–Planck equations. This system couples the transport of ions and the electric field in a domain separated by a thin membrane with periodically distributed channels. In the channels, partially fixed charges control the permeability of the membrane. Here again, the derivation of the transmission conditions is the crucial topic. A formulation of the results and a sketch of the steps necessary in the analysis are given in [14], and the full mathematical derivation is presented in the forthcoming paper [15].

2. Statement of the problem and the results.

2.1. Setting of the problem. Let $\varepsilon > 0$ be a sequence of strictly positive numbers tending to zero, with the property that $\frac{1}{\varepsilon} \in \mathbb{N}$, and let $H > 0$ be a fixed real number.

We consider a bounded domain $\Omega =]0,1[^{n-1} \times]-H,H[\subset \mathbb{R}^n,\ n \geq 2$, consisting of three subdomains: the bulk regions Ω_ε^+, Ω_ε^- and the thin heterogeneous layer Ω_ε^M, separated by the interfaces S_ε^+ and S_ε^-; see Figure 2.1. Thus we have

$$\Omega = \Omega_\varepsilon^+ \cup \Omega_\varepsilon^- \cup \Omega_\varepsilon^M \cup S_\varepsilon^+ \cup S_\varepsilon^-,$$

where $\Omega_\varepsilon^+ =]0,1[^{n-1} \times]\varepsilon, H[$, $\Omega_\varepsilon^- =]0,1[^{n-1} \times]-H,-\varepsilon[$, $\Omega_\varepsilon^M =]0,1[^{n-1} \times]-\varepsilon,\varepsilon[$, $S_\varepsilon^+ =]0,1[^{n-1} \times \{\varepsilon\}$, $S_\varepsilon^- =]0,1[^{n-1} \times \{-\varepsilon\}$. We denote

$$\Sigma =]0,1[^{n-1} \times \{0\}.$$

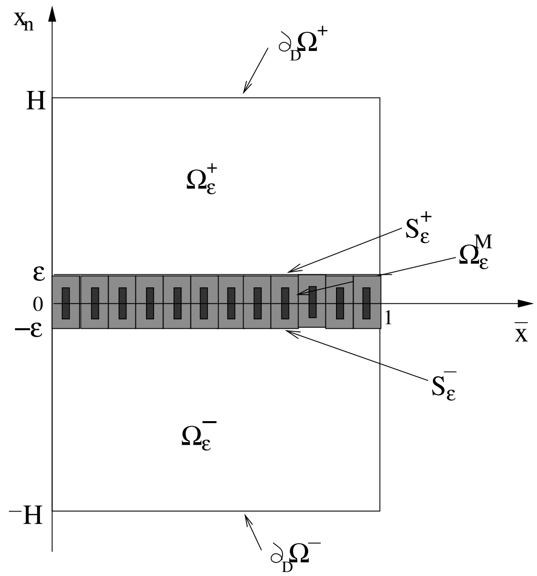


FIG. 2.1. *The domain Ω including the thin heterogeneous layer Ω_ε^M.*

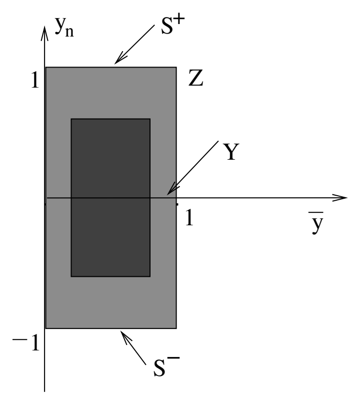


FIG. 2.2. *The standard cell $Z = Y \times [-1,1] = [0,1]^{n-1} \times [-1,1]$.*

The microscopic structure of the layer Ω_ε^M is obtained by the periodic repetition of the standard cell Z (see Figure 2.2), scaled with ε. Here

$$Z = Y \times [-1,1] = [0,1]^{n-1} \times [-1,1],$$

and we denote by

$$S^\pm = \{y \in \mathbb{R}^n : \bar{y} \in Y, y_n = \pm 1\}$$

the upper and lower boundaries of Z. The outer unit normal at the boundaries of the domains Ω and Ω_ε^M is denoted by ν. The restrictions of functions defined on Ω to the subdomains Ω_ε^+, Ω_ε^-, and Ω_ε^M are denoted by the superscripts $+$, $-$, and M, respectively.

In the domain Ω we consider the following system of reaction-diffusion equations for the unknown vector $u_\varepsilon = (u_{1\varepsilon}, \ldots, u_{m\varepsilon}) : (0,T) \times \Omega \to \mathbb{R}^m$,

$$
\begin{aligned}
\partial_t u^+_{j\varepsilon} - D^+_j \Delta u^+_{j\varepsilon} &= f_j(x, u^+_\varepsilon) \qquad \text{in } (0,T)\times\Omega^+_\varepsilon,\\
\partial_t u^-_{j\varepsilon} - D^-_j \Delta u^-_{j\varepsilon} &= f_j(x, u^-_\varepsilon) \qquad \text{in } (0,T)\times\Omega^-_\varepsilon,\\
\frac{1}{\varepsilon}\partial_t u^M_{j\varepsilon} - \nabla\cdot\left(\varepsilon D^M_j\left(\frac{x}{\varepsilon}\right)\nabla u^M_{j\varepsilon}\right) &= \frac{1}{\varepsilon}\, g_j\left(\frac{x}{\varepsilon}, u^M_\varepsilon\right) \text{ in } (0,T)\times\Omega^M_\varepsilon,
\end{aligned}
\tag{2.1}
$$

subjected to the boundary conditions

$$
\begin{aligned}
u^\pm_\varepsilon &= u^\pm_D \quad \text{on } \partial_D\Omega^\pm,\\
\nabla u_{j\varepsilon}\cdot\nu &= 0 \quad\ \ \text{on } \partial_N\Omega
\end{aligned}
\tag{2.2}
$$

and initial conditions

$$
u_\varepsilon(0,x) = \begin{cases} U_0(x), & x\in\Omega^+_\varepsilon,\\ U^M_0\left(\bar{x}, \frac{x_n}{\varepsilon}\right), & x\in\Omega^M_\varepsilon,\\ U_0(x), & x\in\Omega^-_\varepsilon. \end{cases}
\tag{2.3}
$$

Here the boundaries $\partial_D\Omega^\pm$, respectively, $\partial_N\Omega$, are defined as follows:

$$
\begin{aligned}
\partial_D\Omega^\pm &= \partial\Omega \cap \{x\in\mathbb{R}^n, x_n = \pm H\},\\
\partial_N\Omega &= \partial\Omega \setminus \{\partial_D\Omega^+ \cup \partial_D\Omega^-\}.
\end{aligned}
$$

On the interfaces S^+_ε and S^-_ε we require the natural transmission conditions, i.e., the continuity of the solutions and of the normal fluxes:

$$
\begin{aligned}
u^\pm_\varepsilon &= u^M_\varepsilon \qquad\qquad\qquad\quad \text{on } S^\pm_\varepsilon,\\
D^\pm_j \nabla u^\pm_{j\varepsilon}\cdot\nu &= \varepsilon D^M_j\left(\frac{x}{\varepsilon}\right)\nabla u^M_{j\varepsilon}\cdot\nu \quad \text{on } S^\pm_\varepsilon.
\end{aligned}
\tag{2.4}
$$

Assumptions on the data. For the diffusion coefficients $D_j : \Omega \to \mathbb{R}$, $j = 1, \ldots, m$, given by

$$
D_j(x) = \begin{cases} D^+_j, & x\in\Omega^+_\varepsilon,\\ D^M_j\left(\frac{x}{\varepsilon}\right), & x\in\Omega^M_\varepsilon,\\ D^-_j, & x\in\Omega^-_\varepsilon, \end{cases}
$$

we assume the following:

- $D^+_j > 0, D^-_j > 0,\ j = 1, \ldots, m$.
- D^M_j is defined on the standard cell Z and belongs to $C^1_{per}([0,1]^{n-1}, C^1([-1,1]))$. We also assume that it is strictly positive.

Concerning the reaction terms we suppose the following:

- $f = f(x,z) : \bar{\Omega}\times\mathbb{R}^m \to \mathbb{R}^m$ is continuous and Lipschitz continuous with respect to z, with a Lipschitz constant independent of x.
- $g = g(\bar{y}, y_n, z) : [0,1]^{n-1}\times[-1,1]\times\mathbb{R}^m \to \mathbb{R}^m$ is continuous, Lipschitz continuous in z, and periodic in $\bar{y} = (y_1, \ldots, y_{n-1})$.

The assumptions on the reaction terms imply that there exist positive constants c_1 and c_2 such that for $j = 1, \ldots, m$

$$
|f_j(x,z)| \le c_1(1+|z|) \quad \text{for all } z\in\mathbb{R}^m, x\in\bar{\Omega},
\tag{2.5}
$$

$$
|g_j(y,z)| \le c_2(1+|z|) \quad \text{for all } z\in\mathbb{R}^m, y\in Z.
\tag{2.6}
$$

Additionally, we have to impose on f and g structural conditions which guarantee L^∞-estimates of the solutions u_ε. A possible choice of such conditions is given in the following.

Let $M_j \in \mathbb{R}, M_j > 0$, $j = 1, \dots, m$, be given. We consider

$$f_j(\cdot, z) \le A_j z_j \quad \text{for } z_j \ge M_j, \tag{2.7}$$

$$g_j(\cdot, z) \le A_j z_j \quad \text{for } z_j \ge M_j, \tag{2.8}$$

where $A_j \in \mathbb{R}$, $A_j \ge 0, j = 1, \dots, m$. We also require that

$$\sum_{j=1}^m f_j(\cdot, z)(z_j)_- \le C \sum_{j=1}^m |(z_j)_-|^2, \tag{2.9}$$

$$\sum_{j=1}^m g_j(\cdot, z)(z_j)_- \le C \sum_{j=1}^m |(z_j)_-|^2, \tag{2.10}$$

where $(z_j)_- = \min\{z_j, 0\}$. For the initial functions we assume that $U_0 \in H^2(\Omega^+, \mathbb{R}^m) \cap H^2(\Omega^-, \mathbb{R}^m)$, $U_0^M \in H^2(\Sigma \times]-1, 1[, \mathbb{R}^m)$, such that

$$\frac{1}{\sqrt{\varepsilon}} \left\| U_0^M \left(\cdot, \frac{\cdot}{\varepsilon}\right) \right\|_{L^2(\Omega_\varepsilon^M, \mathbb{R}^m)} + \sqrt{\varepsilon} \left\| \nabla U_0^M \left(\cdot, \frac{\cdot}{\varepsilon}\right) \right\|_{L^2(\Omega_\varepsilon^M, \mathbb{R}^m)} \le C, \tag{2.11}$$

and that they satisfy the compatibility conditions

$$\begin{aligned} U_0(x) &= U_0^M \left(\bar{x}, \frac{x_n}{\varepsilon}\right) && \text{on } S_\varepsilon^\pm, \\ D_j^\pm \nabla U_{j0} \cdot \overrightarrow{\nu} &= \varepsilon D_j^M \left(\frac{x}{\varepsilon}\right) \nabla U_{j0}^M \cdot \overrightarrow{\nu} && \text{on } S_\varepsilon^\pm, \\ U_0(x) &= u_D(0, x) && \text{on } \partial_D \Omega^\pm. \end{aligned} \tag{2.12}$$

For the Dirichlet boundary data we require

$$u_D \in L^2((0,T), H^2(\Omega, \mathbb{R}^m)), \ \operatorname{supp}(u_D) \cap \Omega_\varepsilon^M = \emptyset, \tag{2.13}$$

$$\partial_t u_D^\pm \in L^2((0,T), H^1(\Omega, \mathbb{R}^m)) \cap L^\infty((0,T) \times \Omega, \mathbb{R}^m). \tag{2.14}$$

In order to obtain the L^∞-estimates for the solution u_ε, we have to assume that the initial and boundary functions also satisfy corresponding bounds. For the example of reaction terms given above, we assume that

$$0 \le U_{j0} \le M_j, \ 0 \le U_{j0}^M \le M_j, \ 0 \le u_{jD} \le M_j, \quad j = 1, \dots, m. \tag{2.15}$$

Variational formulation of the microscopic problem. We denote by X the function space

$$X = \{u \in H^1(\Omega, \mathbb{R}^m) : u = 0 \text{ on } \partial_D \Omega^+ \cup \partial_D \Omega^-\}.$$

The variational formulation of problem (2.1)–(2.4) is given as follows: Find $u_\varepsilon : (0,T) \times \Omega \to \mathbb{R}^m$, such that $u_\varepsilon - u_D \in L^2((0,T), X)$, $\partial_t(u_\varepsilon - u_D) \in L^2((0,T), L^2(\Omega))$, and for all $\varphi \in L^2((0,T), X)$ and a.e. $t \in (0,T)$ we have

$$\begin{aligned} &\int_{\Omega_\varepsilon^+} \partial_t u_{j\varepsilon}^+ \varphi_j \, dx + \int_{\Omega_\varepsilon^-} \partial_t u_{j\varepsilon}^- \varphi_j \, dx + \frac{1}{\varepsilon} \int_{\Omega_\varepsilon^M} \partial_t u_{j\varepsilon}^M \varphi_j \, dx \\ &+ D_j^+ \int_{\Omega_\varepsilon^+} \nabla u_{j\varepsilon}^+ \nabla \varphi_j \, dx + D_j^- \int_{\Omega_\varepsilon^-} \nabla u_{j\varepsilon}^- \nabla \varphi_j \, dx + \int_{\Omega_\varepsilon^M} \varepsilon D_j^M \nabla u_{j\varepsilon}^M \nabla \varphi_j \, dx \\ &= \int_{\Omega_\varepsilon^+} f_j(x, u_\varepsilon^+) \varphi_j \, dx + \int_{\Omega_\varepsilon^-} f_j(x, u_\varepsilon^-) \varphi_j \, dx + \frac{1}{\varepsilon} \int_{\Omega_\varepsilon^M} g_j \left(\frac{x}{\varepsilon}, u_\varepsilon^M\right) \varphi_j \, dx \end{aligned} \tag{2.16}$$

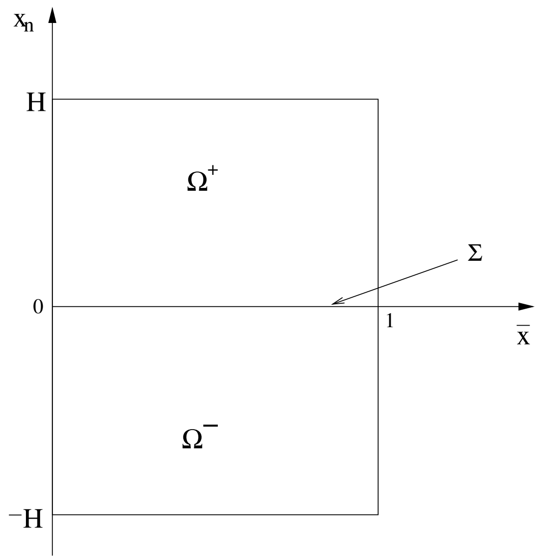


FIG. 2.3. *The structure of the domain* Ω *in the limit* $\varepsilon = 0$.

and

$$u_\varepsilon(0,x) = \begin{cases} U_0(x), & x \in \Omega_\varepsilon^+, \\ U_0^M\left(\bar{x}, \frac{x_n}{\varepsilon}\right), & x \in \Omega_\varepsilon^M, \\ U_0(x), & x \in \Omega_\varepsilon^-. \end{cases}$$

The existence and uniqueness of weak solutions for the problem (2.1)–(2.4) for every fixed $\varepsilon > 0$ is standard, e.g., by using the Galerkin method based on estimates similar to those in section 3.

Our aim is now to study the behavior of the solutions u_ε for small values of the parameter ε. We will do this by studying the asymptotic behavior of the sequence u_ε for $\varepsilon \to 0$.

When ε tends to zero, the thin layer Ω_ε^M approaches the interface Σ. The domains Ω_ε^+ and Ω_ε^- tend to the domains Ω^+ and Ω^-, respectively, defined below:

$$\Omega^+ =]0,1[^{n-1} \times]0,H[, \tag{2.17}$$

$$\Omega^- =]0,1[^{n-1} \times]-H,0[. \tag{2.18}$$

Thus the macroscopic limit of the sequence u_ε (if it exists) will be defined on the domain Ω consisting of (see Figure 2.3)

$$\Omega = \Omega^+ \cup \Omega^- \cup \Sigma.$$

2.2. Main results. From the a priori estimates (given in section 3), it is obvious that different convergence concepts have to be used for studying the asymptotic behavior of the solutions u_ε in the bulk and thin layer regions. Whereas in $\Omega_\varepsilon^\pm$ classical compactness results can be used, in Ω_ε^M compactness needs to be considered with respect to the weak and strong two-scale convergences adapted to the thin layer. The concepts of multiscale convergence in the weak and strong sense, also for thin and periodic structures, are crucial for formulating and proving the main results of this paper. For the definition and properties of two-scale convergence for thin heterogeneous layers, see section 4.

In the following two propositions, we state the convergence results in the bulk and thin layer regions as well as the convergence of the traces on the interfaces $S_\varepsilon^\pm$. For the layer region, we obtain in a first step weak two-scale convergence.

PROPOSITION 2.1. *There exists a subsequence denoted again u_ε and limit functions $u_0^\pm \in L^2((0,T), H^1(\Omega^\pm, \mathbb{R}^m))$, with $\partial_t u_0^\pm \in L^2((0,T), L^2(\Omega^\pm, \mathbb{R}^m))$, and $u_0^M \in L^2((0,T)\times\Sigma, H^1_{per}(Y, H^1(]-1,1[))^m$, with $\partial_t u_0^M \in L^2((0,T)\times\Sigma, L^2(Z))^m$, such that*

1. $\chi_{\Omega_\varepsilon^\pm} u_{j\varepsilon} \to u_{j0}^\pm$ *strongly in* $L^2((0,T), L^2(\Omega^\pm))$;
2. $\chi_{\Omega_\varepsilon^\pm} \nabla u_{j\varepsilon} \to \nabla u_{j0}^\pm$ *weakly in* $L^2((0,T), L^2(\Omega^\pm))$;
3. $\chi_{\Omega_\varepsilon^\pm} \partial_t u_{j\varepsilon} \to \partial_t u_{j0}^\pm$ *weakly in* $L^2((0,T), L^2(\Omega^\pm))$;
4. $u_{j\varepsilon}^M \overset{t.s.}{\to} u_{j0}^M(t,\bar{x},y)$ *weakly in the two-scale sense*;
5. $\varepsilon\,\nabla u_{j\varepsilon}^M \overset{t.s.}{\to} \nabla_y u_{j0}^M(t,\bar{x},y)$ *weakly in the two-scale sense*;
6. $\partial_t u_{j\varepsilon}^M \overset{t.s.}{\to} \partial_t u_{j0}^M(t,\bar{x},y)$ *weakly in the two-scale sense.*

Furthermore, we have

$$u_0^M(t,\bar{x},\bar{y},\pm 1) = u_0^\pm(t,\bar{x},0) \quad a.e.\ (t,\bar{x}) \in (0,T)\times\Sigma,\ \bar{y}\in Y. \tag{2.19}$$

PROPOSITION 2.2. *There exists a subsequence denoted again by $u_\varepsilon^\pm$, such that*

$$\tilde{u}_{j\varepsilon}^\pm \to u_{j0}^\pm \ \textit{strongly in}\, L^2((0,T)\times\Sigma), \tag{2.20}$$

$$\lim_{\varepsilon\to 0}\int_0^T \int_{S_\varepsilon^\pm} u_{j\varepsilon}^\pm(t,x)\varphi_j\left(t,\bar{x},\frac{x}{\varepsilon}\right)dxdt = \int_0^T\int_\Sigma\int_Y u_{j0}^\pm(t,\bar{x},0)\varphi_j(t,\bar{x},\bar{y},\pm 1)\,d\bar{y}\,d\bar{x}dt \tag{2.21}$$

for all $\varphi_j \in C^\infty([0,T]\times\Sigma, C^\infty_{per}(Y, C^\infty([-1,1])))$, where $u_0^\pm$ are the limit functions given in Proposition 2.1*, and the scaled functions $\tilde{u}_{j\varepsilon}^\pm$ are defined in* (5.2).

A central contribution of this paper is formulated in Theorem 2.3 below, where a strong two-scale convergence of the solutions u_ε in the layer Ω_ε^M is obtained. This strong two-scale convergence is not based on extension properties but uses the reaction-diffusion equations in the layer and the compactness of the traces on $S_\varepsilon^\pm$.

THEOREM 2.3. *The extension $\hat{u}_\varepsilon^M$ of u_ε^M to $\hat{\Omega}_\varepsilon^M$ defined in section* 5.4 *can be estimated in terms of its boundary values on $\hat{S}_\varepsilon^\pm$ and its initial values $\hat{U}_0^M$ in the following way: Fix $h \in \left(0,\frac{1}{4}\right)$ and assume $l \in \mathbb{Z}^{n-1}$ such that $|l\varepsilon| < h$. Then there exists a constant C independent of ε and l, such that*

$$\begin{aligned}
&\frac{1}{\sqrt{\varepsilon}}||\hat{u}_\varepsilon^M(t,x+(l,0)\varepsilon) - \hat{u}_\varepsilon^M(t,x)||_{L^2((0,T)\times\Omega_\varepsilon^M,\mathbb{R}^m)}\\
&\le C\Bigg(||\hat{u}_\varepsilon^+(t,x+(l,0)\varepsilon) - \hat{u}_\varepsilon^+(t,x)||_{L^2((0,T)\times\hat{S}_\varepsilon^+,\mathbb{R}^m)}\\
&\quad + ||\hat{u}_\varepsilon^-(t,x+(l,0)\varepsilon) - \hat{u}_\varepsilon^-(t,x)||_{L^2((0,T)\times\hat{S}_\varepsilon^-,\mathbb{R}^m)}\\
&\quad + \frac{1}{\sqrt{\varepsilon}}\left\|\hat{U}_0^M\left(\bar{x}+l\varepsilon,\frac{x_n}{\varepsilon}\right) - \hat{U}_0^M\left(\bar{x},\frac{x_n}{\varepsilon}\right)\right\|_{L^2(\hat{\Omega}_\varepsilon^M,\mathbb{R}^m)}\Bigg)\\
&\quad + C\frac{\varepsilon}{h} + Ch^{\frac{1}{2}}.
\end{aligned} \tag{2.22}$$

Thus, up to subsequence, u_ε^M converges also strongly in the two-scale sense to the limit function u_0^M.

The limit functions satisfy the macroscopic problem formulated in the following theorem.

THEOREM 2.4. *The limit functions $u_0^\pm$ given in Proposition 2.1 satisfy in a distributional sense the initial boundary value problem on Ω^+ and Ω^-, respectively,*

$$\partial_t u_{j0}^+ - D_j^+ \Delta u_{j0}^+ = f_j(x, u_0^+), \qquad (t,x) \in (0,T) \times \Omega^+, \tag{2.23}$$

$$\partial_t u_{j0}^- - D_j^- \Delta u_{j0}^- = f_j(x, u_0^-), \qquad (t,x) \in (0,T) \times \Omega^-, \tag{2.24}$$

$$u_0^\pm(t,x) = u_D^\pm, \qquad (t,x) \in (0,T) \times \partial_D\Omega^\pm, \tag{2.25}$$

$$\frac{\partial u_0^\pm}{\partial \nu} = 0, \qquad (t,x) \in (0,T) \times \partial_N\Omega^\pm, \tag{2.26}$$

$$u_0^\pm(0,x) = U_0(x), \qquad x \in \Omega^\pm, \tag{2.27}$$

together with the effective transmission conditions on the interface Σ,

$$\begin{aligned} [u_{j0}]_\Sigma(t,\bar{x}) = &\int_Z (g_j(y, u_0^M(t,\bar{x},y)) - \partial_t u_{j0}^M(t,\bar{x},y))\eta_j(y)\, dy \\ &+ D_j^+ \eta_j^+ \partial_n u_{j0}^+(t,\bar{x},0) - D_j^- \eta_j^- \partial_n u_{j0}^-(t,\bar{x},0), \end{aligned} \tag{2.28}$$

$$\begin{aligned} &(D_j^+ \partial_n u_{j0}^+ - D_j^- \partial_n u_{j0}^-)(t,\bar{x},0) \\ &\qquad = \int_Z (\partial_t u_{j0}^M(t,\bar{x},y) - g_j(y, u_0^M(t,\bar{x},y)))\, dy. \end{aligned} \tag{2.29}$$

The limit function u_0^M, which enters the transmission conditions, is the weak solution of the local problem

$$\begin{aligned} &\partial_t u_{j0}^M(t,\bar{x},y) - \nabla_y(D_j^M(y)\nabla_y u_{j0}^M(t,\bar{x},y)) \\ &\qquad = g_j(y, u_0^M(t,\bar{x},y)) \text{ in } (0,T)\times Z, \end{aligned} \tag{2.30}$$

$$u_0^M(t,\bar{x},y) = u_0^\pm(t,\bar{x},0) \text{ on } (0,T) \times S^\pm, \tag{2.31}$$

$$u_0^M \text{ is periodic in } Y, \tag{2.32}$$

$$u_0^M(0,\bar{x},y) = U_0^M(\bar{x}, y_n) \text{ in } Z \tag{2.33}$$

for a.e. $\bar{x} \in \Sigma$. The constants $\eta^\pm = (\eta_1^\pm, \dots, \eta_m^\pm)$ are the constant boundary values on $S^\pm$ of the solution $\eta = (\eta_1, \dots, \eta_m)$ to the following boundary value problem on the standard cell: Find

$$\eta = (\eta_1, \dots, \eta_m) \in V := \{\varphi \in H^1(Z, \mathbb{R}^m), \varphi \text{ periodic in } Y, \varphi \equiv \text{const on } S^+ \cup S^-\}$$

such that $\frac{1}{|Z|}\int_Z \eta(y)\, dy = 0$ and, for all $\varphi \in V$,

$$\int_Z D_j^M(y)\nabla\eta_j(y)\nabla\varphi_j(y)\, dy = \int_{S^+} \varphi_j(y)\, ds - \int_{S^-} \varphi_j(y)\, ds. \tag{2.34}$$

In the final step, we prove uniqueness for the macroscopic problem, and therefore we obtain that the sequence of solutions to the microscopic problems converges to the solution of the macroscopic problem in the corresponding topology on every subdomain.

THEOREM 2.5. *The solution (u_0^+, u_0^-, u_0^M) of the macroscopic system* (2.23)–(2.33) *is unique.*

3. A priori estimates for the microscopic model. In order to get some information about the compactness properties of the sequence u_ε, we have to control the dependence of the solutions on the scale parameter ε.

LEMMA 3.1. *For the solutions of problem* (2.1)–(2.4)*, the following estimates hold, with a generic constant C independent of ε:*

$$||u_{j\varepsilon}^{\pm}||_{L^\infty((0,T),H^1(\Omega_\varepsilon^{\pm}))} \le C, \tag{3.1}$$

$$\frac{1}{\sqrt{\varepsilon}}||u_{j\varepsilon}^{M}||_{L^\infty((0,T),L^2(\Omega_\varepsilon^{M}))} + \sqrt{\varepsilon}||\nabla u_{j\varepsilon}^{M}||_{L^\infty((0,T),L^2(\Omega_\varepsilon^{M}))} \le C, \tag{3.2}$$

$$||\partial_t\, u_{j\varepsilon}^{\pm}||_{L^2((0,T),L^2(\Omega_\varepsilon^{\pm}))} \le C, \quad \frac{1}{\sqrt{\varepsilon}}||\partial_t u_{j\varepsilon}^{M}||_{L^2((0,T),L^2(\Omega_\varepsilon^{M}))} \le C. \tag{3.3}$$

In the case that the reaction terms satisfy the additional structural conditions (2.7)–(2.9)*, we obtain the following pointwise bounds for the solution:*

$$0 \le u_{j\varepsilon} \le M_j e^{A_j t} \quad \textit{for a.e.}\,(t,x) \in (0,T)\times\Omega. \tag{3.4}$$

Proof. Let us first set

$$U_\varepsilon = u_\varepsilon - u_D = \begin{cases} u_\varepsilon^{\pm} - u_D, & x \in \Omega_\varepsilon^{\pm}, \\ u_\varepsilon^{M}, & x \in \Omega_\varepsilon^{M}, \end{cases}$$

and use it as a test function in (2.16). We obtain

$$\begin{aligned}
&\int_{\Omega_\varepsilon^{\pm}} \partial_t U_{j\varepsilon}^{\pm}(t) U_{j\varepsilon}^{\pm}(t)\,dx \;+\; \int_{\Omega_\varepsilon^{\pm}} \partial_t u_{jD}(t) U_{j\varepsilon}^{\pm}(t)\,dx \\
&+ D_j^{\pm}\int_{\Omega_\varepsilon^{\pm}} \nabla U_{j\varepsilon}^{\pm}(t)\nabla U_{j\varepsilon}^{\pm}(t)\,dx \;+\; D_j^{\pm}\int_{\Omega_\varepsilon^{\pm}} \nabla u_{jD}(t)\nabla U_{j\varepsilon}^{\pm}(t)\,dx \\
&+ \frac{1}{\varepsilon}\int_{\Omega_\varepsilon^{M}} \partial_t u_{j\varepsilon}^{M}(t) u_{j\varepsilon}^{M}(t)\,dx + \varepsilon\int_{\Omega_\varepsilon^{M}} D_j^{M}\left(\frac{x}{\varepsilon}\right)\nabla u_{j\varepsilon}^{M}(t)\nabla u_{j\varepsilon}^{M}(t)\,dx \\
&= \int_{\Omega_\varepsilon^{\pm}} f_j(x, u_\varepsilon^{\pm}(t,x)) U_{j\varepsilon}^{\pm}(t)dx \;+\; \frac{1}{\varepsilon}\int_{\Omega_\varepsilon^{M}} g_j\left(\frac{x}{\varepsilon}, u_\varepsilon^{M}(t,x)\right) u_{j\varepsilon}^{M}(t)dx \\
&\le C_1\int_{\Omega_\varepsilon^{\pm}} (1+|u_\varepsilon^{\pm}|)\,|U_{j\varepsilon}^{\pm}(t)|\,dx + \frac{C_2}{\varepsilon}\int_{\Omega_\varepsilon^{M}} (1+|u_\varepsilon^{M}|)\,|u_{j\varepsilon}^{M}(t)|\,dx
\end{aligned} \tag{3.5}$$

a.e. in $(0,T)$. Here the integrals over $\Omega_\varepsilon^{\pm}$ stand for the sum of the integrals over Ω_ε^{+} and Ω_ε^{-}. For the last inequality we used the growth conditions (2.5) and (2.6) on the reaction terms. Adding up the estimates in (3.5) for $j = 1, \dots, m$ and integrating with respect to time yields

$$\begin{aligned}
&\frac{1}{2}||U_\varepsilon^{\pm}(t)||^2_{L^2(\Omega_\varepsilon^{\pm},\mathbb{R}^m)} + \frac{1}{2\varepsilon}||u_\varepsilon^{M}(t)||^2_{L^2(\Omega_\varepsilon^{M},\mathbb{R}^m)} \\
&+ \int_0^t\int_{\Omega_\varepsilon^{\pm}} |\nabla U_\varepsilon^{\pm}|^2 dxdt + \varepsilon\int_0^t\int_{\Omega_\varepsilon^{M}} |\nabla u_\varepsilon^{M}(t)|^2 dxdt \\
&\le C\left(\int_0^t\int_{\Omega_\varepsilon^{\pm}} |\partial_t u_D(t) U_\varepsilon^{\pm}| dxdt + \int_0^t\int_{\Omega_\varepsilon^{\pm}} |\nabla u_D \nabla U_\varepsilon^{\pm}|\,dxdt\right) \\
&\quad + C\left(1 + \int_0^t\int_{\Omega_\varepsilon^{\pm}} |U_\varepsilon^{\pm}|^2 dxdt + \frac{1}{\varepsilon}\int_0^t\int_{\Omega_\varepsilon^{M}} |u_\varepsilon^{M}|^2 dxdt\right) \\
&\quad + \frac{1}{2}||U_\varepsilon^{\pm}(0)||^2_{L^2(\Omega_\varepsilon^{\pm},\mathbb{R}^m)} + \frac{1}{2\varepsilon}||u_\varepsilon^{M}(0)||^2_{L^2(\Omega_\varepsilon^{M},\mathbb{R}^m)}.
\end{aligned} \tag{3.6}$$

Here we also used the regularity properties (2.14) of the boundary data u_D. To estimate the second term on the right-hand side, we make use of the inequality

$$2ab \leq \delta a^2 + \frac{1}{\delta} b^2$$

to get

$$\int_0^t \int_{\Omega_\varepsilon^\pm} |\nabla u_D \nabla U_\varepsilon^\pm| \, dx dt \leq C \left(\frac{1}{\delta} + \delta ||\nabla U_\varepsilon^\pm||^2_{L^2((0,T)\times\Omega_\varepsilon^\pm, \mathbb{R}^m)} \right).$$

If δ is small enough, the term involving $\nabla U_\varepsilon^\pm$ can be absorbed on the left-hand side. Now, using Gronwall's lemma and the assumptions (2.11) on the initial conditions, we obtain

$$||U_\varepsilon^\pm||_{L^\infty((0,T),L^2(\Omega_\varepsilon^\pm,\mathbb{R}^m))} + \frac{1}{\sqrt{\varepsilon}} ||u_\varepsilon^M||_{L^\infty((0,T),L^2(\Omega_\varepsilon^M,\mathbb{R}^m))} \leq C$$

and

$$||\nabla U_\varepsilon^\pm(t)||_{L^2((0,T),L^2(\Omega_\varepsilon^\pm,\mathbb{R}^m))} + \sqrt{\varepsilon} ||\nabla u_\varepsilon^M(t)||_{L^2((0,T),L^2(\Omega_\varepsilon^M,\mathbb{R}^m))} \leq C.$$

To obtain the L^∞-estimates with respect to time for the gradients and the estimates for the time derivatives, we take $\varphi = \partial_t U_\varepsilon$ as a test function in (2.16). It yields

$$\begin{aligned} &\int_{\Omega_\varepsilon^\pm} \partial_t U_{j\varepsilon}^\pm(t) \partial_t U_{j\varepsilon}^\pm(t) \, dx \; + \int_{\Omega_\varepsilon^\pm} \partial_t u_{jD}(t) \partial_t U_{j\varepsilon}^\pm(t) \, dx \\ &+ D_j^\pm \int_{\Omega_\varepsilon^\pm} \nabla U_{j\varepsilon}^\pm(t) \nabla \partial_t U_{j\varepsilon}^\pm(t) dx + D_j^\pm \int_{\Omega_\varepsilon^\pm} \nabla u_{jD}(t) \nabla \partial_t U_{j\varepsilon}^\pm(t) \, dx \\ &+ \frac{1}{\varepsilon} \int_{\Omega_\varepsilon^M} \partial_t u_{j\varepsilon}^M(t) \partial_t u_{j\varepsilon}^M(t) \, dx + \varepsilon \int_{\Omega_\varepsilon^M} D_j^M \left(\frac{x}{\varepsilon}\right) \nabla u_{j\varepsilon}^M(t) \nabla \partial_t u_{j\varepsilon}^M(t) \, dx \\ &= \int_{\Omega_\varepsilon^\pm} f_j(x, u_\varepsilon^\pm) \partial_t U_{j\varepsilon}^\pm(t) dx \; + \frac{1}{\varepsilon} \int_{\Omega_\varepsilon^M} g_j \left(\frac{x}{\varepsilon}, u_\varepsilon^M\right) \partial_t u_{j\varepsilon}^M(t) dx \end{aligned}$$

a.e. on $(0,T)$. First, we have to transform the energy integral on Ω_ε^M as follows:

$$\begin{aligned} \varepsilon \int_{\Omega_\varepsilon^M} D_j^M \left(\frac{x}{\varepsilon}\right) \nabla u_{j\varepsilon}^M(t) \nabla \partial_t u_{j\varepsilon}^M(t) \, dx &= \frac{d}{dt} \int_{\Omega_\varepsilon^M} \varepsilon D_j^M \left(\frac{x}{\varepsilon}\right) \nabla u_{j\varepsilon}^M(t) \nabla u_{j\varepsilon}^M(t) \, dx \\ - \varepsilon \int_{\Omega_\varepsilon^M} D_j^M \left(\frac{x}{\varepsilon}\right) \nabla \partial_t u_{j\varepsilon}^M(t) \nabla u_{j\varepsilon}^M(t) \, dx. \end{aligned}$$

Thus,

(3.7)
$$\varepsilon \int_{\Omega_\varepsilon^M} D_j^M \left(\frac{x}{\varepsilon}\right) \nabla u_{j\varepsilon}^M(t) \nabla \partial_t u_{j\varepsilon}^M(t) \, dx = \frac{1}{2} \frac{d}{dt} \int_{\Omega_\varepsilon^M} \varepsilon D_j^M \left(\frac{x}{\varepsilon}\right) \nabla u_{j\varepsilon}^M(t) \nabla u_{j\varepsilon}^M(t) \, dx.$$

Adding up the equations for $j = 1, \ldots, m$, taking into account (3.7) and the growth conditions (2.5), (2.6), and integrating with respect to time, we obtain

$$\begin{aligned} &||\partial_t U_\varepsilon^\pm||^2_{L^2((0,T),L^2(\Omega_\varepsilon^\pm,\mathbb{R}^m))} + \frac{1}{\varepsilon} ||\partial_t u_\varepsilon^M||^2_{L^2((0,T),L^2(\Omega_\varepsilon^M,\mathbb{R}^m))} \\ &+ ||\nabla U_\varepsilon^\pm(t)||^2_{L^2(\Omega_\varepsilon^\pm,\mathbb{R}^m)} + \varepsilon ||\nabla u_\varepsilon^M(t)||^2_{L^2(\Omega_\varepsilon^M,\mathbb{R}^m)} \\ &\leq C \left(1 + ||U_\varepsilon^\pm||^2_{L^2((0,T),L^2(\Omega_\varepsilon^\pm,\mathbb{R}^m))} + \frac{1}{\varepsilon} ||u_\varepsilon^M||^2_{L^2((0,T),L^2(\Omega_\varepsilon^M,\mathbb{R}^m))} \right) \\ &+ C ||\nabla U_\varepsilon^\pm||^2_{L^2((0,T),L^2(\Omega_\varepsilon^\pm,\mathbb{R}^m))}. \end{aligned} \tag{3.8}$$

Using the estimates obtained in the first part of the proof, it follows that the right-hand side of (3.8) is bounded independently of ε. Thus estimates (3.1)–(3.3) are proved.

Now, it remains to show the L^∞-bounds for the solution under the hypothesis (2.7)–(2.10) on the reaction terms. We first show positivity of the solutions. Let us test our system (2.16) with the test function φ given by

$$\varphi_j = (u_{j\varepsilon})_- = \min\{u_{j\varepsilon}, 0\} \quad \text{a.e. on } [0,T]\times\Omega.$$

Due to the the assumptions (2.15), our test function has zero boundary values on the parabolic boundary. Thus we obtain

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\Omega_\varepsilon^+} |(u_{j\varepsilon}^+)_-|^2\,dx + \frac{1}{2}\frac{d}{dt}\int_{\Omega_\varepsilon^-} |(u_{j\varepsilon}^-)_-|^2\,dx + \frac{1}{2\varepsilon}\frac{d}{dt}\int_{\Omega_\varepsilon^M} |(u_{j\varepsilon}^M)_+|^2\,dx \\
&+ D_j^+\int_{\Omega_\varepsilon^+} |\nabla(u_{j\varepsilon}^+)_-|^2\,dx + D_j^-\int_{\Omega_\varepsilon^-} |\nabla(u_{j\varepsilon}^-)_-|^2\,dx + \int_{\Omega_\varepsilon^M} \varepsilon D_j^M |\nabla(u_{j\varepsilon}^M)_-|^2\,dx \\
&= \int_{\Omega_\varepsilon^+} f_j(x,u_\varepsilon^+)(u_{j\varepsilon}^+)_- + \int_{\Omega_\varepsilon^-} f_j(x,u_\varepsilon^-)(u_{j\varepsilon}^-)_- + \frac{1}{\varepsilon}\int_{\Omega_\varepsilon^M} g_j\left(\frac{x}{\varepsilon}, u_\varepsilon^M\right)(u_{j\varepsilon}^M)_-.
\end{aligned}$$

Integrating with respect to time, adding up the equations for $j = 1,\dots,m$, and using the assumptions (2.9), (2.10) on the reaction terms leads to

$$\begin{aligned}
&\int_{\Omega_\varepsilon^+}\sum_{j=1}^m |(u_{j\varepsilon}^+)_-(t)|^2 dx + \int_{\Omega_\varepsilon^-}\sum_{j=1}^m |(u_{j\varepsilon}^-)_-(t)|^2 dx + \frac{1}{\varepsilon}\int_{\Omega_\varepsilon^M}\sum_{j=1}^m |(u_{j\varepsilon}^M)_-(t)|^2 dx \\
&\le C\int_0^t \left\{ \int_{\Omega_\varepsilon^+}\sum_{j=1}^m |(u_{j\varepsilon}^+)_-|^2 dx + \int_{\Omega_\varepsilon^-}\sum_{j=1}^m |(u_{j\varepsilon}^-)_-|^2 dx + \frac{1}{\varepsilon}\int_{\Omega_\varepsilon^M}\sum_{j=1}^m |(u_{j\varepsilon}^M)_-|^2 dx \right\} dt.
\end{aligned}$$

Now, Gronwall's inequality implies that $(u_{j\varepsilon})_- = 0$. Thus the positivity of the solution is proved. To obtain the upper bound, we first test (2.16) with the test function

$$\varphi_j(t,x) = e^{-A_j t}\psi_j(t,x),$$

where $\psi \in L^2((0,T),H^1(\Omega))$ and has zero boundary values on the parabolic boundary. We obtain

$$\begin{aligned}
&\int_{\Omega_\varepsilon^+} \partial_t u_{j\varepsilon}^+ e^{-A_j t}\psi_j dx + \int_{\Omega_\varepsilon^-} \partial_t u_{j\varepsilon}^- e^{-A_j t}\psi_j dx + \frac{1}{\varepsilon}\int_{\Omega_\varepsilon^M} \partial_t u_{j\varepsilon}^M e^{-A_j t}\psi_j dx \\
&+ D_j^+\int_{\Omega_\varepsilon^+} e^{-A_j t}\nabla u_{j\varepsilon}^+ \nabla\psi_j dx + D_j^-\int_{\Omega_\varepsilon^-} e^{-A_j t}\nabla u_{j\varepsilon}^- \nabla\psi_j dx \\
&+ \int_{\Omega_\varepsilon^M} \varepsilon D_j^M e^{-A_j t}\nabla u_{j\varepsilon}^M \nabla\psi_j dx = \int_{\Omega_\varepsilon^+} f_j(x,u_\varepsilon^+) e^{-A_j t}\psi_j dx \\
&+ \int_{\Omega_\varepsilon^-} f_j(x,u_\varepsilon^-) e^{-A_j t}\psi_j dx + \frac{1}{\varepsilon}\int_{\Omega_\varepsilon^M} g_j\left(\frac{x}{\varepsilon}, u_\varepsilon^M\right) e^{-A_j t}\psi_j dx.
\end{aligned}$$

Now we intend to set

$$\psi_j = (e^{-A_j t}u_{j\varepsilon} - M_j)_+ = \max\{e^{-A_j t}u_{j\varepsilon} - M_j, 0\} \quad \text{a.e. on } [0,T]\times\Omega.$$

Therefore, we write the terms containing the time derivative as

$$\int_{\Omega_\varepsilon^+} \partial_t u_{j\varepsilon}^+ e^{-A_j t}\psi_j dx = \int_{\Omega_\varepsilon^+} \partial_t(e^{-A_j t}u_{j\varepsilon}^+ - M_j)\psi_j dx + \int_{\Omega_\varepsilon^+} A_j e^{-A_j t}u_{j\varepsilon}^+\psi_j dx,$$

and analogously the terms on Ω_ε^- and Ω_ε^M. We obtain

$$
\begin{aligned}
(3.9)\quad & \frac{1}{2}\frac{d}{dt}\int_{\Omega_\varepsilon^+} |(e^{-A_jt}u_{j\varepsilon}^+ - M_j)_+|^2\,dx + \frac{1}{2}\frac{d}{dt}\int_{\Omega_\varepsilon^-} |(e^{-A_jt}u_{j\varepsilon}^- - M_j)_+|^2\,dx \\
& + \frac{1}{2\varepsilon}\frac{d}{dt}\int_{\Omega_\varepsilon^M} |(e^{-A_jt}u_{j\varepsilon}^M - M_j)_+|^2\,dx + \int_{\Omega_\varepsilon^+} A_j e^{-A_jt}u_{j\varepsilon}^+(e^{-A_jt}u_{j\varepsilon}^+ - M_j)_+ dx \\
& + \int_{\Omega_\varepsilon^-} A_j e^{-A_jt}u_{j\varepsilon}^-(e^{-A_jt}u_{j\varepsilon}^- - M_j)_+ dx + \frac{1}{\varepsilon}\int_{\Omega_\varepsilon^M} A_j e^{-A_jt}u_{j\varepsilon}^M(e^{-A_jt}u_{j\varepsilon}^M - M_j)_+ dx \\
\le & \int_{\Omega_\varepsilon^+} f_j(x,u_\varepsilon^+)e^{-A_jt}(e^{-A_jt}u_{j\varepsilon}^+ - M_j)_+ dx + \int_{\Omega_\varepsilon^-} f_j(x,u_\varepsilon^-)e^{-A_jt}(e^{-A_jt}u_{j\varepsilon}^- - M_j)_+ dx \\
& + \frac{1}{\varepsilon}\int_{\Omega_\varepsilon^M} g_j\left(\frac{x}{\varepsilon},u_\varepsilon^M\right)e^{-A_jt}(e^{-A_jt}u_{j\varepsilon}^M - M_j)_+ dx.
\end{aligned}
$$

Now, due to assumptions (2.7), (2.8) on the reaction terms, the right-hand side in the above inequality can be estimated from above by

$$
\begin{aligned}
& \int_{\Omega_\varepsilon^+} A_j u_{j\varepsilon}^+ e^{-A_jt}(e^{-A_jt}u_{j\varepsilon}^+ - M_j)_+ dx + \int_{\Omega_\varepsilon^-} A_j u_{j\varepsilon}^- e^{-A_jt}(e^{-A_jt}u_{j\varepsilon}^- - M_j)_+ dx \\
+ & \frac{1}{\varepsilon}\int_{\Omega_\varepsilon^M} A_j u_{j\varepsilon}^M e^{-A_jt}(e^{-A_jt}u_{j\varepsilon}^M - M_j)_+ dx.
\end{aligned}
$$

However, these terms cancel with the corresponding terms on the left-hand side in (3.9), and thus, after integration with respect to time, we get

$$
\begin{aligned}
& \int_{\Omega_\varepsilon^+} |(e^{-A_jt}u_{j\varepsilon}^+ - M_j)_+|^2\,dx + \int_{\Omega_\varepsilon^-} |(e^{-A_jt}u_{j\varepsilon}^- - M_j)_+|^2\,dx \\
+ & \frac{1}{\varepsilon}\int_{\Omega_\varepsilon^M} |(e^{-A_jt}u_{j\varepsilon}^M - M_j)_+|^2\,dx \le 0.
\end{aligned}
$$

Finally, we have

$$e^{-A_jt}u_{j\varepsilon} - M_j \le 0 \quad \text{a.e. on } [0,T]\times\Omega.$$

This completes the proof. □

4. Two-scale convergence for thin heterogeneous layers. From the a priori estimates in Lemma 3.1, we see that on the subdomain Ω_ε^M we cannot use the classic compactness results for passing to the limit when $\varepsilon \to 0$. Here, we have to consider special convergence concepts which are adapted to sequences of functions varying on different scales.

The so-called two-scale convergence was introduced in [1] and [12] in order to handle two-scale phenomena with periodic structure in all space dimensions. Then it was extended to multiple scales [2]; periodic surfaces [13] and measures [17], [4]; thin domains [11]; and stochastic media [6]. For our problem, we need to generalize the concept of two-scale convergence to thin heterogeneous domains. Thus let $G \subset \mathbb{R}^{n-1}$ be a bounded domain and let $Y = [0,1]^{n-1}$ be the closed unit cube in $\mathbb{R}^{n-1}$. Let G_ε be a thin domain defined by

$$G_\varepsilon = G \times]-\varepsilon, \varepsilon[.$$

Let Σ be the interface

$$\Sigma = G \times \{0\}$$

and, as before, let us denote by Z the standard cell

$$Z = Y \times [-1, 1].$$

Let $C_{per}(Y)$ be the space of continuous functions in $\mathbb{R}^{n-1}$ which are periodic of period Y. Let $L^2_{per}(Y)$ (respectively, $H^1_{per}(Y)$) be the completion of $C_{per}(Y)$ in the norm of $L^2(Y)$ (respectively, $H^1(Y)$).

DEFINITION 4.1. *A sequence of functions $u_\varepsilon \in L^2((0,T)\times G_\varepsilon)$ is said to two-scale converge weakly to $u_0(t,\bar{x},y)$ belonging to $L^2((0,T)\times\Sigma\times Z)$ if, for any*

$$\psi(t,\bar{x},\bar{y},y_n) \in C\left([0,T]\times\bar{\Sigma}, C_{per}([0,1]^{n-1}, C([-1,1]))\right),$$

we have

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T\int_{G_\varepsilon} u_\varepsilon(t,x)\psi\left(t,\bar{x},\frac{x}{\varepsilon}\right)dx\,dt = \int_0^T\int_\Sigma\int_Z u_0(t,\bar{x},y)\psi(t,\bar{x},y)dy\,d\bar{x}\,dt. \tag{4.1}$$

A sequence $u_\varepsilon \in L^2((0,T)\times G_\varepsilon)$ which converges weakly to $u_0 \in L^2((0,T)\times\Sigma\times Z)$ is said to converge strongly in the two-scale sense to the limit u_0 if

$$\lim_{\varepsilon\to 0}\frac{1}{\sqrt{\varepsilon}}||u_\varepsilon||_{L^2((0,T)\times G_\varepsilon)} = ||u_0||_{L^2((0,T)\times\Sigma\times Z)}. \tag{4.2}$$

REMARK 1. *If u_0 in* (4.2) *has the property that $u_0(\cdot,\cdot,\frac{\cdot}{\varepsilon}) \in L^2((0,T)\times G_\varepsilon)$, then the relation* (4.2) *is equivalent to*

$$\lim_{\varepsilon\to 0}\frac{1}{\sqrt{\varepsilon}}\left\|u_\varepsilon(t,x) - u_0\left(t,\bar{x},\frac{x}{\varepsilon}\right)\right\|_{L^2((0,T)\times G_\varepsilon)} = 0.$$

A sufficient condition for $u_0(\cdot,\cdot,\frac{\cdot}{\varepsilon})$ to be in $L^2((0,T)\times G_\varepsilon)$ is, e.g., that $u_0 \in L^2((0,T)\times\Sigma, C_{per}(Y, C([-1,1])))$. For more details concerning this topic, see [1, Remark 1.10].

The main compactness result obtained for standard two-scale convergence in [12] and [1] can be generalized for the case of sequences defined on thin domains with microstructure.

PROPOSITION 4.2. *Let u_ε be a sequence in $L^2((0,T)\times G_\varepsilon)$, such that*

$$\frac{1}{\sqrt{\varepsilon}}||u_\varepsilon||_{L^2((0,T)\times G_\varepsilon)} \le C$$

with a positive constant C, independent of ε. Then there exists a subsequence (which we still denote by ε) and a limit function $u_0 \in L^2((0,T)\times\Sigma\times Z)$, such that

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T\int_{G_\varepsilon} u_\varepsilon(t,x)\varphi\left(t,\bar{x},\frac{x}{\varepsilon}\right)dx\,dt = \int_0^T\int_\Sigma\int_Z u_0(t,\bar{x},y)\varphi(t,\bar{x},y)dy\,d\bar{x}\,dt \tag{4.3}$$

for every test function $\varphi \in C([0,T]\times\bar{\Sigma}, C_{per}([0,1]^{n-1}, C([-1,1])))$.

Proof. Using Lemma 4.3 below, this result can be proved analogously to Theorem 1.2 in [1]. However, we have to take into account new aspects like time dependence and domains shrinking to a hypersurface.

Let us first consider functions u_ε, which do not vary with respect to time. Let $\varphi \in C\left(\bar{\Sigma}, C_{per}(Y, C([-1,1]))\right)$ and define

$$\mu_\varepsilon(\varphi) := \frac{1}{\varepsilon}\int_{G_\varepsilon} u_\varepsilon(x)\varphi\left(\bar{x}, \frac{x}{\varepsilon}\right) dx.$$

Since

$$|\mu_\varepsilon(\varphi)| \leq \frac{1}{\sqrt{\varepsilon}}||u_\varepsilon||_{L^2(G_\varepsilon)} \cdot \left\{\frac{1}{\varepsilon}\int_{G_\varepsilon}\left|\varphi\left(\bar{x}, \frac{x}{\varepsilon}\right)\right|^2 dx\right\}^{\frac{1}{2}} \leq C||\varphi||_B,$$

μ_ε is a bounded sequence of functionals on $B = C(\bar{\Sigma}, C_{per}(Y, C([-1,1])))$. Since this space is a separable Banach space, one can extract a subsequence of μ_ε (denoted μ_ε again) which weak*-converges to a limit functional $\mu_0 \in B$. Using now the boundedness of u_ε and Lemma 4.3, we obtain for every $\varphi \in B$

$$|\mu_0(\varphi)|^2 = \lim_{\varepsilon\to 0}|\mu_\varepsilon(\varphi)|^2 \leq C\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_{G_\varepsilon}\left|\varphi\left(\bar{x}, \frac{x}{\varepsilon}\right)\right|^2 dx = C||\varphi||^2_{L^2(\Sigma\times Z)}.$$

From the density of B in $L^2(\Sigma\times Z)$, it follows that μ_0 is a bounded functional on the Hilbert space $L^2(\Sigma\times Z)$. Thus the Riesz representation theorem implies the existence of a function $u_0 \in L^2(\Sigma\times Z)$ such that (4.3) is satisfied.

The proof of the theorem for the case of time-dependent functions can be reduced to the previous one by considering functions defined on the spatial domain with values in the separable Banach space $L^2((0,T))$. $\square$

Lemma 4.3. *Let*

$$B = C\left([0,T]\times\bar{\Sigma}, C_{per}(Y, C([-1,1]))\right)$$

be the space of continuous functions on $[0,T]\times\bar{\Sigma}$ *with values in the space* $C_{per}(Y, C([-1,1]))$ *of continuous functions on* Z *and* Y *periodic.* B *is a separable Banach space, which is dense in* $L^p((0,T)\times\Sigma\times Z)$ *for* $1 \leq p < \infty$, *and for every* $\varphi \in B$ *the following assertions hold:*

$$\frac{1}{\varepsilon}\int_0^T\int_{G_\varepsilon}\left|\varphi\left(t,\bar{x},\frac{x}{\varepsilon}\right)\right|^p dx\,dt \leq C||\varphi||^p_B \tag{4.4}$$

and

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T\int_{G_\varepsilon}\left|\varphi\left(t,\bar{x},\frac{x}{\varepsilon}\right)\right|^p dx\,dt = \int_0^T\int_\Sigma\int_Z|\varphi(t,\bar{x},y)|^p\ dy\,d\bar{x}\,dt. \tag{4.5}$$

The proof of (4.4) is obvious. To prove (4.5), we consider a paving of Σ with ε-cells and approximate φ by step functions with respect to the variable $\bar{x}\in\Sigma$. Using then the periodicity of φ with respect to the variable $\bar{y}\in Y$ and taking the limit for $\varepsilon\to 0$, the assertion follows.

Next we investigate the situation where we also have bounds on the gradients.

Proposition 4.4.

(i) *Let* u_ε *be a sequence of functions in* $L^2((0,T),H^1(G_\varepsilon))$, *such that*

$$\frac{1}{\sqrt{\varepsilon}}||u_\varepsilon||_{L^2((0,T)\times G_\varepsilon)} + \frac{1}{\sqrt{\varepsilon}}||\nabla u_\varepsilon||_{L^2((0,T)\times G_\varepsilon)} \leq C.$$

Then there exist functions $u_0 \in L^2((0,T), H^1(\Sigma))$ *and* $u_1 \in L^2((0,T) \times \Sigma, H^1_{per}(Y, H^1[-1,1])/\mathbb{R})$*, such that*

$$u_\varepsilon \xrightarrow{t.s.} u_0(t,\bar{x}) \text{ weakly in the two-scale sense,}$$
$$\nabla u_\varepsilon \xrightarrow{t.s.} \nabla_{\bar{x}} u_0(t,\bar{x}) + \nabla_y u_1(t,\bar{x},y) \text{ weakly in the two-scale sense.}$$

(ii) *Let* u_ε *be a sequence in* $L^2((0,T), H^1(G_\varepsilon))$*, such that*

$$\frac{1}{\sqrt{\varepsilon}}||u_\varepsilon||_{L^2((0,T)\times G_\varepsilon)} + \sqrt{\varepsilon}||\nabla u_\varepsilon||_{L^2((0,T)\times G_\varepsilon)} \leq C.$$

Then there exists $u_0 \in L^2((0,T) \times \Sigma, H^1_{per}(Y, H^1[-1,1]))$ *such that*

$$u_\varepsilon \xrightarrow{t.s.} u_0(t,\bar{x},y) \text{ weakly in the two-scale sense,}$$
$$\varepsilon\nabla u_\varepsilon \xrightarrow{t.s.} \nabla_y u_0(t,\bar{x},y) \text{ weakly in the two-scale sense.}$$

The proof of this theorem is given by using Theorem 4.2 and arguments similar to the ones in Proposition 1.14 in [1].

Equivalent formulation. When we are dealing with nonlinear problems, the weak two-scale convergence is no longer sufficient for passing to the limit in the nonlinear terms. Here one needs strongly two-scale convergent sequences. However, it is very difficult to show directly the strong two-scale convergence for sequences defined on varying domains, e.g., thin heterogeneous layers. In such cases we use an equivalent characterization of the two-scale convergence described below. This reformulation has the strong advantage that sequences of functions defined on varying domains are transformed to sequences on fixed domains.

Contributions to the development of this method are given in [3], [5], [8], and [9]. In our paper we adapt this method to the case of thin heterogeneous layers and nonlinear problems. Recently, in [16], a proof for the "thick" Neumann sieve was given using multiscale techniques based on [9].

For each $\varepsilon > 0$, let us consider the lattice

$$\mathcal{A}_\varepsilon = \{\bar{x} = \varepsilon i, i \in \mathbb{Z}^{n-1}\} = \varepsilon\mathbb{Z}^{n-1}.$$

To every $\bar{x} \in \Sigma$ we can associate a unique lattice point $c_\varepsilon(\bar{x}) := \varepsilon\left[\frac{\bar{x}}{\varepsilon}\right] \in \mathcal{A}_\varepsilon$, such that $\bar{x} \in c_\varepsilon(\bar{x}) + \varepsilon Y$. For simplicity, from now on we consider domains G_ε of the form $G_\varepsilon =]0,1[^{n-1} \times]-\varepsilon,\varepsilon[$.

Definition 4.5. *We define the operator* L_ε*, mapping measurable functions* u_ε *on* $(0,T) \times G_\varepsilon$ *to measurable functions* $L_\varepsilon u_\varepsilon$ *on* $(0,T) \times \Sigma \times Z$*, with*

$$(L_\varepsilon u_\varepsilon)(t,\bar{x},y) = u_\varepsilon(t,(c_\varepsilon(\bar{x}),0) + \varepsilon y) \tag{4.6}$$

for a.e. $\bar{x} \in c_\varepsilon(\bar{x}) + \varepsilon Y,\ (t,y) \in (0,T) \times Z$.

The following properties of the operator L_ε can be proved in analogy to Lemma 2 in [3].

Lemma 4.6. *For* $u_\varepsilon, v_\varepsilon \in L^2((0,T) \times G_\varepsilon)$*, we have*

$$(L_\varepsilon u_\varepsilon, L_\varepsilon v_\varepsilon)_{L^2((0,T)\times\Sigma\times Z)} = \frac{1}{\varepsilon}(u_\varepsilon, v_\varepsilon)_{L^2((0,T)\times G_\varepsilon)},$$
$$||L_\varepsilon u_\varepsilon||_{L^2((0,T)\times\Sigma\times Z)} = \frac{1}{\sqrt{\varepsilon}}||u_\varepsilon||_{L^2((0,T)\times G_\varepsilon)},$$
$$\nabla_y L_\varepsilon u_\varepsilon = \varepsilon L_\varepsilon(\nabla_x u_\varepsilon) \quad \text{a.e. in } (0,T)\times\Sigma\times Z.$$

The next proposition shows that weak (strong) two-scale convergence for a sequence $u_\varepsilon \in L^2((0,T)\times G_\varepsilon)$ is equivalent to weak (strong) convergence for the sequence $L_\varepsilon u_\varepsilon$ in $L^2((0,T)\times\Sigma\times Z)$.

PROPOSITION 4.7. *Let $u_\varepsilon \in L^2((0,T)\times G_\varepsilon)$ be a sequence, such that*

$$\frac{1}{\sqrt{\varepsilon}}||u_\varepsilon||_{L^2((0,T)\times G_\varepsilon)} \leq C.$$

Then there exists a subsequence (again denoted by u_ε) and a limit function $u_0 \in L^2((0,T)\times\Sigma\times Z)$, such that the following statements are equivalent:

(i) $u_\varepsilon \xrightarrow{t.s.} u_0$ *weakly (strongly) in the two-scale sense.*
(ii) $L_\varepsilon u_\varepsilon \longrightarrow u_0$ *weakly (strongly) in* $L^2((0,T)\times\Sigma\times Z)$.

Proof. Since by Lemma 4.6

$$||L_\varepsilon u_\varepsilon||_{L^2((0,T)\times\Sigma\times Z)} = \frac{1}{\sqrt{\varepsilon}}||u_\varepsilon||_{L^2((0,T)\times G_\varepsilon)} \leq C$$

it follows that there exists a subsequence of u_ε (again denoted by u_ε), and there exist $u_0, u_* \in L^2((0,T)\times\Sigma\times Z)$, such that

$$\begin{aligned} u_\varepsilon &\xrightarrow{\text{t.s.}} u_0 \quad \text{weakly in the two-scale sense,}\\ L_\varepsilon u_\varepsilon &\longrightarrow u_* \quad \text{weakly in } L^2((0,T)\times\Sigma\times Z). \end{aligned}$$

Then a proof analogous to that of Proposition 4.6 in [5] shows that $u_0 \equiv u_*$. To prove the equivalence of statements (i) and (ii) with respect to the strong convergences let us remark that

$$\begin{aligned} ||L_\varepsilon u_\varepsilon - u_0||^2_{L^2((0,T)\times\Sigma\times Z)} &= \int_0^T\int_\Sigma\int_Z |L_\varepsilon u_\varepsilon - u_0|^2\, dy\, d\bar{x}\, dt\\ = ||L_\varepsilon u_\varepsilon||^2_{L^2((0,T)\times\Sigma\times Z)} &- 2\int_0^T\int_\Sigma\int_Z (L_\varepsilon u_\varepsilon)u_0\, dy\, d\bar{x}\, dt + ||u_0||^2_{L^2((0,T)\times\Sigma\times Z)}\\ = \frac{1}{\varepsilon}||u_\varepsilon||^2_{L^2((0,T)\times G_\varepsilon)} &- 2\int_0^T\int_\Sigma\int_Z (L_\varepsilon u_\varepsilon)u_0\, dy\, d\bar{x}\, dt + ||u_0||^2_{L^2((0,T)\times\Sigma\times Z)}. \end{aligned}$$

Taking now the limits $\varepsilon \to 0$ on both sides of this equality, the equivalence of (i) and (ii) is proved. □

5. Proofs for the convergence results stated in Propositions 2.1 and 2.2 and Theorem 2.3. From the a priori estimates we see that we have different compactness properties for the solutions u_ε on the subdomains $\Omega^\pm_\varepsilon$ and Ω^M_ε. Thus we have to study the convergence of the sequences $u^\pm_\varepsilon$ and u^M_ε separately.

5.1. Convergence in the bulk. In this subsection, we give the proof of the first three convergence results from Proposition 2.1.

Proof. Let us consider the transformations

$$\Omega^\pm \mapsto \Omega^\pm_\varepsilon, \quad (\bar{x}, \tilde{x}_n) \mapsto (\bar{x}, x_n) = \left(\bar{x}, \frac{H-\varepsilon}{H}\tilde{x}_n \pm \varepsilon\right) \tag{5.1}$$

and define

$$\tilde{u}_\varepsilon^\pm : [0,T] \times \Omega^\pm \to \mathbb{R}^m, \quad \tilde{u}_\varepsilon^\pm(t,\bar{x},\tilde{x}_n) = u_\varepsilon^\pm\left(t,\bar{x},\frac{H-\varepsilon}{H}\tilde{x}_n \pm \varepsilon\right). \tag{5.2}$$

Using the transformation formula for integrals, we can easily show the estimates for the functions $\tilde{u}_\varepsilon^\pm$,

$$||\tilde{u}_{j\varepsilon}^\pm||_{L^2(0,T;H^1(\Omega^\pm))} \le C\,||u_{j\varepsilon}^\pm||_{L^2((0,T),H^1(\Omega_\varepsilon^\pm))}, \tag{5.3}$$

$$||\partial_t\tilde{u}_{j\varepsilon}^\pm||_{L^2((0,T),L^2(\Omega^\pm))} \le C\,||\partial_t u_{j\varepsilon}^\pm||_{L^2((0,T),L^2(\Omega_\varepsilon^\pm))}, \tag{5.4}$$

with a constant C independent of ε. Now, since the functions $\tilde{u}_\varepsilon^\pm$ are defined on fixed domains $\Omega^\pm$, standard compactness results together with the estimates (5.3) and (5.4) imply that there exist $u_0^\pm \in L^2((0,T),H^1(\Omega^\pm,\mathbb{R}^m))$ with $\partial_t u_0^\pm \in L^2((0,T),L^2(\Omega^\pm,\mathbb{R}^m))$, such that up to a subsequence

$$\begin{aligned}
&\tilde{u}_{j\varepsilon}^\pm \to u_{j0}^\pm \quad \text{weakly in } L^2((0,T),H^1(\Omega^\pm)),\\
&\partial_t\tilde{u}j_\varepsilon{}^\pm \to \partial_t u_{j0}^\pm \quad \text{weakly in } L^2((0,T),L^2(\Omega^\pm)),\\
&\tilde{u}_{j\varepsilon}^\pm \to u_{j0}^\pm \quad \text{strongly in } L^2((0,T),L^2(\Omega^\pm)).
\end{aligned}$$

The strong convergence follows from the estimate

$$||\tilde{u}_{j\varepsilon}^\pm||_{L^2(0,T;H^1(\Omega^\pm))} + ||\partial_t\tilde{u}_{j\varepsilon}^\pm||_{L^2(0,T;L^2(\Omega^\pm))} \le C$$

and a compactness theorem of Lions; see [10, Theorem 1, p. 58].

Now let $\varphi \in C_0^\infty([0,T]\times\Omega^\pm,\mathbb{R}^m)$. Using the transformations (5.1), (5.2), we have

$$\begin{aligned}
&\int_0^T\int_{\Omega^\pm}(\chi_{\Omega_\varepsilon^\pm}u_{j\varepsilon})(t,x)\varphi_j(t,x)\,dx\,dt = \int_0^T\int_{\Omega_\varepsilon^\pm}u_{j\varepsilon}^\pm(t,\bar{x},x_3)\varphi_j(t,\bar{x},x_3)d\bar{x}\,dx_3\,dt\\
&= \int_0^T\int_{\Omega^\pm}u_{j\varepsilon}^\pm\left(t,\bar{x},\frac{H-\varepsilon}{H}\tilde{x}_3\pm\varepsilon\right)\varphi_j\left(t,\bar{x},\frac{H-\varepsilon}{H}\tilde{x}_3\pm\varepsilon\right)\frac{H-\varepsilon}{H}d\bar{x}\,d\tilde{x}_3\,dt\\
&= \frac{H-\varepsilon}{H}\int_0^T\int_{\Omega^\pm}\tilde{u}_{j\varepsilon}^\pm(t,\bar{x},\tilde{x}_3)\varphi_j(t,\bar{x},\tilde{x}_3)d\bar{x}\,d\tilde{x}_3\,dt\\
&\quad+\frac{H-\varepsilon}{H}\int_0^T\int_{\Omega^\pm}\tilde{u}_{j\varepsilon}^\pm(t,\bar{x},\tilde{x}_3)\left[\varphi_j\left(t,\bar{x},\frac{H-\varepsilon}{H}\tilde{x}_3\pm\varepsilon\right)-\varphi_j(t,\bar{x},\tilde{x}_3)\right]d\bar{x}\,d\tilde{x}_3\,dt\\
&\to \int_0^T\int_{\Omega^\pm}u_{j0}^\pm(t,\bar{x},\tilde{x})\varphi_j(t,\bar{x},\tilde{x})\,d\bar{x}\,d\tilde{x}\,dt
\end{aligned}$$

due to the convergence properties of $\tilde{u}_\varepsilon^\pm$ and the smoothness of φ. Thus

$$\chi_{\Omega_\varepsilon^\pm}u_{j\varepsilon} \to u_{j0}^\pm \quad \text{weakly in } L^2((0,T),L^2(\Omega^\pm)).$$

Additionally, we have that

$$||\chi_{\Omega_\varepsilon^\pm}u_{j\varepsilon}||_{L^2((0,T),L^2(\Omega^\pm))} = \sqrt{1-\varepsilon/H}||\tilde{u}_{j\varepsilon}^\pm||_{L^2(0,T;L^2(\Omega^\pm))} \to ||u_{j0}^\pm||_{L^2(0,T;L^2(\Omega^\pm))},$$

since $\tilde{u}_\varepsilon^\pm$ converges to $u_0^\pm$ strongly in $L^2((0,T)\times\Omega^\pm,\mathbb{R}^m)$. Thus, statement 1 of Proposition 2.1 is proved. The proofs of statements 2 and 3 follow along the same line. $\square$

5.2. Convergence for the traces on the interfaces bulk/layer. The compactness of the traces of u_ε on $S_\varepsilon^\pm$ is crucial for the control of the solutions in the layer. In the following, we give the proof of Proposition 2.2.

Proof. Since

$$||\tilde{u}_{j\varepsilon}^\pm||_{L^2((0,T),H^1(\Omega^\pm))} + ||\partial_t \tilde{u}_{j\varepsilon}^\pm||_{L^2((0,T),L^2(\Omega^\pm))} \le C$$

and the embedding

$$H^1(\Omega^\pm) \hookrightarrow H^\beta(\Omega^\pm)$$

is compact for every $\frac{1}{2} < \beta < 1$, it follows from Lions compactness theorem [10, Theorem 1, p. 58] that there exists a subsequence such that

$$\tilde{u}_{j\varepsilon}^\pm \to u_{j0}^\pm \quad \text{strongly in } L^2((0,T), H^\beta(\Omega^\pm)).$$

Due to the continuity of the embedding

$$H^\beta(\Omega^\pm) \hookrightarrow L^2(\partial\Omega^\pm) \quad \text{for } \frac{1}{2} < \beta < 1,$$

it follows that

$$||\tilde{u}_{j\varepsilon}^\pm - u_{j0}^\pm||_{L^2((0,T)\times\Sigma)} \le C||\tilde{u}_{j\varepsilon}^\pm - u_{j0}^\pm||_{L^2((0,T),H^\beta(\Omega^\pm))} \to 0 \quad \text{for } \varepsilon \to 0.$$

Thus the first assertion is proved. To prove the second one, we notice that, since $\tilde{u}_\varepsilon|_\Sigma$ is strongly convergent in $L^2((0,T)\times\Sigma, \mathbb{R}^m)$, it is also weakly two-scale convergent to the same limit; see [1]. Then, using again (5.1) and (5.2), we obtain for $\varepsilon \to 0$

$$\begin{aligned}\int_0^T \int_{S_\varepsilon^\pm} u_{j\varepsilon}^\pm(t,x)\varphi_j\left(t,\bar{x},\frac{x}{\varepsilon}\right) dxdt &= \frac{H-\varepsilon}{H}\int_0^T \int_\Sigma \tilde{u}_{j\varepsilon}^\pm(t,\bar{x},0)\varphi_j\left(t,\bar{x},\frac{\bar{x}}{\varepsilon},\pm 1\right) d\bar{x}dt \\ &\to \int_0^T \int_\Sigma \int_Y u_{j0}^\pm(t,\bar{x},0)\varphi_j\left(t,\bar{x},\bar{y},\pm 1\right) d\bar{y}d\bar{x}dt. \qquad \square\end{aligned}$$

5.3. Weak two-scale convergence in the layer. The compactness results with respect to the weak two-scale convergence of u_ε^M in the layer stated in statements 4–6 of Proposition 2.1 follow directly from Proposition 4.4 and the a priori estimates for u_ε^M. It remains to prove the coupling condition between the effective solutions in the bulk regions and the local solution in the layer given in (2.19).

Proof. Let us start from the identity

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^T\int_{\Omega_\varepsilon^M}\varepsilon\nabla u_{j\varepsilon}^M(t,x)\varphi_j\left(t,\bar{x},\frac{x}{\varepsilon}\right)dx\,dt = \int_0^T\int_\Sigma\int_Z \nabla_y u_{j0}^M(t,\bar{x},y)\varphi_j(t,\bar{x},y)\,dy\,d\bar{x}\,dt$$

for any $\varphi_j \in C^\infty([0,T]\times\Sigma\times Z, \mathbb{R}^n)$ periodic in $Y = [0,1]^{n-1}$ and with compact support with respect to $\bar{x} \in \Sigma$. Integrating by parts on the left-hand side and using the continuity of the solution u_ε on S_ε^+, respectively, S_ε^-, and Proposition 2.2, we obtain

$$\lim_{\varepsilon\to 0}\left\{-\frac{1}{\varepsilon}\int_0^T\int_{\Omega_\varepsilon^M} u_{j\varepsilon}^M(t,x)\left(\varepsilon\nabla_{\bar{x}}\,\varphi_j\left(t,\bar{x},\frac{x}{\varepsilon}\right)+\nabla_y\,\varphi_j\left(t,\bar{x},\frac{x}{\varepsilon}\right)\right)dx\,dt\right.$$
$$\left.+\int_0^T\int_{S_\varepsilon^+} u_{j\varepsilon}^+(t,x)\varphi_j\left(t,\bar{x},\frac{x}{\varepsilon}\right)\cdot\nu\,dx\,dt+\int_0^T\int_{S_\varepsilon^-} u_{j\varepsilon}^-(t,x)\varphi_j\left(t,\bar{x},\frac{x}{\varepsilon}\right)\cdot\nu\,dx\,dt\right\}$$
$$=-\int_0^T\int_\Sigma\int_Z u_{j0}^M(t,\bar{x},y)\nabla_y\,\varphi_j(t,\bar{x},y)\,dy\,d\bar{x}\,dt$$
$$+\int_0^T\int_\Sigma\int_Y u_{j0}^+(t,\bar{x},0)\varphi_j(t,\bar{x},\bar{y},1)\cdot n\,dy\,d\bar{x}\,dt$$
$$-\int_0^T\int_\Sigma\int_Y u_{j0}^-(t,\bar{x},0)\varphi_j(t,\bar{x},\bar{y},-1)\cdot n\,dy\,d\bar{x}\,dt.$$

Here $n=(0,\dots,0,1)$. By equality between the two limits, we obtain statement (2.19) of the theorem. ∎

5.4. Strong two-scale convergence in the layer. In the following, we prove strong two-scale convergence for the sequence u_ε^M. The method to be used is of general interest. Here, we are in the situation that the standard estimates of the solutions u_ε^M imply only weak two-scale convergence, and due to the scaling in the diffusion coefficients we cannot use the method of bounded extensions to get strong convergence for u_ε^M in L^2. The best compactness result we get for u_ε^M is strong two-scale convergence. To prove this, we show the strong convergence of the transformed solution $L_\varepsilon u_\varepsilon^M$ in $L^2((0,T)\times\Sigma\times Z)$ (see Proposition 4.7). The control of the dependence of $L_\varepsilon u_\varepsilon^M$ on y and t is standard since we can use the differential equations in the cells. However, the dependence on x poses more serious problems since the functions $L_\varepsilon u_\varepsilon^M$ are only step functions with respect to x. To get equicontinuity in L^2 also with respect to shifts in x, we mainly have to compare solutions of the differential equations in different cells. A similar argument, although in a quite different situation, can be found in [7], which deals with the Neumann problem for rapidly oscillatory boundaries.

Verifying the compactness criteria requires the extension of u_ε^M into a neighborhood of Ω_ε^M. We construct the extension

$$\hat{u}_\varepsilon^M:(0,T)\times\mathbb{R}^{n-1}\times\,]-\varepsilon,\varepsilon[\,\to\mathbb{R}^m$$

as follows.

First, we extend u_ε^M by reflection with respect to the plane $\{(x_1,x_2,\dots,x_n):x_1=0\}$ to

$$\hat{u}_\varepsilon^M(t,x)=\begin{cases} u_\varepsilon^M(t,x), & (t,x)\in(0,T)\times\Omega_\varepsilon^M,\\ u_\varepsilon^M(t,-x_1,x_2,\dots,x_{n-1},x_n), & (t,-x_1,x_2,\dots,x_{n-1},x_n)\in(0,T)\times\Omega_\varepsilon^M.\end{cases}$$

Then we repeat this extension procedure with respect to the planes $\{(x_1,x_2,\dots,x_n):x_2=0\},\dots,\{(x_1,x_2,\dots,x_n):x_{n-1}=0\}$ and obtain the extension $\hat{u}_\varepsilon^M$ on the domain

$$(0,T)\times\hat{\Omega}_\varepsilon^M=(0,T)\times\,]-1,1[^{n-1}\times\,]-\varepsilon,\varepsilon[.$$

Now, we further extend $\hat{u}_\varepsilon^M$ by periodicity to $(0,T)\times\mathbb{R}^{n-1}\times\,]-\varepsilon,\varepsilon[$. Due to our extension procedure, we have that

$$\hat{u}_\varepsilon^M\in L^2((0,T),H^1_{per}(]-1,1[^{n-1},H^1(]-\varepsilon,\varepsilon[,\mathbb{R}^m)))$$

and

$$\|\hat{u}_\varepsilon^M\|_{L^2((0,T),H^1(\hat{\Omega}_\varepsilon^M,\mathbb{R}^m))} \le C\|u_\varepsilon^M\|_{L^2((0,T),H^1(\Omega_\varepsilon^M,\mathbb{R}^m))}.$$

In an analogous way, we extend $u_\varepsilon^\pm$ to $\hat{u}_\varepsilon^\pm$, satisfying

$$\hat{u}_\varepsilon^+ \in L^2((0,T), H^1_{per}(]-1,1[^{n-1}, H^1(]\varepsilon, H[,\mathbb{R}^m))),$$

$$\hat{u}_\varepsilon^- \in L^2((0,T), H^1_{per}(]-1,1[^{n-1}, H^1(]-H,-\varepsilon[,\mathbb{R}^m))).$$

We also define

$$\hat{S}_\varepsilon^\pm = \{x = (\bar{x}, x_n) : \bar{x} \in \,]-1,1[^{n-1}, x_n = \pm\varepsilon\}$$

and remark that the traces $\hat{u}_\varepsilon^\pm|_{\hat{S}_\varepsilon^\pm}$ are periodic with respect to $\bar{x}$ with period $]-1,1[^{n-1}$, and the continuity condition

$$\hat{u}_\varepsilon^M|_{\hat{S}_\varepsilon^\pm} = \hat{u}_\varepsilon^\pm|_{\hat{S}_\varepsilon^\pm}$$

holds. Let us now give the proof of Theorem 2.3.

Proof. We denote

$$\begin{aligned}
\delta\hat{u}_\varepsilon^M(t,x) &= \hat{u}_\varepsilon^M(t, x+(l,0)\varepsilon) - \hat{u}_\varepsilon^M(t,x),\\
\delta\hat{u}_\varepsilon^\pm(t,x) &= \hat{u}_\varepsilon^\pm(t, x+(l,0)\varepsilon) - \hat{u}_\varepsilon^\pm(t,x),\\
\delta g(t,x) &= g\left(\frac{x}{\varepsilon}, u_\varepsilon^M(t, x+(l,0)\varepsilon)\right) - g\left(\frac{x}{\varepsilon}, u_\varepsilon^M(t,x)\right)\\
\delta U_0^M(x) &= U_0^M\left(\bar{x}+l\varepsilon, \frac{x_n}{\varepsilon}\right) - U_0^M\left(\bar{x}, \frac{x_n}{\varepsilon}\right).
\end{aligned}$$

Now let $\delta v_\varepsilon = (\delta v_{1\varepsilon}, \dots, \delta v_{m\varepsilon})$ be the solution to the following problem:

$$\begin{aligned}
\frac{1}{\varepsilon}\partial_t(\delta v_{j\varepsilon})(t,x) - \varepsilon\nabla\left(D_j^M\left(\frac{x}{\varepsilon}\right)\nabla(\delta v_{j\varepsilon})(t,x)\right) &= 0 \quad \text{in } (0,T)\times\hat{\Omega}_\varepsilon^M,\\
\delta v_\varepsilon(t,x) &= \delta\hat{u}_\varepsilon^\pm(t,x) \quad \text{on } (0,T)\times\hat{S}_\varepsilon^\pm,\\
\delta v_\varepsilon(0,x) &= 0 \quad \text{in } \hat{\Omega}_\varepsilon^M\\
\delta v_\varepsilon \text{ is periodic with period } \ ]-1,1[^{n-1}.&
\end{aligned}$$

The following estimates hold:

$$\text{(5.5)}\quad \frac{1}{\sqrt{\varepsilon}}||\delta v_\varepsilon||_{L^\infty((0,T),L^2(\hat{\Omega}_\varepsilon^M,\mathbb{R}^m))} + \sqrt{\varepsilon}||\nabla(\delta v_\varepsilon)||_{L^2((0,T)\times\hat{\Omega}_\varepsilon^M,\mathbb{R}^m)} \le C,$$

$$\text{(5.6)}\quad \frac{1}{\sqrt{\varepsilon}}||\delta v_\varepsilon||_{L^2((0,T)\times\hat{\Omega}_\varepsilon^M,\mathbb{R}^m)} \le C(||\delta\hat{u}_\varepsilon^+||_{L^2((0,T)\times\hat{S}_\varepsilon^+,\mathbb{R}^m)} + ||\delta\hat{u}_\varepsilon^-||_{L^2((0,T)\times\hat{S}_\varepsilon^-,\mathbb{R}^m)}).$$

To prove (5.5), we test the equation for δv_ε with $\delta v_\varepsilon - \delta\hat{u}_\varepsilon^M$ and use the same techniques as in the first part the proof of Lemma 3.1, together with the a priori estimates for u_ε^M. For the proof of (5.6), we use the solution $h = (h_1, \dots, h_m)$ to the adjoint problem

$$\begin{aligned}
-\frac{1}{\varepsilon}\partial_t h_j(t,x) - \varepsilon\nabla(D_j^M(\frac{x}{\varepsilon})\nabla h_j(t,x)) &= \frac{1}{\varepsilon}\delta v_{j\varepsilon}(t,x) \quad \text{in } (0,T)\times\hat{\Omega}_\varepsilon^M,\\
h(t,x) &= 0 \quad \text{on } (0,T)\times\hat{S}_\varepsilon^\pm,\\
h(T,x) &= 0 \quad \text{in } \hat{\Omega}_\varepsilon^M,\\
h \text{ is periodic with period } \ ]-1,1[^{n-1}.&
\end{aligned}$$

Regularity theory for parabolic systems together with a scaling argument implies that $h \in L^2((0,T), H^2(\hat{\Omega}_\varepsilon^M, \mathbb{R}^m))$ and we have

$$
(5.7)\quad \frac{1}{\sqrt{\varepsilon}}||h_j||_{L^2((0,T)\times\hat{\Omega}_\varepsilon^M)} + \sqrt{\varepsilon}||\nabla h_j||_{L^2((0,T)\times\hat{\Omega}_\varepsilon^M,\mathbb{R}^n)} + \varepsilon\sqrt{\varepsilon}||\nabla^2 h_j||_{L^2((0,T)\times\hat{\Omega}_\varepsilon^M,\mathbb{R}^{n^2})} \leq \frac{C}{\sqrt{\varepsilon}}||\delta v_{j\varepsilon}||_{L^2((0,T)\times\hat{\Omega}_\varepsilon^M,\mathbb{R}^m)}.
$$

Let us now multiply the equation for h by $\delta v_{j\varepsilon}$ and integrate. We obtain

$$
\begin{aligned}
&\frac{1}{\varepsilon}\int_0^T\int_{\hat{\Omega}_\varepsilon^M}|\delta v_{j\varepsilon}|^2 dxdt \\
&= \int_0^T\int_{\hat{\Omega}_\varepsilon^M}\left(-\frac{1}{\varepsilon}\partial_t h_j - \varepsilon\nabla\left(D_j^M\left(\frac{x}{\varepsilon}\right)\nabla h_j\right)(\delta v_{j\varepsilon})\right)dxdt \\
&= \int_0^T\int_{\hat{\Omega}_\varepsilon^M}\frac{1}{\varepsilon}h_j\partial_t(\delta v_{j\varepsilon}) + \varepsilon D_j^M\left(\frac{x}{\varepsilon}\right)\nabla h_j\nabla(\delta v_{j\varepsilon})dxdt \\
&\quad - \varepsilon\int_0^T\int_{\hat{S}_\varepsilon^+}D_j^M\left(\frac{x}{\varepsilon}\right)\delta v_{j\varepsilon}\nabla h_j\cdot\nu dxdt - \varepsilon\int_0^T\int_{\hat{S}_\varepsilon^-}D_j^M\left(\frac{x}{\varepsilon}\right)\delta v_{j\varepsilon}\nabla h_j\cdot\nu dxdt.
\end{aligned}
$$

Now, using the problem for δv_ε and the fact that h is equal to zero on $(0,T)\times\hat{S}_\varepsilon^\pm$, we obtain

$$
\begin{aligned}
&\frac{1}{\varepsilon}||\delta v_{j\varepsilon}||^2_{L^2((0,T)\times\hat{\Omega}_\varepsilon^M)} \leq C\varepsilon||\delta v_{j\varepsilon}||_{L^2((0,T)\times\hat{S}_\varepsilon^+)}||\nabla h_j\cdot\nu||_{L^2((0,T)\times\hat{S}_\varepsilon^+)} \\
&\quad + C\varepsilon||\delta v_{j\varepsilon}||_{L^2((0,T)\times\hat{S}_\varepsilon^-)}||\nabla h_j\cdot\nu||_{L^2((0,T)\times\hat{S}_\varepsilon^-)} \\
&\leq C\varepsilon\left(\frac{1}{\sqrt{\varepsilon}}||\nabla h_j\cdot\nu||_{L^2((0,T)\times\hat{\Omega}_\varepsilon^M,\mathbb{R}^n)} + \sqrt{\varepsilon}||\nabla(\nabla h_j\cdot\nu)||_{L^2((0,T)\times\hat{\Omega}_\varepsilon^M,\mathbb{R}^{n^2})}\right) \\
&\quad \times(||\delta v_{j\varepsilon}||_{L^2((0,T)\times\hat{S}_\varepsilon^+)} + ||\delta v_{j\varepsilon}||_{L^2((0,T)\times\hat{S}_\varepsilon^-)}).
\end{aligned}
$$

For the last inequality, we used the trace estimate for the thin domain $\hat{\Omega}_\varepsilon^M$ given in Lemma 5.1. The estimate (5.7) and the boundary condition for δv_ε on $\hat{S}_\varepsilon^\pm$ imply the assertion (5.6).

Let us now consider the function $\delta w_\varepsilon = (\delta w_{1\varepsilon}, \dots, \delta w_{m\varepsilon})$ defined by

$$
\delta w_\varepsilon(t,x) = \delta\hat{u}_\varepsilon^M(t,x) - \delta v_\varepsilon(t,x).
$$

In order to avoid the boundary values of δw_ε on the lateral boundary of Ω_ε^M, we will cut off this part of the boundary and estimate $\frac{1}{\sqrt{\varepsilon}}||\delta w_\varepsilon||_{L^2((0,T)\times\Omega_{2h}^M,\mathbb{R}^m)}$, where

$$
(5.8)\qquad \Omega_{2h}^M = \{x\in\Omega_\varepsilon^M : 2h < x_i < 1-2h,\ i = 1,\dots,n-1\}.
$$

Defining Ω_h^M and $S_h^\pm$ analogously to (5.8), we have that δw_ε satisfies

$$
\begin{aligned}
\frac{1}{\varepsilon}\partial_t(\delta w_{j\varepsilon})(t,x) - \varepsilon\nabla\left(D_j^M\left(\frac{x}{\varepsilon}\right)\nabla(\delta w_{j\varepsilon})(t,x)\right) &= \frac{1}{\varepsilon}\delta g_j(t,x) && \text{in } (0,T)\times\Omega_h^M, \\
(\delta w_\varepsilon)(t,x) &= 0 && \text{on } (0,T)\times S_h^\pm, \\
(\delta w_\varepsilon)(0,x) &= \delta U_0^M(x) && \text{in } \Omega_h^M.
\end{aligned}
$$

Now, we test the equation for δw_ε with a function which vanishes on the lateral boundary of Ω_ε^M. We consider the following cut-off function $\varphi \in C_0^\infty((h, 1-h)^{n-1}, C^\infty(]-\varepsilon, \varepsilon[, [0,1]))$ with the properties

$$\varphi \equiv 1 \quad \text{in } \Omega_{2h}^M, \quad ||\nabla\varphi||_{L^\infty(\Omega_h^M)} \leq \frac{C}{h}.$$

Multiplying the equation for $\delta w_{j\varepsilon}$ by $\delta w_{j\varepsilon}\varphi^2$ and integrating over Ω_h^M, we obtain

$$\frac{1}{\varepsilon}\int_{\Omega_h^M} \partial_t(\delta w_{j\varepsilon})\delta w_{j\varepsilon}\varphi^2 + \varepsilon\int_{\Omega_h^M} D_j^M\left(\frac{x}{\varepsilon}\right)\nabla(\delta w_{j\varepsilon})\nabla(\delta w_{j\varepsilon})\varphi^2 \tag{5.9}$$

$$+\,\varepsilon\int_{\Omega_h^M} D_j^M\left(\frac{x}{\varepsilon}\right)\nabla(\delta w_{j\varepsilon})\delta w_{j\varepsilon}2\varphi\nabla\varphi = \frac{1}{\varepsilon}\int_{\Omega_h^M}\delta g_j\delta w_{j\varepsilon}\varphi^2. \tag{5.10}$$

Integration with respect to time and the Lipschitz continuity of g in the second argument imply

$$\begin{aligned}
&\frac{1}{\varepsilon}\int_{\Omega_h^M}|\delta w_{j\varepsilon}(t)|^2\varphi^2 + \varepsilon\int_0^t\int_{\Omega_h^M}|\nabla(\delta w_{j\varepsilon})|^2\varphi^2 \\
&\leq \frac{C}{\varepsilon}\left(\int_0^t\int_{\Omega_h^M}|\delta u_{j\varepsilon}^M||\delta w_{j\varepsilon}|\varphi^2 + \int_{\Omega_h^M}|\delta U_{j0}^M|^2\varphi^2\right) \\
&+ C\left(\eta\varepsilon\int_0^t\int_{\Omega_h^M}|\nabla(\delta w_{j\varepsilon})\varphi|^2 + \frac{1}{\eta}\varepsilon\int_0^t\int_{\Omega_h^M}|\delta w_{j\varepsilon}|^2|\nabla\varphi|^2\right).
\end{aligned}\tag{5.11}$$

Let us now estimate the last term on the right-hand side of (5.11). Since $\varphi \equiv 1$ in Ω_{2h}^M, the support of $\nabla\varphi$ is contained in the domain

$$T_h^M := \Omega_h^M \setminus \Omega_{2h}^M.$$

Using the estimate for $\nabla\varphi$ and Poincaré's inequality (since $\delta w_\varepsilon|_{S_\varepsilon^\pm} = 0$), we have

$$\begin{aligned}
&\varepsilon\int_0^t\int_{T_h^M}|\delta w_{j\varepsilon}|^2|\nabla\varphi|^2 \leq \frac{C\varepsilon}{h^2}\int_0^t\int_{T_h^M}|\delta w_{j\varepsilon}|^2 \\
&\leq \frac{C\varepsilon^3}{h^2}\int_0^t\int_{T_h^M}|\nabla(\delta w_{j\varepsilon})|^2 \leq \frac{C\varepsilon^3}{h^2}\int_0^t\int_{T_h^M}|\nabla(\delta\hat{u}_{j\varepsilon}^M)|^2 + |\nabla(\delta v_{j\varepsilon})|^2) \\
&\leq \frac{C\varepsilon^2}{h^2}.
\end{aligned}\tag{5.12}$$

For the last inequality, we used the a priori estimates for u_ε^M and the estimate (5.5) for δv_ε.

Going back to relation (5.11), taking η small enough, and using the estimate (5.12) and the fact that $\delta u_\varepsilon^M = \delta v_\varepsilon + \delta w_\varepsilon$, we get

$$\begin{aligned}
&\frac{1}{\varepsilon}\int_{\Omega_h^M}|\delta w_{j\varepsilon}(t)|^2\varphi^2 + \varepsilon\int_0^t\int_{\Omega_h^M}|\nabla(\delta w_{j\varepsilon})|^2\varphi^2 \\
&\leq \frac{C}{\varepsilon}\left(\int_0^t\int_{\Omega_h^M}(|\delta u_{j\varepsilon}^M|^2 + |\delta w_{j\varepsilon}|^2)\varphi^2 + \int_{\Omega_h^M}|\delta U_{j0}^M|^2\varphi^2\right) + \frac{C\varepsilon^2}{h^2} \\
&\leq \frac{C}{\varepsilon}\left(\int_0^t\int_{\Omega_h^M}(|\delta v_{j\varepsilon}|^2 + |\delta w_{j\varepsilon}|^2)\varphi^2 + \int_{\Omega_h^M}|\delta U_{j0}^M|^2\varphi^2\right) + \frac{C\varepsilon^2}{h^2}.
\end{aligned}\tag{5.13}$$

Finally, Gronwall's lemma implies the estimate

$$\frac{1}{\varepsilon}||\delta w_{j\varepsilon}\varphi||^2_{L^2((0,T)\times\Omega_h^M)} \leq \frac{C}{\varepsilon}||\delta v_{j\varepsilon}\varphi||^2_{L^2((0,T)\times\Omega_h^M)} + \frac{C}{\varepsilon}||\delta U_{j0}^M\varphi||^2_{L^2(\Omega_h^M)} + \frac{C\varepsilon^2}{h^2}. \tag{5.14}$$

Now, it remains to estimate $\frac{1}{\sqrt{\varepsilon}}||\delta u_\varepsilon^M||_{L^2((0,T)\times(\Omega_\varepsilon^M\setminus\Omega_{2h}^M),\mathbb{R}^m)}$. For this estimate, we will exploit the fact that the domain $\Omega_\varepsilon^M\setminus\Omega_{2h}^M$ can be decomposed in subdomains which have thickness $O(h)$ at least in one space direction. Using the L^∞-estimate (3.4), we obtain

$$\frac{1}{\sqrt{\varepsilon}}\left\{\int_0^T\int_{\Omega_\varepsilon^M\setminus\Omega_{2h}^M}|u_{j\varepsilon}^M|^2\right\}^{\frac{1}{2}} \leq \frac{M_j e^{A_j T}}{\sqrt{\varepsilon}}\left\{\int_0^T\int_{\Omega_\varepsilon^M\setminus\Omega_{2h}^M}dxdt\right\}^{\frac{1}{2}} \leq Ch^{\frac{1}{2}}. \tag{5.15}$$

The estimates (5.6), (5.14), and (5.15) imply the estimate (2.22) from the first part of the theorem.

To prove the strong two-scale convergence for u_ε^M, we will show that

$$L_\varepsilon u_\varepsilon^M \to u_0^M \quad \text{strongly in} \quad L^2((0,T)\times\Sigma\times Z,\mathbb{R}^m).$$

We first show that for all $\rho > 0$, there exists $\delta > 0$, such that for all $\varepsilon \leq \varepsilon_0$

$$||L_\varepsilon\hat{u}_\varepsilon^M(t,\bar{x}+\bar{\xi},y) - L_\varepsilon\hat{u}_\varepsilon^M(t,\bar{x},y)||^2_{L^2((0,T)\times\Sigma\times Z,\mathbb{R}^m)} < \rho \tag{5.16}$$

for all $\bar{\xi}\in\mathbb{R}^{n-1}, |\bar{\xi}| < \delta$.

Let $I \subset \mathbb{Z}^{n-1}$, such that

$$\Sigma = \sum_{i\in I}\varepsilon(Y+i) =: \sum_{i\in I}\varepsilon Y_i.$$

Obviously, for $\bar{x}\in\varepsilon Y_i$ we have that $\left[\frac{\bar{x}}{\varepsilon}\right] = i$. For every $i \in I$ we divide the cell εY_i into subsets εY_i^k with $k\in\{0,1\}^{n-1}$, defined as follows:

$$\varepsilon Y_i^k = \left\{\bar{x}\in\varepsilon Y_i, \left[\frac{\bar{x}+\left\{\frac{\bar{\xi}}{\varepsilon}\right\}\varepsilon}{\varepsilon}\right]\varepsilon = \varepsilon(i+k)\right\}.$$

Then $\varepsilon Y_i = \bigcup_{k\in\{0,1\}^{n-1}}\varepsilon Y_i^k$. In Figure 5.1, we sketch the subsets εY_i^k of εY_i in the case $n = 3$, $Y = [0,1]^2$ and the translation $\bar{\xi}$ is of the form $\bar{\xi} = (\xi_1, 0),\ \xi_1 > 0$.

Now, let us calculate

$$\begin{aligned}
&||L_\varepsilon\hat{u}_{j\varepsilon}^M(t,\bar{x}+\bar{\xi},y) - L_\varepsilon\hat{u}_{j\varepsilon}^M(t,\bar{x},y)||^2_{L^2((0,T)\times\Sigma\times Z)}\\
&= \sum_{i\in I}\int_0^T\int_{\varepsilon Y_i}\int_Z\left|\hat{u}_{j\varepsilon}^M\left(t,\varepsilon\left(\left[\frac{\bar{x}+\bar{\xi}}{\varepsilon}\right],0\right)+\varepsilon y\right) - \hat{u}_{j\varepsilon}^M\left(t,\varepsilon\left(\left[\frac{\bar{x}}{\varepsilon}\right],0\right)+\varepsilon y\right)\right|^2 dyd\bar{x}dt\\
&= \sum_{i\in I}\sum_{k\in\{0,1\}^{n-1}}\int_0^T\int_{\varepsilon Y_i^k}\int_Z\left|\hat{u}_{j\varepsilon}^M\left(t,\varepsilon\left(i+\left[\frac{\bar{\xi}}{\varepsilon}\right]+k,0\right)+\varepsilon y\right) - \hat{u}_{j\varepsilon}^M\left(t,\varepsilon(i,0)+\varepsilon y\right)\right|^2 dyd\bar{x}dt\\
&\leq \sum_{i\in I}\sum_{k\in\{0,1\}^{n-1}}\int_0^T\int_{\varepsilon Y_i}\int_Z\left|\hat{u}_{j\varepsilon}^M\left(t,\varepsilon\left(i+\left[\frac{\bar{\xi}}{\varepsilon}\right]+k,0\right)+\varepsilon y\right) - \hat{u}_{j\varepsilon}^M\left(t,\varepsilon(i,0)+\varepsilon y\right)\right|^2 dyd\bar{x}dt\\
&= \sum_{k\in\{0,1\}^{n-1}}\frac{1}{\varepsilon}\int_0^T\int_{\Omega_\varepsilon^M}\left|\hat{u}_{j\varepsilon}^M\left(t,x+\varepsilon\left(\left[\frac{\bar{\xi}}{\varepsilon}\right]+k,0\right)\right) - \hat{u}_{j\varepsilon}^M(t,x)\right|^2 dxdt.
\end{aligned}$$

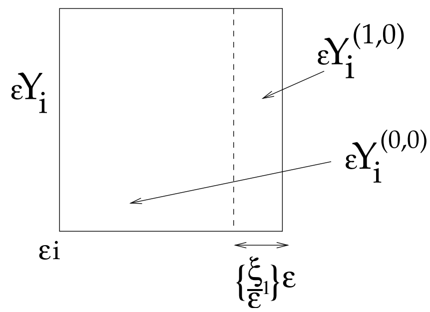


FIG. 5.1. *The subsets* εY_i^k *for* $\bar{\xi} = (\xi_1, 0)$, $\xi_1 > 0$.

Let us fix $h \in \left(0, \min\left\{\frac{1}{4}, \frac{\rho}{3}\frac{1}{2^{(n-1)}C}\right\}\right)$, where C is the constant in the estimate (2.22). Let δ_1 and ε_1 be such that for $|\bar{\xi}| < \delta_1$ and $\varepsilon < \varepsilon_1$ we have

$$\left| \left[\frac{\bar{\xi}}{\varepsilon}\right] \varepsilon + k\varepsilon \right| < h.$$

Then we can apply the first part of the theorem to estimate

$$\begin{aligned}
&\sum_{k\in\{0,1\}^{n-1}} \frac{1}{\varepsilon}\int_0^T \int_{\Omega_\varepsilon^M} \left| \hat{u}_{j\varepsilon}^M \left(t, x + \varepsilon\left(\left[\frac{\bar{\xi}}{\varepsilon}\right] + k, 0\right)\right) - \hat{u}_{j\varepsilon}^M(t,x)\right|^2 dxdt \\
\leq \quad & \sum_{k\in\{0,1\}^{n-1}} C \left\| \hat{u}_\varepsilon^+ \left(t, x + \varepsilon\left(\left[\frac{\bar{\xi}}{\varepsilon}\right] + k, 0\right)\right) - \hat{u}_\varepsilon^+(t,x)\right\|^2_{L^2((0,T)\times \hat{S}_\varepsilon^+, \mathbb{R}^m)} \\
+ & \sum_{k\in\{0,1\}^{n-1}} C \left\| \hat{u}_\varepsilon^- \left(t, x + \varepsilon\left(\left[\frac{\bar{\xi}}{\varepsilon}\right] + k, 0\right)\right) - \hat{u}_\varepsilon^-(t,x)\right\|^2_{L^2((0,T)\times \hat{S}_\varepsilon^-, \mathbb{R}^m)} \\
+ & \sum_{k\in\{0,1\}^{n-1}} \frac{C}{\varepsilon} \left\| \hat{U}_0^M \left(\bar{x} + \varepsilon\left(\left[\frac{\bar{\xi}}{\varepsilon}\right] + k\right), \frac{x_n}{\varepsilon}\right) - \hat{U}_0^M\left(\bar{x}, \frac{x_n}{\varepsilon}\right)\right\|^2_{L^2(\hat{\Omega}_\varepsilon^M, \mathbb{R}^m)} \\
+ & \sum_{k\in\{0,1\}^{n-1}} C\left\{\frac{\varepsilon^2}{h^2} + h\right\}.
\end{aligned} \tag{5.17}$$

The properties of the initial concentration U_0^M ensure that $\hat{U}_0^M$ is a function in H^1 on the rescaled domain obtained from $\hat{\Omega}_\varepsilon^M$, and thus it satisfies the Kolmogorov criterion. This fact and the strong convergence of the traces of $u_\varepsilon^\pm$ on $S_\varepsilon^\pm$ from Proposition 2.2 imply that there exists $\delta^* > 0$, such that the sum over the differences of the boundary values and of the initial values in (5.17) is smaller than $\frac{\rho}{3}$ for

$$\left| \left[\frac{\bar{\xi}}{\varepsilon}\right] \varepsilon + k\varepsilon \right| < \delta^*.$$

This condition is fulfilled for all $\bar{\xi} = (\xi_1, \dots, \xi_{n-1}) \in \mathbb{R}^{n-1}$, such that $|\xi_i| < \frac{\delta^*}{2\sqrt{n-1}}$ for $i = 1, \dots, n-1$ and all $\varepsilon < \frac{\delta^*}{2\sqrt{n-1}}$.

Now, we set $\delta_2 = \min\{\delta_1, \frac{\delta^*}{2\sqrt{n-1}}\}$. Then, for all $\bar{\xi}$ with $|\bar{\xi}| < \delta_2$ and all $\varepsilon < \varepsilon_2 = \min\{\varepsilon_0, \varepsilon_1, \frac{\delta^*}{2\sqrt{n-1}}, \sqrt{\frac{\rho}{3C\cdot 2^{n-1}}}h\}$, we have that

$$||L_\varepsilon \hat{u}_\varepsilon^M(t, \bar{x} + \bar{\xi}, y) - L_\varepsilon \hat{u}_\varepsilon^M(t, \bar{x}, y)||_{L^2((0,T)\times\Sigma\times Z, \mathbb{R}^m)} < \rho. \tag{5.18}$$

For $\varepsilon \in (\varepsilon_2, \varepsilon_0)$, the estimate (5.16) holds for every ε if $|\bar{\xi}| < \delta(\varepsilon)$ due to the continuity in the mean of L^2-functions. Since we consider sequences ε of the form $\varepsilon_k = \frac{1}{k}$, $k \in \mathbb{N}$, there are finitely many elements ε_k in the interval $(\varepsilon_2, \varepsilon_0)$. Thus choosing

$$\delta = \min\left\{\delta_2, \delta(\varepsilon_k), \varepsilon_k \in (\varepsilon_2, \varepsilon_0)\right\},$$

the property (5.16) is proved.

In addition, for $L_\varepsilon u_\varepsilon^M$ the following conditions are satisfied:

$$||\nabla_y L_\varepsilon u_\varepsilon^M||_{L_2((0,T)\times\Sigma\times Z,\mathbb{R}^m)} = \sqrt{\varepsilon}||\nabla u_\varepsilon^M||_{L_2((0,T)\times\Omega_\varepsilon^M,\mathbb{R}^m)} \leq C, \tag{5.19}$$

$$||\partial_t L_\varepsilon u_\varepsilon^M||_{L_2((0,T)\times\Sigma\times Z,\mathbb{R}^m)} = \frac{1}{\sqrt{\varepsilon}}||\partial_t u_\varepsilon^M||_{L_2(0,T)\times\Omega_\varepsilon^M,\mathbb{R}^m)} \leq C. \tag{5.20}$$

The conditions (5.16), (5.19), and (5.20) imply that the Kolmogorov criterion for $L_\varepsilon u_\varepsilon^M$ holds true in $L_2((0,T)\times\Sigma\times Z,\mathbb{R}^m)$. This concludes the proof of our theorem. $\square$

LEMMA 5.1. *For $u_\varepsilon \in H^1(\Omega_\varepsilon^M)$, the following trace estimate holds:*

$$||u_\varepsilon||_{L^2(S_\varepsilon^+\cup S_\varepsilon^-)} \leq C\left(\frac{1}{\sqrt{\varepsilon}}||u_\varepsilon||_{L^2(\Omega_\varepsilon^M)} + \sqrt{\varepsilon}||\nabla u_\varepsilon||_{L^2(\Omega_\varepsilon^M)}\right). \tag{5.21}$$

Proof. The proof follows by a scaling argument and by the standard trace estimate for H^1-functions. $\square$

6. Derivation of the macroscopic model. Using the convergence results proved in section 5, we are able to pass to the limit in the weak formulation of the microscopic problem.

6.1. Derivation of the equations in the bulk. First, we derive the macroscopic problem satisfied by the limit functions $u_0^\pm$.

Proof. Let us consider test functions $\varphi^\pm \in C_0^\infty((0,T)\times\Omega,\mathbb{R}^m)$ with

$$\operatorname{supp}\varphi^\pm \subset (0,T)\times\Omega^\pm.$$

Choose ε_0 such that

$$\min\left\{\operatorname{dist}\{\Sigma, \operatorname{supp}\varphi^+\},\ \operatorname{dist}\{\Sigma, \operatorname{supp}\varphi^-\}\right\} \geq \varepsilon_0. \tag{6.1}$$

Then for every $\varepsilon < \varepsilon_0$ we have

$$\operatorname{supp}\varphi^\pm \cap \Omega_\varepsilon^M = \emptyset.$$

Testing now (2.16) with φ^+ and φ^-, we obtain

$$\begin{aligned}&\int_0^T\int_{\Omega_\varepsilon^\pm}\partial_t u_{j\varepsilon}^\pm(t,x)\varphi_j^\pm(t,x)dx\,dt + D_j^\pm\int_0^T\int_{\Omega_\varepsilon^\pm}\nabla u_{j\varepsilon}^\pm(t,x)\nabla\varphi_j^\pm(t,x)dx\,dt\\ &= \int_0^T\int_{\Omega_\varepsilon^\pm} f_j(x,u_\varepsilon^\pm(t,x))\varphi_j^\pm(t,x)dx\,dt.\end{aligned} \tag{6.2}$$

For $\varepsilon \to 0$ the terms on the left-hand side converge due to the weak compactness results from Proposition 2.1. The convergence of the right-hand side follows due to the following argument: By Proposition 2.1

$$\chi_{\Omega_\varepsilon^\pm} u_{j\varepsilon} \to u_{j0}^\pm \quad \text{strongly in } L^2((0,T), L^2(\Omega^\pm)),$$

and there exists a subsequence, again denoted by $u_\varepsilon^\pm$, such that

$$\chi_{\Omega_\varepsilon^\pm} u_{j\varepsilon} \to u_{j0}^\pm \quad \text{a.e. in } (0,T) \times \Omega^\pm.$$

Then the continuity of f implies

$$f(\cdot, \chi_{\Omega_\varepsilon^\pm} u_\varepsilon(\cdot,\cdot)) \to f(\cdot, u_0^\pm(\cdot,\cdot)) \quad \text{a.e. in } (0,T) \times \Omega^\pm.$$

On the other hand, from the growth conditions (2.5) and the a priori estimates on $u_\varepsilon^\pm$, we have

$$\begin{aligned} &||f_j(\cdot, \chi_{\Omega_\varepsilon^\pm} u_\varepsilon(\cdot,\cdot))||^2_{L^2((0,T)\times\Omega^\pm)} = \int_0^T \int_{\Omega^\pm} |f_j(x, \chi_{\Omega_\varepsilon^\pm} u_\varepsilon(t,x))|^2 dx\, dt \\ &\leq \int_0^T \int_{\Omega^\pm} C(1 + |\chi_{\Omega_\varepsilon^\pm} u_\varepsilon(t,x)|^2) dx\, dt \leq C. \end{aligned}$$

Thus,

$$f_j(\cdot, \chi_{\Omega_\varepsilon^\pm} u_\varepsilon(\cdot,\cdot)) \to f_j(\cdot, u_0^\pm(\cdot,\cdot)) \quad \text{weakly in } L^2((0,T)\times\Omega^\pm).$$

Now, taking the limit $\varepsilon \to 0$ in (6.2), we obtain

$$\begin{aligned} &\int_0^T \int_{\Omega^\pm} \partial_t u_{j0}^\pm(t,x) \varphi_j^\pm(t,x) dx\, dt + D_j^\pm \int_0^T \int_{\Omega^\pm} \nabla u_{j0}^\pm(t,x) \nabla \varphi_j^\pm(t,x) dx\, dt \\ &= \int_0^T \int_{\Omega^\pm} f_j(x, u_0^\pm(t,x)) \varphi_j^\pm(t,x) dx\, dt, \end{aligned}$$

which is exactly the variational formulation for equations (2.23) and (2.24). The Dirichlet boundary condition can be deduced very easily, since $u_\varepsilon^\pm|_{\partial_D\Omega^\pm} = u_D^\pm|_{\partial_D\Omega^\pm}$. The Neumann boundary conditions are obtained by testing (2.16) with test functions $\varphi^\pm \in C_0^\infty((0,T), C^\infty(\Omega, \mathbb{R}^m))$ with $\operatorname{supp} \varphi^\pm \subset (0,T) \times \{\Omega^\pm \cup \partial_N \Omega^\pm\}$ and by repeating the arguments from the first part of the proof.

It remains to derive the initial conditions for $u_0^\pm$. Thus, let us consider $\varphi^\pm \in C_0^\infty(\Omega^\pm, \mathbb{R}^m)$ and again choose $\varepsilon_0 > 0$ as in (6.1). Then let $\xi \in C^\infty([0,T])$, such that $\xi(T) = 0$. Then for every $\varepsilon < \varepsilon_0$ we have

$$\begin{aligned} \int_0^T \int_{\Omega_\varepsilon^\pm} \partial_t u_{j\varepsilon}^\pm(t,x) \varphi_j^\pm(x) \xi(t)\, dx\, dt = &- \int_{\Omega_\varepsilon^\pm} U_{j0}(x) \varphi_j^\pm(x) \xi(0)\, dx \\ &- \int_0^T \int_{\Omega_\varepsilon^\pm} u_{j\varepsilon}^\pm(t,x) \varphi_j^\pm(x) \partial_t \xi(t)\, dx\, dt. \end{aligned}$$

Proposition 2.1 and the choice of our test function allow us to pass to the limit for $\varepsilon \to 0$ and to obtain

$$\begin{aligned} \int_0^T \int_{\Omega^\pm} \partial_t u_{j0}^\pm \varphi_j^\pm(x) \xi(t)\, dx\, dt = &- \int_{\Omega^\pm} U_{j0}(x) \varphi_j^\pm(x) \xi(0)\, dx \\ &- \int_0^T \int_{\Omega^\pm} u_{j0}^\pm(t,x) \varphi_j^\pm(x) \partial_t \xi(t)\, dx\, dt, \end{aligned}$$

which is equivalent to the initial conditions (2.27). $\square$

6.2. Derivation of the local equations in the layer. We now derive the local problem for the limit function u_0^M which enters the transmission conditions.

Proof. We start from the weak formulation (2.16) and use as test function

$$\varphi^\varepsilon(t,x) = \begin{cases} 0, \ (t,x) \in (0,T) \times (\Omega_\varepsilon^+ \cup \Omega_\varepsilon^-), \\ \varphi(t,\bar{x},\frac{x}{\varepsilon}), \ (t,x) \in (0,T) \times \Omega_\varepsilon^M, \end{cases}$$

where $\varphi \in \left\{ C_0^\infty((0,T)\times\Sigma, C_{per}^\infty(Y, C_0^\infty(]-1,1[))) \right\}^m$. Thus, we obtain

$$\begin{aligned} &\frac{1}{\varepsilon}\int_0^T \int_{\Omega_\varepsilon^M} \partial_t u_{j\varepsilon}^M \, \varphi_j\left(t,\bar{x},\frac{x}{\varepsilon}\right) dx\,dt \\ &+\frac{1}{\varepsilon}\int_0^T \int_{\Omega_\varepsilon^M} D_j^M\left(\frac{x}{\varepsilon}\right) \varepsilon \nabla u_{j\varepsilon}^M \left(\varepsilon \nabla_{\bar{x}}\varphi_j\left(t,\bar{x},\frac{x}{\varepsilon}\right) + \nabla_y \varphi_j\left(t,\bar{x},\frac{x}{\varepsilon}\right)\right) dx\,dt \\ &= \frac{1}{\varepsilon}\int_0^T \int_{\Omega_\varepsilon^M} g_j\left(\frac{x}{\varepsilon}, u_\varepsilon^M\right)\varphi_j\left(t,\bar{x},\frac{x}{\varepsilon}\right) dx\,dt. \end{aligned} \tag{6.3}$$

Using the weak two-scale convergence properties from Proposition 2.1, we can pass to the limit for $\varepsilon \to 0$ on the left-hand side of the equation above. In order to pass to the limit in the nonlinear term on the right-hand side, we have to use the strong two-scale convergence given in Theorem 2.3 to show that

$$g_j\left(\frac{\cdot}{\varepsilon}, u_\varepsilon^M\right) \overset{t.s.}{\to} g_j(y, u_0^M) \quad \text{weakly in the two-scale sense.}$$

Thus, let us consider

$$\begin{aligned} L_\varepsilon\left[g_j\left(\frac{\cdot}{\varepsilon}, u_\varepsilon^M\right)\right](t,\bar{x},y) &= g_j\left(\frac{(c_\varepsilon(\bar{x})+\varepsilon\bar{y},\varepsilon y_n)}{\varepsilon}, u_\varepsilon^M(t,(c_\varepsilon(\bar{x})+\varepsilon\bar{y},\varepsilon y_n))\right) \\ &= g_j(y, L_\varepsilon u_\varepsilon^M(t,\bar{x},y)). \end{aligned}$$

Since g is continuous and $L_\varepsilon u_\varepsilon^M$ converges strongly to u_0^M in $L^2((0,T)\times\Sigma\times Z,\mathbb{R}^m)$, we have that

$$g_j(\cdot, L_\varepsilon u_\varepsilon^M) \longrightarrow g_j(\cdot, u_0^M) \quad \text{a.e. in } (0,T)\times\Sigma\times Z.$$

From the growth conditions (2.6) and the a priori estimates on u_ε^M we obtain

$$\begin{aligned} ||g_j(\cdot, L_\varepsilon u_\varepsilon^M)||_{L^2((0,T)\times\Sigma\times Z)} &= \int_0^T \int_\Sigma \int_Z |g_j(y, u_\varepsilon^M(t,(c_\varepsilon(\bar{x})+\varepsilon\bar{y},\varepsilon y_n)))|^2 dy d\bar{x} dt \\ &\le C \int_0^T \int_\Sigma \int_Z 1 + |u_\varepsilon^M(t,(c_\varepsilon(\bar{x})+\varepsilon\bar{y},\varepsilon y_n))|^2 dy d\bar{x} dt \\ &\le C_1 + \frac{C_2}{\varepsilon}||u_\varepsilon^M||^2_{L^2((0,T)\times\Omega_\varepsilon^M,\mathbb{R}^m)} \le C. \end{aligned} \tag{6.4}$$

Thus,

$$g_j(\cdot, L_\varepsilon u_\varepsilon^M) \longrightarrow g_j(\cdot, u_0^M) \quad \text{weakly in } L^2((0,T)\times\Sigma\times Z).$$

This now implies that

$$L_\varepsilon\left[g_j\left(\frac{\cdot}{\varepsilon}, u_\varepsilon^M\right)\right] \to g_j(\cdot, u_0^M) \quad \text{weakly in } L^2((0,T)\times\Sigma\times Z),$$

and by Theorem 4.7 it follows that

$$g_j\left(\frac{\cdot}{\varepsilon}, u_\varepsilon^M\right) \stackrel{t.s.}{\to} g_j(y, u_0^M) \quad \text{weakly in the two-scale sense.}$$

Thus, we can pass to the limit in (6.3) and obtain

$$\begin{aligned} \int_0^T \int_\Sigma \int_Z \partial_t u_{j0}^M(t,\bar{x},y)\varphi_j(t,\bar{x},y)\,dy\,d\bar{x}\,dt \\ + \int_0^T \int_\Sigma \int_Z D_j^M(y)\nabla_y u_{j0}^M(t,\bar{x},y)\nabla_y\varphi_j(t,\bar{x},y)\,dy\,d\bar{x}\,dt \\ = \int_0^T \int_\Sigma \int_Z g_j(y, u_0^M(t,\bar{x},y))\varphi_j(t,\bar{x},y)\,dy\,d\bar{x}\,dt. \end{aligned} \tag{6.5}$$

This is just the weak formulation of (2.30). The boundary conditions have already been proved in Proposition 2.1; see (2.19). It remains to prove the initial condition for u_0^M. We will proceed as in section 6.1.

Let $\varphi \in \{C_0^\infty(\Sigma, C_{per}(Y, C_0^\infty(]-1,1[)))\}^m$ and $\xi \in C^\infty([0,T])$ such that $\xi(T) = 0$. Then the following relation holds:

$$\begin{aligned} \frac{1}{\varepsilon}\int_0^T \int_{\Omega_\varepsilon^M} \partial_t u_{j\varepsilon}^M(t,x)\varphi_j\left(\bar{x},\frac{x}{\varepsilon}\right)\xi(t)\,dx\,dt = -\frac{1}{\varepsilon}\int_{\Omega_\varepsilon^M} U_{j0}^M\left(\bar{x},\frac{x_n}{\varepsilon}\right)\varphi_j\left(\bar{x},\frac{x}{\varepsilon}\right)\xi(0)\,dx \\ -\frac{1}{\varepsilon}\int_0^T \int_{\Omega_\varepsilon^M} u_{j\varepsilon}^M(t,x)\varphi_j\left(\bar{x},\frac{x}{\varepsilon}\right)\partial_t\xi(t)\,dx\,dt. \end{aligned}$$

Using the convergence results from Proposition 2.1 together with Lemma 4.3, we can pass to the limit for $\varepsilon \to 0$ and obtain

$$\begin{aligned} \int_0^T \int_\Sigma \int_Z \partial_t u_{j0}^M(t,\bar{x},y)\varphi(\bar{x},y)\xi(t)\,dy\,d\bar{x}\,dt = -\int_\Sigma \int_Z U_{j0}^M(\bar{x},y_n)\varphi_j(\bar{x},y)\xi(0)\,dy\,d\bar{x} \\ -\int_0^T \int_\Sigma \int_Z u_{j0}^M(t,\bar{x},y)\varphi_j(\bar{x},y)\partial_t\xi(t)\,dy\,d\bar{x}\,dt, \end{aligned}$$

which immediately implies the initial conditions (2.33) for the local problem. ∎

6.3. Derivation of the transmission conditions across Σ. Measurements and experiments indicate jumps of the concentrations and their normal derivatives across the interface Σ. We derive formulas for these jumps which involve the solutions of the local cell problems. Thus, the jumps become computable.

DEFINITION 6.1. *For $u \in L^2((0,T)\times\Omega,\mathbb{R}^m)$ with $u^\pm \in L^2((0,T), H^1(\Omega^\pm,\mathbb{R}^m))$, the jump of u across Σ is defined as*

$$[u]_\Sigma(t,\bar{x}) = u^+(t,\bar{x},0) - u^-(t,\bar{x},0), \quad (t,\bar{x}) \in (0,T)\times\Sigma.$$

To find a relation for the jump in the concentrations, we use the boundary layer function constructed below.

Consider the Hilbert space

$$V = \{\varphi \in H^1(Z,\mathbb{R}^m), \varphi \text{ periodic in } Y, \varphi \equiv \text{ const on } S^+ \cup S^-\}.$$

Find $\eta \in V$ such that

$$\frac{1}{|Z|}\int_Z \eta(y)\,dy = 0 \tag{6.6}$$

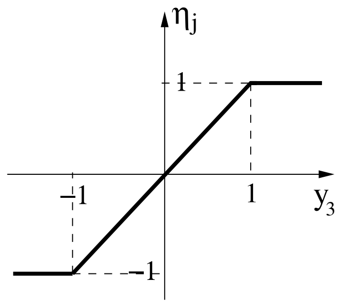


FIG. 6.1. *The boundary layer $\eta_j(y)$ for $D_j(y) \equiv 1$.*

and, for all $\varphi \in V$,

$$\int_Z D_j^M(y)\nabla\eta_j(y)\nabla\varphi_j(y)\,dy = \int_{S^+}\varphi_j(y)\,ds - \int_{S^-}\varphi_j(y)\,ds. \tag{6.7}$$

Since the right-hand side of (6.7) defines a linear, continuous functional on V, the Lax–Milgram lemma implies the existence of a unique solution $\eta \in V$ to problem (6.7) with (6.6). The strong formulation for the problem for η is given as follows: Find $\eta \in V$ with $\frac{1}{|Z|}\int_Z \eta(y)\,dy = 0$ such that

$$\nabla_y(D_j^M(y)\nabla_y\eta_j(y)) = 0, \quad y \in Z, \tag{6.8}$$

$$\int_{S^+} D_j^M(y)\partial_n\eta_j(y)\,dy = 1, \tag{6.9}$$

$$-\int_{S^-} D_j^M(y)\partial_n\eta_j(y)\,dy = -1. \tag{6.10}$$

In the case $D_j^M(y) \equiv 1$, the function η is given by (see Figure 6.1)

$$\eta_j(y) = y_n, \quad j = 1, \dots, m.$$

We denote

$$\eta^\pm \in \mathbb{R}^m \text{ the constant values of } \eta \text{ on } S^\pm. \tag{6.11}$$

The derivation of the transmission condition is now possible by using a test function of boundary layer type as shown below.

Proof. Consider the function

$$\eta_\varepsilon(x) = \begin{cases} \eta^+, & x \in \Omega_\varepsilon^+, \\ \eta\left(\frac{x}{\varepsilon}\right), & x \in \Omega_\varepsilon^M, \\ \eta^-, & x \in \Omega_\varepsilon^-, \end{cases} \tag{6.12}$$

and let $\xi(t, \bar{x})$ be a smooth function, $\xi \in C_0^\infty((0,T) \times \Sigma, \mathbb{R}^m)$. To get the jump condition for u_{j0} across Σ, we evaluate the integral

$$I_\varepsilon = \int_0^T \int_{\Omega_\varepsilon^M} \varepsilon D_j^M\left(\frac{x}{\varepsilon}\right)\nabla\eta_{j\varepsilon}(x)\nabla(u_{j\varepsilon}^M(t,x)\xi_j(t,\bar{x}))dx\,dt \tag{6.13}$$

in two different ways and then pass to the limit for $\varepsilon \to 0$. First, integrating by parts and taking into account (6.8) and the continuity of u_ε across S_ε^+ and S_ε^-, we obtain

$$\begin{aligned} I_\varepsilon = & \int_0^T \int_{S_\varepsilon^+} \varepsilon D_j^M \left(\frac{x}{\varepsilon}\right) \frac{1}{\varepsilon} \nabla_y \eta_j \left(\frac{x}{\varepsilon}\right) \cdot \vec{\nu}(x)\, u_{j\varepsilon}^+(t,x) \xi_j(t,\bar{x}) dx\, dt \\ & + \int_0^T \int_{S_\varepsilon^-} \varepsilon D_j^M \left(\frac{x}{\varepsilon}\right) \frac{1}{\varepsilon} \nabla_y \eta_j \left(\frac{x}{\varepsilon}\right) \cdot \vec{\nu}(x)\, u_{j\varepsilon}^-(t,x) \xi_j(t,\bar{x}) dx\, dt. \end{aligned}$$

Now, Proposition 2.2 allows us to pass to the limit for $\varepsilon \to 0$ and to obtain

$$\begin{aligned} \lim_{\varepsilon \to 0} I_\varepsilon &= \int_0^T \int_\Sigma \int_Y u_{j0}^+(t,\bar{x},0) \xi_j(t,\bar{x}) D_j^M(\bar{y},1) \partial_n \eta_j(\bar{y},1) d\bar{y}\, d\bar{x}\, dt \\ & \quad - \int_0^T \int_\Sigma \int_Y u_{j0}^-(t,\bar{x},0) \xi_j(t,\bar{x}) D_j^M(\bar{y},-1) \partial_n \eta_j(\bar{y},-1) d\bar{y}\, d\bar{x}\, dt \\ &= \int_0^T \int_\Sigma (u_{j0}^+(t,\bar{x},0) - u_{j0}^-(t,\bar{x},0)) \xi_j(t,\bar{x}) d\bar{x}\, dt, \end{aligned}$$

where the last equality follows from the boundary conditions (6.9) and (6.10) for η.

Second, by differentiation in the second gradient, we have

$$\begin{aligned} I_\varepsilon &= \int_0^T \int_{\Omega_\varepsilon^M} \varepsilon D_j^M \left(\frac{x}{\varepsilon}\right) \nabla \eta_{j\varepsilon}(x) \nabla \xi_j(t,\bar{x}) u_{j\varepsilon}^M(t,x) dx\, dt \\ & \quad + \int_0^T \int_{\Omega_\varepsilon^M} \varepsilon D_j^M \left(\frac{x}{\varepsilon}\right) \nabla(\eta_{j\varepsilon}(x) \xi_j(t,\bar{x})) \nabla u_{j\varepsilon}^M(t,x) dx\, dt \\ & \quad - \int_0^T \int_{\Omega_\varepsilon^M} \varepsilon D_j^M \left(\frac{x}{\varepsilon}\right) \eta_{j\varepsilon}(x) \nabla \xi_j(t,\bar{x}) \nabla u_{j\varepsilon}^M(t,x) dx\, dt \\ &= I_\varepsilon^1 + I_\varepsilon^2 + I_\varepsilon^3. \end{aligned}$$

Since u_ε^M and $\varepsilon \nabla u_\varepsilon^M$ converge weakly in the two-scale sense (see Proposition 2.1), the integrals I_ε^1 and I_ε^3 converge to zero for $\varepsilon \to 0$. To calculate $\lim_{\varepsilon \to 0} I_\varepsilon^2$, we start from the weak formulation (2.16) where we insert as test function

$$\varphi(t,x) = \eta_\varepsilon(x) \xi(t,\bar{x}) \psi(x_n) \tag{6.14}$$

with $\psi \in C_0^\infty((-H,H),\mathbb{R})$, such that $\psi(x_n) \equiv 1$ for $|x_n| < \frac{H}{2}$. Thus, we have

$$\begin{aligned} \lim_{\varepsilon \to 0} I_\varepsilon^2 = & \lim_{\varepsilon \to 0} \int_0^T \int_{\Omega_\varepsilon^+} (f_j(x,u_\varepsilon^+) - \partial_t u_{j\varepsilon}^+) \eta_j^+ \xi_j(t,\bar{x}) \psi(x_n) dx\, dt \\ & + \lim_{\varepsilon \to 0} \int_0^T \int_{\Omega_\varepsilon^-} (f_j(x,u_\varepsilon^-) - \partial_t u_{j\varepsilon}^-) \eta_j^- \xi_j(t,\bar{x}) \psi(x_n) dx\, dt \\ & + \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_0^T \int_{\Omega_\varepsilon^M} \left(g_j\left(\frac{x}{\varepsilon}, u_\varepsilon^M\right) - \partial_t u_{j\varepsilon}^M\right) \eta_j \left(\frac{x}{\varepsilon}\right) \xi_j(t,\bar{x}) dx\, dt \\ & - \lim_{\varepsilon \to 0} \int_0^T \int_{\Omega_\varepsilon^+} D_j^+ \eta_j^+ \nabla u_{j\varepsilon}^+ \nabla(\xi_j(t,\bar{x}) \psi(x_n)) dx\, dt \\ & - \lim_{\varepsilon \to 0} \int_0^T \int_{\Omega_\varepsilon^-} D_j^- \eta_j^- \nabla u_{j\varepsilon}^- \nabla(\xi_j(t,\bar{x}) \psi(x_n)) dx\, dt. \end{aligned}$$

Using the convergence properties from Proposition 2.1, Theorem 2.3, and the macroscopic problem for the limit functions u_0^+ and u_0^-, we obtain

$$\lim_{\varepsilon\to 0} I_\varepsilon^2 = \int_0^T \int_\Sigma \int_Z (g_j(y, u_0^M(t,\bar{x},y)) - \partial_t u_{j0}^M(t,\bar{x},y))\eta_j(y)\,dy\,\xi_j(t,\bar{x})\,d\bar{x}\,dt$$
$$+ \int_0^T \int_\Sigma (D_j^+\eta_j^+\partial_n u_{j0}^+(t,\bar{x},0) - D_j^-\eta_j^-\partial_n u_{j0}^-(t,\bar{x},0))\xi_j(t,\bar{x})\,d\bar{x}\,dt.$$

Thus, the first transmission condition is proved. It remains to derive a transmission condition for the fluxes. We start again from the weak formulation (2.16) and use as test function

$$\varphi(t,x) = \xi(t,\bar{x})\psi(x_n),$$

with ξ and ψ as in (6.14). For $\varepsilon \to 0$ we obtain

$$\int_0^T \int_{\Omega^+} \partial_t u_{j0}^+ \xi_j\psi_j dx\,dt + \int_0^T \int_{\Omega^+} D_j^+\nabla u_{j0}^+ \nabla(\xi_j\psi_j) dx\,dt - \int_0^T \int_{\Omega^+} f_j(x,u_0^+)\xi_j\psi_j dx\,dt$$
$$+ \int_0^T \int_{\Omega^-} \partial_t u_{j0}^- \xi_j\psi_j dx\,dt + \int_0^T \int_{\Omega^-} D_j^-\nabla u_{j0}^- \nabla(\xi_j\psi_j) dx\,dt - \int_0^T \int_{\Omega^-} f_j(x,u_0^-)\xi_j\psi_j dx\,dt$$
$$+ \int_0^T \int_\Sigma \int_Z (\partial_t u_{j0}^M(t,\bar{x},y) - g_j(y,u_0^M(t,\bar{x},y)))\,dy\,\xi_j(t,\bar{x})\,d\bar{x}\,dt = 0.$$

Using again the macroscopic equations for u_0^+ and u_0^-, we get the following transmission condition:

$$\int_0^T \int_\Sigma (D_j^+\partial_n u_{j0}^+(t,\bar{x},0) - D_j^-\partial_n u_{j0}^-(t,\bar{x},0))\xi_j(t,\bar{x})$$
$$= \int_0^T \int_\Sigma \int_Z (\partial_t u_{j0}^M(t,\bar{x},y) - g_j(y,u_0^M(t,\bar{x},y)))\,dy\,\xi_j(t,\bar{x})\,d\bar{x}\,dt. \qquad \square$$

7. Uniqueness for the macroscopic model. In this section, we want to show uniqueness of the solutions (u_0^+, u_0^-, u_0^M) to the effective model presented in Theorem 2.4. To this end, let us first formulate a corollary which gives an equivalent formulation of the transmission conditions (2.28), (2.29).

COROLLARY 7.1. *The transmission conditions* (2.28), (2.29) *are equivalent to*

$$D_j^+\partial_n u_{j0}^+(t,\bar{x},0) = \int_{S^+} D_j^M(y)\partial_n u_{j0}^M(t,\bar{x},y)dy, \tag{7.1}$$

$$D_j^-\partial_n u_{j0}^-(t,\bar{x},0) = \int_{S^-} D_j^M(y)\partial_n u_{j0}^M(t,\bar{x},y)dy. \tag{7.2}$$

These equivalent transmission conditions hold in a distributional sense with respect to t and $\bar{x}$.

REMARK 2. *The physical interpretation of* (7.1) *and* (7.2) *is obvious; it states that the macroscopic flux is given by the microscopic flux averaged over the corresponding part of the cell surface.*

Proof. Using the properties (6.8), (6.9), and (6.10) of the boundary layer η and the boundary conditions

$$u_0^M(t,\bar{x},y) = u_0^\pm(t,\bar{x},0) \quad \text{on } (0,T)\times\Sigma\times S^\pm,$$

we obtain

$$
\begin{aligned}
[u_{j0}]|_\Sigma(t,\bar{x}) &= u_{j0}^+(t,\bar{x},0) - u_{j0}^-(t,\bar{x},0) \\
&= \int_{S^+} u_{j0}^M(t,\bar{x},y)dy - \int_{S^-} u_{j0}^M(t,\bar{x},y)dy \\
&= \int_Z D_j^M(y)\nabla\eta_j(y)\nabla u_{j0}^M(t,\bar{x},y)dy.
\end{aligned}
$$

Equation (2.30) for u_0^M and the fact that η is constant on S^+ and S^- with constants defined in (6.11) yield

$$
\begin{aligned}
[u_{j0}]_\Sigma(t,\bar{x}) = &\int_Z (g_j(y,u_0^M(t,\bar{x},y)) - \partial_t u_{j0}^M(t,\bar{x},y))\eta_j(y)\,dy \\
&+ \int_{S^+} D_j^M(y)\partial_n u_{j0}^M(t,\bar{x},y)\eta_j^+ dy - \int_{S^-} D_j^M(y)\partial_n u_{j0}^M(t,\bar{x},y)\eta_j^- dy.
\end{aligned} \tag{7.3}
$$

From (2.28) and (7.3), we then obtain

$$
\begin{aligned}
&D_j^+\partial_n u_{j0}^+(t,\bar{x},0)\eta_j^+ - D_j^-\partial_n u_{j0}^-(t,\bar{x},0)\eta_j^- \\
&= \int_{S^+} D_j^M(y)\partial_n u_{j0}^M(t,\bar{x},y)\eta_j^+ dy - \int_{S^-} D_j^M(y)\partial_n u_{j0}^M(t,\bar{x},y)\eta_j^- dy.
\end{aligned} \tag{7.4}
$$

On the other hand, using (2.30) for u_0^M, the transmission condition (2.29) transforms to

$$
\begin{aligned}
&D_j^+\partial_n u_{j0}^+(t,\bar{x},0) - D_j^-\partial_n u_{j0}^-(t,\bar{x},0) \\
&= \int_{S^+} D_j^M(y)\partial_n u_{j0}^M(t,\bar{x},y)dy - \int_{S^-} D_j^M(y)\partial_n u_{j0}^M(t,\bar{x},y)dy.
\end{aligned} \tag{7.5}
$$

Relations (7.4) and (7.5) are equivalent with the transmission conditions (7.1) and (7.2) since $\eta_j^+ - \eta_j^- \neq 0$. This follows from the relations

$$\eta_j^+ > 0 \ \text{ and } \ \eta_j^- < 0 \quad \text{for } j = 1, \dots n,$$

which hold by the maximum principle for elliptic equations. ∎

In the following, we prove uniqueness for the macroscopic system by using the equivalent transmission conditions (7.1) and (7.2).

Proof. Assume that (u_i^+, u_i^-, u_i^M), $i = 1, 2$, are solutions of the macroscopic system with the same data. Let δu^+, δu^-, δu^M denote the differences. Now, we consider a test function $\varphi \in V$, where

$$V = \{(C_0^\infty((0,T), C^\infty(\overline{\Omega^+})))^m \cap (C_0^\infty((0,T), C^\infty(\overline{\Omega^-})))^m, \varphi = 0 \,\text{on}\, \partial_D\Omega^+ \cup \partial_D\Omega^-\}.$$

We multiply (2.23) and (2.24) by φ and integrate, obtaining

$$
\begin{aligned}
&\int_{\Omega^+} \partial_t u_{j0}^+ \varphi_j^+ \,dx + \int_{\Omega^-} \partial_t u_{j0}^- \varphi_j^- \,dx + D_j^+ \int_{\Omega^+} \nabla u_{j0}^+ \nabla \varphi_j^+ \,dx \\
&+ D_j^- \int_{\Omega^-} \nabla u_{j0}^- \nabla \varphi_j^- \,dx + \int_\Sigma D_j^+ \partial_n u_{j0}^+ \varphi_j^+ d_\Sigma(x) - \int_\Sigma D_j^- \partial_n u_{j0}^- \varphi_j^- d_\Sigma(x) \\
&= \int_{\Omega^+} f_j(x, u_0^+)\varphi_j^+ \,dx + \int_{\Omega^-} f_j(x, u_0^-)\varphi_j \,dx.
\end{aligned} \tag{7.6}
$$

From (7.6) and the transmission conditions (7.1) and (7.2), we obtain the following equation for the differences $\delta u^+, \delta u^-$:

$$\begin{aligned} &\int_{\Omega^+} \partial_t(\delta u_j^+)\varphi_j^+\,dx + \int_{\Omega^-} \partial_t(\delta u_j^-)\varphi_j^-\,dx + D_j^+ \int_{\Omega^+} \nabla(\delta u_j^+)\nabla\varphi_j^+\,dx \\ &+ D_j^- \int_{\Omega^-} \nabla(\delta u_j^-)\nabla\varphi_j^-\,dx + \int_\Sigma \int_{S^+} D_j^M(y)\partial_n(\delta u_j^M)dy\,\varphi_j^+\,d_\Sigma(x) \\ &- \int_\Sigma \int_{S^-} D_j^M \partial_n(\delta u_j^M)dy\,\varphi_j^-\,d_\Sigma(x) = \int_{\Omega^+} \delta f_j^+ \varphi_j^+\,dx + \int_{\Omega^-} \delta f_j^- \varphi_j\,dx, \end{aligned} \tag{7.7}$$

where we denoted

$$\delta f_j^+ = f_j(x, u_1^+) - f_j(x, u_2^+), \quad \delta f_j^- = f_j(x, u_1^-) - f_j(x, u_2^-).$$

Now, we insert $\varphi^+ = \delta u^+$ and $\varphi^- = \delta u^-$ as test functions. Using the boundary conditions (2.31), we get for the terms on Σ

$$\begin{aligned} &\int_\Sigma \int_{S^+} D_j^M(y)\partial_n(\delta u_j^M)dy\,\delta u_j^+ d_\Sigma - \int_\Sigma \int_{S^-} D_j^M \partial_n(\delta u_j^M)dy\,\delta u_j^- d_\Sigma \\ &= \int_\Sigma \int_{S^+} D_j^M(y)\partial_n(\delta u_j^M)\delta u_j^M dy d_\Sigma - \int_\Sigma \int_{S^-} D_j^M \partial_n(\delta u_j^M)\,\delta u_j^M dy d_\Sigma \\ &= \int_\Sigma \int_Z D_j^M(y)\nabla(\delta u_j^M)\nabla(\delta u_j^M)dy d_\Sigma + \int_\Sigma \int_Z \partial_t(\delta u_j^M)\delta u_j^M dy d_\Sigma \\ &\quad - \int_\Sigma \int_Z \delta g_j \delta u_j^M dy d_\Sigma. \end{aligned} \tag{7.8}$$

Here, we used the notation $\delta g_j = g_j(y, u_1^M) - g_j(y, u_2^M)$. Inserting (7.8) in (7.7), summing up for $j = 1, \ldots, m$, and integrating with respect to t, we obtain

$$\begin{aligned} &\frac{1}{2}||\delta u^+(t)||^2_{L^2(\Omega^+)} + \frac{1}{2}||\delta u^-(t)||^2_{L^2(\Omega^-)} + \frac{1}{2}||\delta u^M(t)||^2_{L^2(\Sigma\times Z)} \\ &+ D_j^+ \int_0^t \int_{\Omega^+} |\nabla(\delta u^+)|^2\,dxdt + D_j^- \int_0^t \int_{\Omega^-} |\nabla(\delta u^-)|^2\,dxdt \\ &+ \int_0^t \int_\Sigma \int_Z D^M(y)|\nabla(\delta u^M)|^2 dy d_\Sigma dt = \int_0^t \int_\Sigma \int_Z \delta g \delta u^M dy d_\Sigma dt \\ &+ \int_0^t \int_{\Omega^+} \delta f^+ \delta u^+\,dxdt + \int_0^t \int_{\Omega^-} \delta f^- \delta u^-\,dxdt. \end{aligned} \tag{7.9}$$

Using the Lipschitz continuity of the reaction terms, the right-hand side in (7.9) can be estimated by

$$C\left(||\delta u^M||^2_{L^2((0,t)\times\Sigma\times Z)} + ||\delta u^+||^2_{L^2((0,t)\times\Omega^+)} + ||\delta u^-||^2_{L^2((0,t)\times\Omega^-)}\right).$$

Then Gronwall's inequality yields

$$\delta u^+ = \delta u^- = \delta u^M = 0$$

and the theorem is proved. ∎

COROLLARY 7.2. *The entire sequence* $(u_\varepsilon^+, u_\varepsilon^-, u_\varepsilon^M)$ *converges to the limit* (u_0^+, u_0^-, u_0^M)*, solving the macroscopic system.*

REMARK 3. *An important aim of our homogenization procedure is to reduce the computational complexity. The algorithms for solving the derived transmission problem numerically will be considered in a forthcoming paper.*

Acknowledgment. The authors appreciate the comments and suggestions of the referees and tried to include them as much as possible.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[2] G. Allaire and M. Briane, *Multi-scale convergence and reiterated homogenization*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 297–342.

[3] T. Arbogast, J. Douglas, Jr., and U. Hornung, *Derivation of the double porosity model of single phase flow via homogenization theory*, SIAM J. Math. Anal., 21 (1990), pp. 823–836.

[4] G. Bouchitté, G. Butazzo, and P. Seppecher, *Energies with respect to a measure and applications to low dimensional structures*, Calc. Var. Partial Differential Equations, 5 (1997), pp. 37–54.

[5] A. Bourgeat, S. Luckhaus, and A. Mikelić, *Convergence of the homogenization process for a double-porosity model of immiscible two-phase flow*, SIAM J. Math. Anal., 27 (1996), pp. 1520–1543.

[6] A. Bourgeat, A. Mikelić, and S. Wright, *Stochastic two-scale convergence in the mean and applications*, J. Reine Angew. Math., 456 (1994), pp. 19–51.

[7] R. Brizzi and J. P. Chalot, *Boundary homogenization and Neumann boundary value problem*, Ricerche Mat., 46 (1997), pp. 341–387.

[8] J. Casado-Díaz, M. Luna-Laynez, and J. D. Martín, *An adaptation of the multi-scale methods for the analysis of very thin reticulated structures*, C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 223–228.

[9] D. Ciorănescu, A. Damlamian, and G. Griso, *Periodic unfolding and homogenization*, C. R. Math. Acad. Sci. Paris I, 335 (2002), pp. 99–104.

[10] J. L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[11] S. Marušić and E. Marušić-Paloka, *Two-scale convergence for thin domains and its applications to some lower-dimensional models in fluid mechanics*, Asymptot. Anal., 23 (2000), pp. 23–57.

[12] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.

[13] M. Neuss-Radu, *Some extensions of two-scale convergence*, C. R. Acad. Sci. Paris Sér. I Math., 322 (1996), pp. 899–904.

[14] M. Neuss-Radu and W. Jäger, *Homogenization of thin porous layers and applications to ion transport through channels of biological membranes*, Oberwolfach Rep., 49 (2005), pp. 2809–2812.

[15] M. Neuss-Radu and W. Jäger, *A multi-scale approach to ion transport through channels*, in preparation.

[16] D. Onofrei, *The unfolding operator near a hyperplane and its applications to the Neumann sieve model*, Adv. Math. Sci. Appl., 23 (2006), pp. 239–258.

[17] V. V. Zhikov, *On an extension of the method of two-scale convergence and its applications*, Math. Sb., 191 (2000), pp. 973–1014.

SIAM J. MATH. ANAL.
Vol. 39, No. 3, pp. 721–741

UNIQUENESS THEOREMS FOR STABLE ANISOTROPIC CAPILLARY SURFACES*

MIYUKI KOISO† AND BENNETT PALMER‡

Abstract. We consider capillary surfaces for certain rotationally invariant elliptic parametric functionals supported on two hydrophobically wetted horizontal plates separated by a fixed distance. It is shown that each such stable capillary surface is uniquely determined by the volume interior to the surface.

Key words. anisotropic, capillary surface, uniqueness, stability

AMS subject classifications. 49Q10, 53C42, 76D45

DOI. 10.1137/060657297

1. Introduction.

When the temperature of a fluid is gradually lowered, it undergoes a process of crystallization in which its constituent atoms, molecules, or ions will align themselves in a regular repeating pattern. It is rare that a single crystal will form, and instead many crystals will form a polycrystal. This is the state in which, for example, most metals occur.

As the fluid cools, the usual isotropic surface energy (surface tension) will no longer be appropriate to model the shape of the interface of the fluid with its environment. Because of the internal structure of the material, the isotropic surface energy must be replaced by an anisotropic one, i.e., an energy that depends on the direction of the surface at each point. In this paper, we will treat a class of capillary problems for the simplest type of anisotropic surface energy: a constant coefficient, elliptic parametric functional.

Particularly, we consider a variational problem whose solution is a mathematical model of a drop of a cooled liquid trapped between two horizontal plates. The plates are hydrophobically wetted and are made of the same material. It is natural to consider the volume of the drop, the distance between the plates, and the wetting constant ω which couples the energy of the fluid-plate interface to the free surface energy as "initial data" and then ask if the shape of the drop is uniquely determined. In our previous paper [6], we obtained a geometric characterization of such drops. In this paper, we show (Theorem 2.1) that under certain assumptions on the energy functional the uniqueness follows if the additional natural condition of stability is imposed. Not only do we have uniqueness but we are able to determine the shape of the drop (Theorem 2.3) to the extent that a parameterization can be easily obtained from our previous work [4].

We wish to emphasize that we have restricted ourselves to the cases of hydrophobic wetting and equal contact angle. This is not to suggest that the other cases are of lesser importance. At present we also ignore gravitational and other external forces.

*Received by the editors April 14, 2006; accepted for publication (in revised form) March 13, 2007; published electronically August 17, 2007.
http://www.siam.org/journals/sima/39-3/65729.html
†Department of Mathematics, Nara Women's University, Nara 630-8506, Japan (koiso@cc.nara-wu.ac.jp). This author was partially supported by Grant-in-Aid for Scientific Research (C) 16540195 of the Japan Society for the Promotion of Science.
‡Department of Mathematics, Idaho State University, Pocatello, ID 83209 (palmbenn@isu.edu).

This paper should be considered as part of a program in which these more general problems will be considered.

Our assumptions imposed on the energy functional are satisfied by the usual area functional. In this important special case, solutions are constant mean curvature (CMC) surfaces which meet each of the supporting planes with constant angle. In the CMC case without wetting, the uniqueness and characterization of stable solutions follow from the results in Athanassenas [1] and Vogel [7]. For hydrophobic wetting, they follow from the results in Vogel [8] and Finn and Vogel [3]. The lower bound for the volume of a stable spanning drop of height h was shown by Finn and Vogel in [3] to be h^3/π, giving an affirmation of Carter's conjecture. We will generalize this result to the anisotropic case with hydrophobic wetting (Theorem 2.4).

The paper is organized as follows. Section 2 contains precise statements of our main results. Sections 3 and 4 will be devoted to proofs of the results stated in section 2. In section 5, we will give a strict examination of the uniqueness for the case without wetting energy. Section 6 contains a summary of results concerning anisotropic Delaunay surfaces (rotationally symmetric surfaces with constant anisotropic mean curvature). These surfaces were introduced in detail in [4] and play a fundamental role in our stability and uniqueness analysis.

2. Statements of results. Let F be a smooth, positive function on S^2. To an immersion $X : \Sigma \to \mathbf{R}^3$ from a two-dimensional oriented, connected, compact, smooth manifold Σ (possibly with boundary $\partial\Sigma$) to the three-dimensional Euclidean space $\mathbf{R}^3$, we assign the free anisotropic energy

$$\mathcal{F}[X] := \int_\Sigma F(\nu)\, d\Sigma\,, \tag{1}$$

where $\nu = (\nu_1, \nu_2, \nu_3) : \Sigma \to S^2$ is the Gauss map of X, and $d\Sigma$ is the area form of the induced metric. We will assume that F satisfies a "convexity condition" in the following sense: Denote by DF and D^2F the gradient and Hessian of F on S^2. We assume that at each point in S^2 the matrix $D^2F + FI$ is positive definite, where I is the identity endomorphism field on TS^2. Such an energy functional $\mathcal{F}$ is then referred to as a constant coefficient elliptic parametric functional.

It is known that the energy $\mathcal{F}$ possesses a canonical critical point which minimizes $\mathcal{F}$ among closed surfaces enclosing a specific three-dimensional volume (see [2]), and it is known as the *Wulff shape* (for $\mathcal{F}$), which we will denote by W. W is a uniformly convex smooth surface and given by the immersion $\chi : S^2 \to \mathbf{R}^3$ defined by $\chi(\nu) = DF(\nu) + F(\nu)\cdot\nu$. In the special case where $F \equiv 1$, $\mathcal{F}$ is the area functional and W is the round sphere of radius 1 with center at the origin.

The property that X is a critical point of $\mathcal{F}$ for all compactly supported volume-preserving variations is characterized by the property that the anisotropic mean curvature Λ of X is constant, where Λ is given by

$$\Lambda := 2HF - \mathrm{div}_\Sigma DF = -\mathrm{trace}_\Sigma(D^2F + FI)\circ d\nu$$

(cf. [4]). Here H is the mean curvature of X. This definition is a generalization of the idea of constant mean curvature which arises from the area functional. Since $D^2F + FI$ is positive definite, the equation $\Lambda =$ constant is absolutely elliptic, and hence a maximum principle analogous to that for CMC surfaces holds for surfaces of constant anisotropic mean curvature.

In this paper, we consider connected, compact surfaces X with nonempty boundary embedded in a region $\Omega := \{z_0 \le z \le z_1\}$ whose interiors are included in the interior of Ω, whose boundary components are restricted to lie on the two supporting (horizontal) planes $\Pi_i := \{z = z_i\}$, $i = 0, 1$, and which are constrained to enclose a fixed volume V in Ω. These considerations require that the surface bounds a connected volume so that we preclude some physically important configurations like a "string of spheres." Also, for simplicity, we are assuming that each boundary component of the considered surface is homeomorphic to a circle. We will call such a surface an *anisotropic capillary surface* if it is in equilibrium for a functional

$$\mathcal{E}[X] := \mathcal{F}[X] + \omega_0 \mathcal{A}_0[X] + \omega_1 \mathcal{A}_1[X]\,. \tag{2}$$

Here $\mathcal{A}_i[X]$ is the area in the plane Π_i which is bounded by the boundary components of X in Π_i (physically, the area which is wetted by the material inside the surface) and the ω_i's are coupling constants. In practice the ω_i's are determined by the materials involved. Throughout this paper we use the term "capillary surface" to mean anisotropic capillary surface. We will use the adjective isotropic when it is needed to denote the special case when the free energy is the surface area.

For an embedding $X : (\Sigma, \partial\Sigma) \to (\Omega, \Pi_0 \cup \Pi_1)$ with outward pointing unit normal ν, the contact angle of X with Π_i at $X(\zeta) \in \Pi_i$ ($\zeta \in \partial\Sigma$) is defined as the angle $\vartheta \in [0, \pi]$, between $\nu(\zeta)$ and $(-1)^i(0, 0, 1)$. The surface X is a capillary surface if and only if the anisotropic mean curvature Λ of X is constant, and the contact angle ϑ of X with each Π_i is a constant $\vartheta(\omega_i)$ (see [6, Propositions 3.1 and 3.2]). The precise value $\vartheta(\omega_i)$ will be given below. Now, by the maximum principle and the Alexandrov reflection methods, X is a surface of revolution with vertical rotation axis, and its genus is zero (see [6, Corollary 3.2]).

A capillary surface is said to be stable if the second variation of the energy functional $\mathcal{E}$ is nonnegative for all volume-preserving variations satisfying the boundary condition.

A natural question to ask is whether one can uniquely determine the shape of the (stable) capillary surface from the "initial data" $\mathcal{F}$, V, $h := z_1 - z_0$, ω_0, and ω_1. We will show that this is possible under certain conditions.

We will first impose conditions on the functional $\mathcal{F}$ which will be described via the corresponding Wulff shape W. We will assume the following:

- (W1) W is a uniformly convex surface of revolution with vertical rotation axis.
- (W2) W is symmetric with respect to reflection through the horizontal plane $z = 0$.
- (W3) The generating curve of W has nondecreasing curvature (with respect to the inward pointing normal) as a function of arc length on $\{z \geq 0\}$ as one moves in an upward direction.

In addition, it will be assumed that $\omega_0 = \omega_1 =: \omega \geq 0$ holds in (2). In the isotropic (liquid) case, the condition $\omega_i > 0$ is known as *hydrophobic wetting* since the material inside the surface will tend to avoid the supporting planes when minimizing energy. The case of the ω_i's being equal would occur (physically) if both supporting planes were made from the same material.

The Wulff shape W can be represented as

$$(x_1, x_2, x_3) = (u(\sigma)\cos\theta, u(\sigma)\sin\theta, v(\sigma)),$$

where σ is the arc length of the generating curve

$$\Gamma_W : (u(\sigma), v(\sigma))$$

of W. Denote by $\bar{\omega}$ the maximum height on W, that is, $\bar{\omega} = \max_\sigma v(\sigma)$. At times we will also represent the generating curve of W as a graph $(u(v), v)$, $-\bar{\omega} \leq v \leq \bar{\omega}$.

For $\omega \in (-\bar{\omega}, \bar{\omega})$, denote by $\vartheta(\omega)$ the contact angle between the region $W \cap \{x_3 \leq \omega\}$ of W and the plane $\{x_3 = \omega\}$. Also we define $\vartheta(-\bar{\omega}) := 0$, $\vartheta(\bar{\omega}) := \pi$. Then $\vartheta(\omega)$ is a continuous strictly increasing function of ω on $[0, \pi]$ with $\vartheta(0) = \pi/2$. An embedding X is a capillary surface for

$$\mathcal{E} := \mathcal{E}_\omega := \mathcal{F} + \omega\mathcal{A}_0 + \omega\mathcal{A}_1 \tag{3}$$

if and only if the anisotropic mean curvature of X is constant, and the contact angle between X and each bounding plane Π_i is constant $\vartheta(\omega)$ along the boundary (see [6, Propositions 3.1 and 3.2]).

We will call an anisotropic capillary surface *spanning* if its intersection with both supporting planes is a circle of positive radius. We denote by $V_0(h, \omega)$ the infimum of the volumes of stable spanning anisotropic capillary surfaces having height h and contact angle $\vartheta(\omega)$.

In keeping with the classical terminology, we will refer to a compact anisotropic capillary surface having nonempty boundary components only on the plane $z = z_0$ (resp., $z = z_1$) as a *sessile drop* (resp., *pendent drop*). Such a surface is necessarily rotationally invariant, and therefore homothetic to a part of the Wulff shape (see [6]).

If $|\omega| > \bar{\omega}$, then there is no capillary surface for the energy $\mathcal{E}_\omega$ (see [6, Corollary 3.1]). For $0 \leq \omega \leq \bar{\omega}$, we will show the following uniqueness theorem.

Theorem 2.1. *We assume* (W1) *through* (W3) *stated above.*

I. *Assume $0 \leq \omega < \bar{\omega}$. Then $V_0(h, \omega) > 0$ holds and*
   (i) *for volume $V < V_0$, any stable capillary surface for the energy $\mathcal{E}_\omega$ with volume V and height h is a sessile or pendent drop;*
   (ii) *for volumes $V \geq V_0$, there exists a unique stable spanning capillary surface for the energy $\mathcal{E}_\omega$ with volume V and height h.*

II. *Assume $\omega = \bar{\omega}$. Then any capillary surface for the energy $\mathcal{E}_\omega$ is tangent to the supporting planes $\Pi_0 \cup \Pi_1$. $V_0(h, \omega) > 0$ holds, and it coincides with the volume of the closed surface homothetic to the Wulff shape which is tangent to both Π_0 and Π_1. Also,*
   (i) *for volume $V \leq V_0$, there is no stable capillary surface for the energy $\mathcal{E}_\omega$ with volume V and height h;*
   (ii) *for volumes $V > V_0$, there exists a unique stable capillary surface for the energy $\mathcal{E}_\omega$ with volume V and height h. Moreover, this surface is spanning.*

Actually, we will later give analytic and geometric characterizations of each of the unique solutions for $V \geq V_0$ in Theorem 2.1. In order to do this, we first recall the classification of surfaces of revolution with constant anisotropic mean curvature (see section 6). Such surfaces were studied in detail by the authors in [4] and are called *anisotropic Delaunay surfaces*. They are classified into six classes: horizontal plane, anisotropic catenoid, Wulff shape (up to translation and homothety), cylinder, anisotropic unduloid, and anisotropic nodoid. Each surface in each of these classes has similar properties to the corresponding Delaunay surface.

We let μ_i, $i = 1, 2$, denote the principal curvatures of the Wulff shape W with respect to the inward pointing normal. Here we let μ_1 denote the curvature of the generating curve of W.

The following characterization of stable anisotropic capillary surfaces was obtained in our previous papers [5], [6].

Theorem 2.2. *Let X be a capillary surface with free boundary on two horizontal planes for the functional* (3) *with $\omega \geq 0$ and with the Wulff shape for the functional satisfying conditions* (W1) *through* (W3) *stated above.*

(i) *If* $\omega = 0$*, then* X *is stable if and only if the surface is either homothetic to a half of the Wulff shape or a cylinder which is perpendicular to* $\Pi_0 \cup \Pi_1$ *and whose height* h *and radius* R *satisfy*

$$\frac{\mu_1(0)}{\mu_2(0)}(1/R^2) \leq (\pi/h)^2 \,,$$

where $\mu_i(0)$*,* $i = 1, 2$*, is the value of* μ_i *along the equator of* W*.*

(ii) *If* $\omega > 0$ *holds, then* X *is stable if and only if* X *is a portion of an anisotropic Delaunay surface whose generating curve has no inflection points in its interior.*

Define

$$V_1 := V_1(h, \omega) := \pi h^3 \frac{\int_{-\omega}^{\omega} u^2 \, dv}{\left(\int_{-\omega}^{\omega} \, dv\right)^3} \,.$$

V_1 is the volume of the capillary surface which is homothetic to the part of the Wulff shape with contact angle $\vartheta(\omega)$ on the plane Π_i, $i = 0, 1$.

$$V_2 := V_2(h, \omega) := \pi h^3 \frac{\int_{-\omega}^{\bar{\omega}} u^2 \, dv}{\left(\int_{-\omega}^{\bar{\omega}} \, dv\right)^3} \,.$$

V_2 is the volume of the surface which is homothetic to the part of the Wulff shape which is tangent to the plane Π_1 and with contact angle $\vartheta(\omega)$ on the plane Π_0.

THEOREM 2.3. *We assume* (W1) *through* (W3) *stated above.*

I. *Assume* $0 < \omega < \bar{\omega}$*. Then the following hold:*

(i) *For volumes* $V_0 \leq V < V_1$*, there exists a unique stable spanning capillary surface with volume* V*, height* h*, and contact angle* $\vartheta(\omega)$*, and the surface is an anisotropic unduloid. For* $V = V_0$*, this surface has inflection points on the boundary, while, for* $V_0 < V < V_1$*, it does not have inflection points.*

(ii) *For* $V = V_1$*, there exists a unique stable capillary surface with volume* V*, height* h*, and contact angle* $\vartheta(\omega)$*, and the surface is homothetic to a part of the Wulff shape.*

(iii) *For* $V_1 < V$*, there exists a unique stable capillary surface with volume* V*, height* h*, and contact angle* $\vartheta(\omega)$*, and the surface is an anisotropic nodoid.*

II. *Assume* $\omega = \bar{\omega}$*. Then, for* $V_0 < V$*, there exists a unique stable capillary surface with volume* V*, height* h*, and contact angle* $\vartheta(\omega)$*, and the surface is an anisotropic nodoid.*

Figure 1 shows the generating curves of examples of Theorem 2.3 (I) (i), (ii), and (iii) for the isotropic case, while Figure 2 shows examples for the anisotropic case.

For a fixed volume V, height h, and $\bar{\omega} \geq \omega > 0$, we let $U(V, h, \omega)$ (resp., $N(V, h, \omega)$) denote the stable anisotropic unduloid (resp., nodoid) with volume V, height $h := z_1 - z_0$, and contact angle $\vartheta(\omega)$ which we obtained in Theorem 2.3.

Remark 2.1. In the theorems above "unique" means "unique up to horizontal translation."

Remark 2.2. Even in the isotropic case, there is no uniqueness without the stability assumption. Figure 3 shows the plots of the volumes of two families of

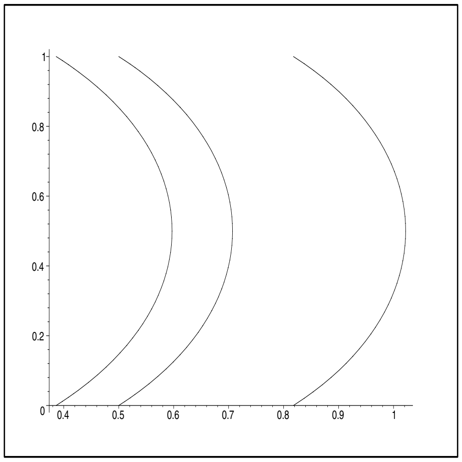


FIG. 1. *The innermost curve generates an (isotropic) unduloid, the middle curve is a sphere, and the outer curve is a nodoid. The height is* 1 *and the contact angle is* $\pi/4$*. The values of* a *are* 0.25*,* 0*, and* -1*.*

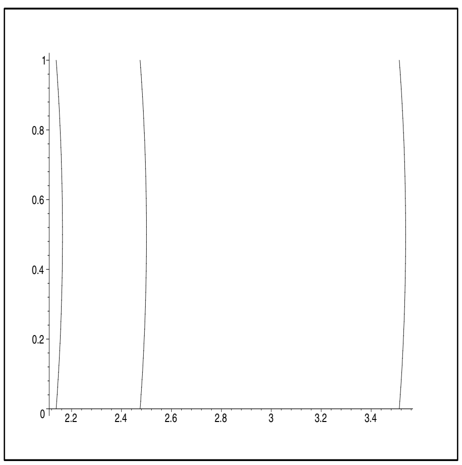


FIG. 2. *The innermost curve generates an anisotropic unduloid, the middle curve is a part of the Wulff shape* $u^2 + v^4 = 1$*, and the outer curve is an anisotropic nodoid. The height is* 1 *and the contact angle is* $\omega = \pi/4$*. The values of* a *are* 0.25*,* 0*, and* -1*.*

capillary surfaces for the area functional. The top curve represents the volumes of stable capillary unduloids with height 1 and contact angle $\pi/4$ with two planes. The bottom curve shows the volumes of unstable capillary unduloids with the same height and contact angles. The generating curves of these surfaces have exactly one interior inflection point, which makes them unstable by Theorem 2.2. This shows that volume does not uniquely determine the surface without the stability assumption.

Also, there is no positive lower bound for the volume without the assumption of stability. For any functional satisfying the conditions above, any vertical round cylinder is a capillary surface for the case $\omega = 0$. However, the volume of the cylinder can be made arbitrarily small. Also, for $0 < \omega < \bar{\omega}$, there is an unstable unduloid with contact angle $\vartheta(\omega)$ and an arbitrarily small volume.

Remark 2.3. For $V \geq V_2$, the capillary surface is unique. If $V_0 < V_2$, then, for

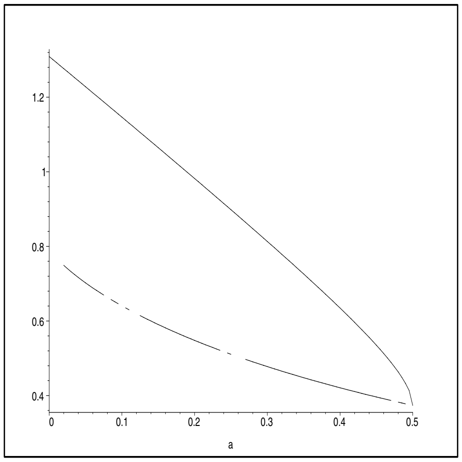


FIG. 3. *Plot of the volumes of stable (upper) and unstable (lower) unduloids as a function of* a*. The height is* 1 *and the contact angle is* $\pi/4$*. The generating curves of the unstable unduloids have exactly one inflection point.*

$V_0 \le V < V_2$, there exist exactly two stable capillary surfaces (up to translation) with volume V, height h, and contact angle $\vartheta(\omega)$. One of them is a sessile or pendent drop, while the other has two boundary components. In the isotropic case, these results follow from Chapter 6 of [10]. It would be interesting to know whether $V_0 \le V < V_2$ holds in general. This inequality is proved in the case $\omega = 0$ in section 5.

The next result yields a numerical lower bound on the volume of a stable, spanning capillary surface.

THEOREM 2.4. *Assume that the Wulff shape satisfies conditions* (W1) *through* (W3)*. If* $0 < \omega \le \bar{\omega}$*, then*

$$V_0(h,\omega) > \frac{h^3}{\pi}\left(\frac{2u(\omega)(u(0)-u(\omega))}{\omega^2-(u(0)-u(\omega))^2}\right) > 0 \tag{4}$$

holds. If $\omega = 0$*, then*

$$V_0(h,0) \ge \frac{h^3}{\pi}\left(\frac{\mu_1(0)}{\mu_2(0)}\right) \tag{5}$$

holds, and this inequality is sharp in the sense that there is a stable cylinder which satisfies the equality in (5)*.*

Remark 2.4. For the CMC case, Theorem 2.4 implies that if the contact angle $\vartheta = \vartheta(\omega)$ satisfies $\pi/2 \le \vartheta \le \pi$, then

$$V_0(h,\omega) \ge \frac{h^3}{\pi}$$

holds and the equality holds only for the most slender stable cylinder. This is exactly the result proved by Finn and Vogel in [3] for $0 < \vartheta \le \pi$. Zhou [9] proved this for the general case where the contact angle on the lower and upper planes may be different.

Finally, we will show the following.

THEOREM 2.5. *Assume* $0 \le \omega \le \bar{\omega}$ *and that the Wulff shape satisfies conditions* (W1) *through* (W3)*. For* $V \ge V_0$*, let* $\Sigma(V) = \Sigma(V,h,\omega)$ *denote the unique stable* *capillary surface with volume V, height h, and contact angles $\vartheta(\omega)$ with two boundary components. Here, we let $\Sigma(V_0, h, \bar{\omega})$ be the homothety of the Wulff shape with height h which touches both Π_0 and Π_1. Then the family of surfaces $\Sigma(V)$, $V > V_0$, foliate the open region of space which lies exterior to the surface $\Sigma(V_0)$ and lies between the planes $z = z_i$, $i = 0, 1$.*

3. Preliminary results. We introduce the auxiliary quantity

$$V^* := V^*(h, \omega) := \pi h^3 \left(\int_{-\omega}^{\omega} \frac{1}{\sqrt{u^2 - u^2(\omega)}} \, dv \right)^{-2}, \tag{6}$$

which will be used to obtain the lower bound for the volume in Theorem 2.4. The main result of this section is the following technical lemma.

LEMMA 3.1. *Assume $0 < \omega \leq \bar{\omega}$ and that the Wulff shape satisfies conditions* (W1) *through* (W3). *Let $\hat{R}$ denote the radius of the circle through the points $(u(\omega), \pm\omega), (u(0), 0)$. Define*

$$A(\omega) = \left[\frac{u(\omega) + (\hat{R} - u(0))}{u(\omega)} \right]^{1/2}. \tag{7}$$

Then there holds

$$\int_{-\omega}^{\omega} \frac{dv}{\sqrt{u^2(v) - u^2(\omega)}} < \pi A(\omega), \tag{8}$$

and consequently

$$V^*(h, \omega) > \frac{h^3}{\pi} \left(\frac{2u(\omega)(u(0) - u(\omega))}{\omega^2 - (u(0) - u(\omega))^2} \right) > 0 \tag{9}$$

holds.

The generating curve Γ_W^+ of the Wulff shape is represented as

$$\begin{aligned} &(u(\sigma), v(\sigma)), \qquad -L \leq \sigma \leq L, \\ u_0 &:= \max u = u(0) \geq u(\sigma) \geq 0 = u(-L) = u(L) \qquad \forall \sigma \in [-L, L], \\ \bar{\omega} &:= \max v = v(L) \geq v(\sigma) \geq -\bar{\omega} = v(-L) \qquad \forall \sigma \in [-L, L], \end{aligned}$$

where σ is the arc length of (u, v) and $2L$ is the length of Γ_W^+.

$$(u(\sigma), v(\sigma)), \qquad -2L \leq \sigma \leq 2L,$$

gives a convex closed curve Γ_W, which is the section of W by the (x_1, x_3)-plane. For simplicity, we set

$$\kappa := \mu_1,$$

which is the curvature of $(u(\sigma), v(\sigma))$ with respect to the inward pointing normal.

LEMMA 3.2.

$$f(\sigma) := u^2(\sigma) + v^2(\sigma)$$

is a nondecreasing function of σ in $0 \leq \sigma \leq L$.

In order to prove Lemma 3.2, we need the following.

LEMMA 3.3. *Consider the eigenvalue problem*

$$\varphi'' + \kappa^2\varphi = -\lambda\varphi \quad in\ 0 \le \sigma \le L, \qquad \varphi(0) = \varphi(L) = 0. \tag{10}$$

Then the first eigenvalue $\lambda_1[0, L]$ of problem (10) *is nonnegative.*

Proof. Set $\eta := v'$. Let ϕ be a function on $[0, L]$ which satisfies $\phi(0) = \phi(L) = 0$. Since

$$\eta > 0 \quad \text{in } 0 < \sigma < L, \quad \eta(L) = 0, \quad \eta'(L) \neq 0,$$

the function

$$\zeta := \phi/\eta$$

is well-defined on $0 \le \sigma \le L$. Elementary calculations show

$$\eta'' + \kappa^2\eta = \kappa' u'.$$

Using this, we obtain

$$\phi'' + \kappa^2\phi = \zeta''\eta + 2\zeta'\eta' + \zeta\kappa' u'.$$

Therefore,

$$\begin{aligned}
-\int_0^L \phi(\phi'' + \kappa^2\phi)\, d\sigma &= -\int_o^L \zeta\eta(\zeta''\eta + 2\zeta'\eta' + \zeta\kappa' u')\, d\sigma \\
&= -\int_0^L \Big\{(\zeta\zeta'\eta^2)' - (\zeta')^2\eta^2 + \kappa'\zeta^2\eta u'\Big\}\, d\sigma \\
&= -[\zeta'\eta\phi]_0^L + \int_0^L \Big\{(\zeta')^2\eta^2 - \kappa'\zeta^2 u'v'\Big\}\, d\sigma \\
&= \int_0^L \Big\{(\zeta')^2\eta^2 - \kappa'\zeta^2 u'v'\Big\}\, d\sigma \ge 0,
\end{aligned}$$

which implies that $\lambda_1[0, L] \ge 0$. □

Proof of Lemma 3.2. We will prove $f' \ge 0$. Note that $\kappa' = 0$ on interval $[0, L]$ is equivalent to $f =$ constant and so $f' = 0$ on $[0, L]$. Denote by q the support function of the curve (u, v). Then $q = uv' - u'v > 0$. By elementary calculations, we obtain

$$f''' + \kappa^2 f' = -2\kappa' q \le 0.$$

Note that $f'(0) = f'(L) = 0$ holds. Now assume $f'(\sigma) < 0$ at some $\sigma \in (0, L)$. Then there exist some $\sigma_1,\ \sigma_2 \in [0, L]$ such that $0 \le \sigma_1 < \sigma_2 \le L$ and

$$f'(\sigma) < 0 \quad \forall\sigma \in (\sigma_1, \sigma_2), \qquad f'(\sigma_1) = f'(\sigma_2) = 0$$

holds. If $[\sigma_1, \sigma_2] = [0, L]$, then

$$-\int_0^L f'(f''' + \kappa^2 f')\, d\sigma = 2\int_0^L \kappa' q f'\, d\sigma < 0.$$

Therefore, $\lambda_1[0, L] < 0$ holds. This contradicts Lemma 3.3. If $[\sigma_1, \sigma_2] \neq [0, L]$, then

$$-\int_{\sigma_1}^{\sigma_2} f'(f''' + \kappa^2 f')\, d\sigma = 2\int_{\sigma_1}^{\sigma_2} \kappa' q f'\, d\sigma \le 0. \tag{11}$$

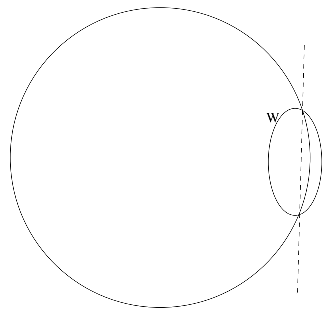


FIG. 4.

Since the eigenvalues of problem (10) have monotonicity with respect to the region, (11) implies that $\lambda_1[0, L] < \lambda_1[\sigma_1, \sigma_2] \le 0$. Again this contradicts Lemma 3.3. $\square$

We assume $0 < \omega \le \bar{\omega}$. Γ_W can be regarded as the graph $(u(v), v)$, $-\bar{\omega} \le v \le \bar{\omega}$, of a function $u(v)$ of v.

The line segment ℓ through the points $(u(\omega), \pm\omega)$ is the limit as $R \to \infty$ of a family of arcs α_R through $(u(\omega), \pm\omega)$ of circles C_R of radius R having centers $(-z_R, 0)$ on the real axis. Let Γ denote the arc of Γ_W with $u > 0$ and $-\omega < v < \omega$. It is clear that for $R \gg 0$, α_R lies strictly between ℓ and Γ. From now on we will consider only these values of R. Thus, if α_R is given by $(U_R(v), v)$ with $-\omega < v < \omega$, then $0 \le U_R(v) \le u(v)$ holds. It is also clear that $0 \le u(\omega) \le U_R$ holds and so $U_R^2 - u^2(\omega) \ge 0$ holds. (See Figure 4.) It follows that

$$\int_{-\omega}^{\omega} \frac{1}{\sqrt{u^2 - u^2(\omega)}}\, dv < \int_{-\omega}^{\omega} \frac{1}{\sqrt{U_R^2 - u^2(\omega)}}\, dv\,. \tag{12}$$

We will try to obtain a lower bound of $U_R^2 - u^2(\omega)$.

The equation of C_R is

$$(U + z_R)^2 + v^2 = R^2\,, \tag{13}$$

and so

$$(u(\omega) + z_R)^2 + \omega^2 = R^2\,. \tag{14}$$

Subtracting these equations and performing elementary manipulations leads to

$$(U_R - u(\omega))(U_R + u(\omega) + 2z_R) = (\omega^2 - v^2)\,.$$

Letting $\hat{z}_R = \max(0, z_R)$, we have

$$U_R^2 - u^2(\omega) \ge \left[\frac{U_R + u(\omega)}{U_R + u(\omega) + 2\hat{z}_R}\right](\omega^2 - v^2)\,.$$

Since the function on the right is a nondecreasing function of U_R ($\ge u(\omega)$), we have

$$U_R^2 - u^2(\omega) \ge \left[\frac{2u(\omega)}{2u(\omega) + 2\hat{z}_R}\right](\omega^2 - v^2)\,. \tag{15}$$

In order to obtain a lower bound of $U_R^2 - u^2(\omega)$ from (15), we will need a lower bound on $\hat{z}_R$.

LEMMA 3.4. *We consider circles*

$$C_+(a,r) : (U-a)^2 + V^2 = r^2, \qquad U \geq a.$$

If a circle $C_+(a,r)$ touches the right half

$$\Gamma_W^+ := \{(u(\sigma), v(\sigma)) \mid u(\sigma) \geq 0\}$$

of Γ_W at a point (u_0, v_0) $(v_0 \neq 0)$ from the left-hand side, then Γ_W is a circle.

Proof. We denote by $\kappa(v) > 0$ the curvature of Γ_W at (u, v). Then $\kappa(v)$ is an even function and

$$\kappa'(v) \geq 0, \qquad 0 \leq \forall v \leq \bar{\omega}. \tag{16}$$

Set

$$\Gamma := \{(u,v) \in \Gamma_W^+ \mid -|v_0| \leq v \leq |v_0|\},$$
$$C := \{(U,V) \in C_+(a,r) \mid -|v_0| \leq V \leq |v_0|\}.$$

Because of symmetry, C touches Γ on the boundary from the left-hand side.

About the curvatures of these two curves at the point (u_0, v_0), it holds that

$$1/r \geq \kappa(v_0).$$

Therefore, by assumption (W3),

$$1/r \geq \kappa(v), \qquad -v_0 \leq \forall v \leq v_0, \tag{17}$$

holds.

We now move Γ in the negative direction of the u axis so that it does not intersect C. Then we move Γ toward the positive direction of the u axis until it intersects C for the first time, and we denote by $\tilde{\Gamma}$ the translated curve at this time. $\tilde{\Gamma}$ is tangent to C at an interior point $(\tilde{u}, \tilde{v})$. Because of the symmetry of $\tilde{\Gamma}$ and C with respect to the v axis, we may assume that $0 < \tilde{v} \leq v_0$. Since $\tilde{\Gamma}$ lies in the negative side of C with respect to u,

$$1/r \leq \kappa(\tilde{v}) \tag{18}$$

holds. It holds from (16), (17), and (18) that

$$1/r = \kappa(v) \qquad \forall v \in [-v_0, -\tilde{v}] \cup [\tilde{v}, v_0].$$

Therefore, $\tilde{\Gamma}$ is tangent to C at the point (u_0, v_0) and lies on the negative side of C with respect to u. On the other hand, since both $\tilde{\Gamma}$ and Γ are tangent to C at point (u_0, v_0), $\tilde{\Gamma}$ coincides with Γ. Recall that Γ lies on the positive side of C with respect to u. Therefore, Γ coincides with C. Again by assumption (W3), Γ_W is a circle. ∎

LEMMA 3.5. *If we decrease the radius of the circle C_R, then the inequalities*

$$u(\omega) \leq U_R \leq u$$

are satisfied until a value $R = \hat{R}$ is reached, at which the curve $(U_{\hat{R}}(v), v)$ is tangent to the curve (u, v) at the point $(u(0), 0)$. Moreover, $\hat{R} \geq u(0)$ holds. Here, the equality holds if and only if Γ_W is a circle.

Proof. If the Wulff shape W is a sphere, then the statement clearly holds. Hence we assume that W is not a sphere.

Now, assume that the circle $(U_R(v), v)$ is tangent to the curve (u, v) at a point $(u(v_0), v_0)$ $(0 < v_0 \le \omega)$ and

$$u(\omega) \le U_R(v) \le u(v), \qquad -\omega \le \forall v \le \omega,$$

holds. Then, by Lemma 3.4, W must be a sphere, which is a contradiction. Therefore, we have proved the first statement.

Next, we prove $\hat{R} \ge u(0)$. We consider circles $C(r)$ with center at the origin. If $r > 0$ is small, then $C(r)$ is contained in the domain bounded by Γ_W. If we increase r, then, at a certain r_1, $C(r)$ touches Γ_W for the first time from the inside of Γ_W. Because of Lemma 3.4, $C(r)$ touches Γ_W at $(\pm u(0), v(0))$. This implies $\hat{R} \ge u(0)$.

If $\hat{R} = u(0)$, then $C(u(0)) = C_{\hat{R}}$, and this circle touches Γ_W at $(u(\omega), \pm\omega)$. Therefore, by Lemma 3.4, Γ_W is a circle. ∎

Proof of Lemma 3.1. Lemma 3.5 supplies a lower bound for z_R which we denote by $\hat{z}(\omega)$; that is, if $\hat{R}$ is the radius of the circle passing the three points $(u(\omega), \pm\omega), (u(0), 0)$, then

$$\hat{z}(\omega) = \hat{R} - u(0).$$

Therefore,

$$A(\omega) := \left[\frac{u(\omega) + \hat{R} - u(0)}{u(\omega)}\right]^{1/2} = \left[\frac{u(\omega) + \hat{z}(\omega)}{u(\omega)}\right]^{1/2} . \tag{19}$$

We obtain from (12) and (15),

$$\int_{-\omega}^{\omega} \frac{1}{\sqrt{u^2 - u^2(\omega)}}\, dv < A(\omega) \int_{-\omega}^{\omega} \frac{1}{\sqrt{\omega^2 - v^2}}\, dv = A(\omega)\pi . \tag{20}$$

This implies (8). Since the points $(u(\omega), \pm\omega), (u(0), 0)$ lie on the circle given by (13) with center $(-\hat{z}(\omega), 0)$ and radius $\hat{R}$, we have

$$(u(\omega) + \hat{z}(\omega))^2 + \omega^2 = \hat{R}^2 = (u(0) + \hat{z}(\omega))^2 .$$

This leads to

$$\hat{z}(\omega) = \frac{u^2(\omega) + \omega^2 - u^2(0)}{2(u(0) - u(\omega))} > 0 . \tag{21}$$

The last inequality is because the numerator above is nonnegative by Lemma 3.2. By substituting (21) into (19), we obtain

$$A(\omega) = \left[\frac{\omega^2 - (u(0) - u(\omega))^2}{2u(\omega)(u(0) - u(\omega))}\right]^{1/2} . \tag{22}$$

The first inequality in (9) follows from (6), (8), and (22). ∎

4. Proofs of Theorems 2.1, 2.3, 2.4, and 2.5. First, we give a lemma.

LEMMA 4.1. *$V_1 > V_2$ holds.*

Proof.

$$V_1 = \pi h^3 \frac{\int_{-\omega}^{\omega} u^2\, dv}{(2\omega)^3} , \tag{23}$$

$$V_2 = \pi h^3 \frac{\int_{-\omega}^{\bar{\omega}} u^2\, dv}{(\bar{\omega}+\omega)^3} = \pi h^3 \frac{1}{2\omega(\bar{\omega}+\omega)^2} \int_{\frac{-2\omega^2}{\bar{\omega}+\omega}}^{\frac{2\omega\bar{\omega}}{\bar{\omega}+\omega}} u^2(v(\eta))\, d\eta\,, \quad \eta := \frac{2\omega}{\bar{\omega}+\omega} v\,. \tag{24}$$

Since

$$v(\eta) = \frac{\bar{\omega}+\omega}{2\omega}\eta, \qquad \frac{\bar{\omega}+\omega}{2\omega} > 1$$

holds,

$$u(v(\eta)) < u(\eta)$$

holds. Hence,

$$\int_{\frac{-2\omega^2}{\bar{\omega}+\omega}}^{\omega} u^2(v(\eta))\, d\eta < \int_{\frac{-2\omega^2}{\bar{\omega}+\omega}}^{\omega} u^2(v)\, dv \tag{25}$$

holds. Set

$$A := \int_{\omega}^{\frac{2\omega\bar{\omega}}{\bar{\omega}+\omega}} u^2(v(\eta))\, d\eta, \qquad B := \int_{-\omega}^{\frac{-2\omega^2}{\bar{\omega}+\omega}} u^2(v)\, dv.$$

We will show $A < B$. By the symmetry of $u(v)$ with respect to v, we have

$$B = \int_{\frac{2\omega^2}{\bar{\omega}+\omega}}^{\omega} u^2(v)\, dv.$$

Set

$$\xi(\eta) := \eta - \frac{\omega(\bar{\omega}-\omega)}{\bar{\omega}+\omega}.$$

Then

$$v(\eta) = \frac{\bar{\omega}+\omega}{2\omega}\xi + \frac{\bar{\omega}-\omega}{2} > \xi.$$

Therefore, we have

$$A = \int_{\frac{2\omega^2}{\bar{\omega}+\omega}}^{\omega} u^2(v(\eta(\xi)))\, d\xi < \int_{\frac{2\omega^2}{\bar{\omega}+\omega}}^{\omega} u^2(\xi)\, d\xi = B. \tag{26}$$

Formulas (23)–(26) imply that $V_2 < V_1$ holds. ∎

Proof of Theorems 2.1, 2.3, *and* 2.4. First note that, for $V > V_2$, there is no sessile or pendent drop. Especially, by Lemma 4.1, for $V \geq V_1$, there is no sessile or pendent drop.

Assume $0 < \omega \leq \bar{\omega}$. Let $X(s, \theta) = (x(s)e^{i\theta}, z(s))$ be a stable spanning capillary surface. Then, by Theorem 2.2 (ii), X is a convex part of an anisotropic Delaunay surface. From the representation formula (Theorem 6.2) for anisotropic Delaunay surfaces and Remark 6.1, it follows that for stable capillary surfaces, the height h and volume V are given as follows: First note that

$$dz = \frac{z_s}{x_s} dx = \frac{v_\sigma}{u_\sigma} dx = x_u dv$$

holds. Therefore,

$$h = \int_{v=-\omega}^{v=\omega} dz = \int_{-\omega}^{\omega} x_u \, dv = \frac{1}{-\Lambda} \int_{-\omega}^{\omega} 1 + \frac{u}{\sqrt{u^2 + \Lambda c}} \, dv, \tag{27}$$

$$V = \pi \int_{v=-\omega}^{v=\omega} x^2 \, dz = \pi(-\Lambda)^{-3} \int_{-\omega}^{\omega} \frac{(u + \sqrt{u^2 + \Lambda c})^3}{\sqrt{u^2 + \Lambda c}} \, dv,$$

where $\Lambda \leq 0$ is the anisotropic mean curvature of X and c is the flux parameter for X.

We consider the scale invariant quantity,

$$a := -\Lambda c.$$

Then

$$h = \frac{1}{-\Lambda} \int_{-\omega}^{\omega} 1 + \frac{u}{\sqrt{u^2 - a}} \, dv, \tag{28}$$

$$V = \pi(-\Lambda)^{-3} \int_{-\omega}^{\omega} \frac{(u + \sqrt{u^2 - a})^3}{\sqrt{u^2 - a}} \, dv, \tag{29}$$

and we obtain

$$V = \pi h^3 \left(\int_{-\omega}^{\omega} 1 + \frac{u}{\sqrt{u^2 - a}} \, dv \right)^{-3} \int_{-\omega}^{\omega} \frac{(u + \sqrt{u^2 - a})^3}{\sqrt{u^2 - a}} \, dv \,. \tag{30}$$

By applying Hölder's inequality for the measure $dv/\sqrt{u^2 - 1}$ to the formula for h, we obtain

$$h \leq \frac{1}{-\Lambda} \left(\int_{-\omega}^{\omega} \frac{(u + \sqrt{u^2 - a})^3}{\sqrt{u^2 - a}} \, dv \right)^{1/3} \left(\int_{-\omega}^{\omega} \frac{1}{\sqrt{u^2 - a}} \, dv \right)^{2/3}.$$

It then follows that

$$V/h^3 \geq \pi \left(\int_{-\omega}^{\omega} \frac{1}{\sqrt{u^2 - a}} \, dv \right)^{-2} \geq \pi \left(\int_{-\omega}^{\omega} \frac{1}{\sqrt{u^2 - u^2(\omega)}} \, dv \right)^{-2}. \tag{31}$$

If $V_0(h, \omega)$ is defined as the infimum of the volume of all stable capillary surfaces with the given height and ω having *nonempty* boundary components on both planes, (31) shows that $V_0(h, \omega) \geq V^*(h, \omega)$ holds. The first half of Theorem 2.4 then follows from this inequality and Lemma 3.1. The second half of Theorem 2.4 follows easily from Theorem 2.2 (i).

Now we prove Theorems 2.1 and 2.3. Define

$$\Gamma(a, \omega) := \pi \left(\int_{-\omega}^{\omega} 1 + \frac{u}{\sqrt{u^2 - a}} \, dv \right)^{-3} \int_{-\omega}^{\omega} \frac{(u + \sqrt{u^2 - a})^3}{\sqrt{u^2 - a}} \, dv \,. \tag{32}$$

It follows from (28) and (29) that a necessary and sufficient condition for there to exist a spanning, stable capillary surface with prescribed h and V is that $V/h^3 = \Gamma(a, \omega)$ for some real value a.

Note that

$$a = -\Lambda c \le u^2(\omega) \quad \text{for } 0 < \omega \le \bar{\omega},$$

$$0 = u(\bar{\omega}) < u(\omega) < u(0) \quad \text{for } 0 < \omega < \bar{\omega}$$

hold. Also note (see section 6) that for $a > 0$ the capillary surfaces are anisotropic unduloids, for $a = 0$ they are part of the Wulff shape (up to translation and homothety), while for $a < 0$ they are anisotropic nodoids. The result will then follow by showing that with the height h fixed, the volume is a strictly decreasing function of a for $a \le u^2(\omega)$.

First assume $a < u^2(\omega)$. Differentiating (30) with respect to a, we have

$$\begin{aligned} &-2(\pi h^3)^{-1} V_a \left(\int_{-\omega}^{\omega} 1 + \frac{u}{\sqrt{u^2 - a}}\, dv \right)^4 \\ &= 3 \int_{-\omega}^{\omega} \frac{u}{(u^2 - a)^{3/2}}\, dv \int_{-\omega}^{\omega} \frac{(u + \sqrt{u^2 - a})^3}{\sqrt{u^2 - a}}\, dv \\ &- \int_{-\omega}^{\omega} \frac{u + \sqrt{u^2 - a}}{\sqrt{u^2 - a}}\, dv \int_{-\omega}^{\omega} \frac{(u + \sqrt{u^2 - a})^2 (u - 2\sqrt{u^2 - a})}{(u^2 - a)^{3/2}}\, dv. \end{aligned} \tag{33}$$

We will show that the right-hand side of (33) is positive. If

$$u - 2\sqrt{u^2 - a} \le 0$$

holds for all u for $-\omega \le v \le \omega$, then it is done. Assume now that

$$u - 2\sqrt{u^2 - a} > 0 \tag{34}$$

holds for some u. In particular, a must be positive. Note that

$$\begin{aligned} &-2(\pi h^3)^{-1} V_a \left(\int_{-\omega}^{\omega} 1 + \frac{u}{\sqrt{u^2 - a}}\, dv \right)^4 \\ &= \frac{1}{a} \left(3 \int_{-\omega}^{\omega} \frac{au}{(u^2 - a)^{3/2}}\, dv \int_{-\omega}^{\omega} \frac{(u + \sqrt{u^2 - a})^3}{\sqrt{u^2 - a}}\, dv \right. \\ &\left. - \int_{-\omega}^{\omega} \frac{a(u + \sqrt{u^2 - a})}{\sqrt{u^2 - a}}\, dv \int_{-\omega}^{\omega} \frac{(u + \sqrt{u^2 - a})^2 (u - 2\sqrt{u^2 - a})}{(u^2 - a)^{3/2}}\, dv \right). \end{aligned} \tag{35}$$

We will prove that

$$(u + \sqrt{u^2 - a})^3 > a(u + \sqrt{u^2 - a}), \tag{36}$$

$$au > (u + \sqrt{u^2 - a})^2 (u - 2\sqrt{u^2 - a}) \tag{37}$$

holds for any u satisfying (34). Because $a < u^2$ holds, (36) clearly holds. Inequality (34) is equivalent to

$$a < u^2 < \frac{4}{3} a. \tag{38}$$

Set

$$f(u) := au - (u + \sqrt{u^2 - a})^2(u - 2\sqrt{u^2 - a}).$$

Then

$$f(u) = 2(u + \sqrt{u^2 - a})(u^2 - a) > 0$$

holds. This proves (37). Combining (35) with (36) and (37) gives

$$V_a < 0.$$

This implies that V is a strictly decreasing function of a.

It remains to show that the monotonicity extends to the point $a = u^2(\omega)$. This will follow if we can show that, again with the height fixed, V has an extension to $u^2(\omega)$ which is continuous from below.

Both integrals in (32) are of the form

$$\int_{-\omega}^{\omega} \frac{(u + \sqrt{u^2 - a})^p}{\sqrt{u^2 - a}}\, dv$$

with $p = 1, 3$. Also, both integrands are bounded by constant $\cdot\, (\sqrt{u^2 - a})^{-1} \leq$ constant $\cdot\, (\sqrt{u^2 - u^2(\omega)})^{-1}$. Therefore, the continuity of V from below follows by the dominated convergence theorem since it follows from Lemma 3.1 that the integral

$$I := \int_{-\omega}^{\omega} \frac{1}{\sqrt{u^2 - u^2(\omega)}}\, dv$$

is convergent. □

Proof of Theorem 2.5. In the case where $\omega = 0$, because of Theorem 2.2 (i), the statement clearly holds. So, we will assume $0 < \omega \leq \bar{\omega}$. For convenience we will assume that the height is 1. We may assume $z_0 = -1/2$ and $z_1 = 1/2$. For $-\omega \leq v \leq \omega$, we write the generating curve of the Wulff shape as $(u(v), v)$. By using Theorem 6.2 and (28), we can express the coordinates of each capillary surface as $v \mapsto (x(a, v), z(a, v))$, with

$$x(a, v) = \frac{u(v) + \sqrt{u^2(v) - a}}{\int_{-\omega}^{\omega} 1 + u(v)/\sqrt{u^2(v) - a}\, dv}\,.$$

Then for $a < u^2(\omega)$,

$$\begin{aligned}(\partial_a x)(a, v) = &-(1/2)\left(\sqrt{u^2(v) - a}\int_{-\omega}^{\omega} 1 + u(v)/\sqrt{u^2(v) - a}\, dv\right)^{-1}\\ &-(1/2)(u(v) + \sqrt{u^2(v) - a})\left(\int_{-\omega}^{\omega} u(v)(u^2(v) - a)^{-3/2}\, dv\right)\\ &\times\left(\int_{-\omega}^{\omega} (1 + u(v)/\sqrt{u^2(v) - a})\, dv\right)^{-2}\\ &< 0\,.\end{aligned}$$

Thus, for v fixed, $x(a, v)$ is strictly decreasing as a function of a for $a < u^2(\omega)$. Note that the generating curve of each capillary surface can also be represented as a graph $x = \underline{x}(a, z)$, $-1/2 \leq z \leq 1/2$.

We now assume that two generating curves $(x(a,v),z(a,v))$, $(x(b,v),z(b,v))$, $a<b<u^2(\omega)$, intersect. By the above, it is clear that $x(b,0)<x(a,0)$ and equivalently $\underline{x}(b,0)<\underline{x}(a,0)$. Similarly, $x(b,\pm\omega)<x(a,\pm\omega)$ holds, and so equivalently $\underline{x}(b,\pm 1/2)<\underline{x}(a,\pm 1/2)$.

Note that these two curves cannot have any nontransversal intersections for $-1/2<z<1/2$. If they did, then at the point of intersection, the values of v (which depends only on the tangent at each point) for both curves must agree, contradicting the fact that $x(A,v)$ is strictly decreasing in A for $A<u^2(\omega)$.

It follows from the inequalities given above that the two curves have at least two transversal intersections at heights $0<z=\zeta_1<\zeta_2<z_1$. We will assume that ζ_1 is the height of the "first" such intersection and that ζ_2 is the next such intersection.

Since $\underline{x}(b,0)<\underline{x}(a,0)$ holds, we must have

$$\partial_z\underline{x}(a,\zeta_1)\le\partial_z\underline{x}(b,\zeta_1)<0\,,$$

$$\partial_z\underline{x}(b,\zeta_2)\le\partial_z\underline{x}(a,\zeta_2)<0\,.$$

By the intermediate value theorem, $\partial_z\underline{x}(a,\zeta^*)=\partial_z\underline{x}(b,\zeta^*)$ must hold for some $\zeta^*\in[\zeta_1,\zeta_2]$. Note that for $z\in(\zeta_1,\zeta_2)$, $\underline{x}(a,z)<\underline{x}(b,z)$ holds. A contradiction is reached because at the points $(\underline{x}(a,\zeta^*),\zeta^*)$ and $(\underline{x}(b,\zeta^*),\zeta^*)$, the tangent vectors agree and hence the values of v at both points agree. Thus $\underline{x}(a,\zeta^*)>\underline{x}(b,\zeta^*)$ must hold by monotonicity of $x(A,v)$ with respect to A. This shows that distinct generating curves do not intersect.

By (28), it follows that $-\Lambda\to 2\omega$ as $a\to-\infty$. It then follows from the formula for $x(a,v)$ that

$$x(a,0)\to\infty,\quad x(a,\pm\omega)\to\infty,\quad \text{as}\quad a\to-\infty\,.$$

It then follows that since the deformation of the generating curves depends continuously on a, the family of surfaces $\Sigma(V)$ fill out the region exterior to $\Sigma(V_0)$. □

5. Deflating a cylinder. Theorem 2.1 (I) asserts that for each $\omega\in[0,\bar{\omega})$ there is a least volume, stable capillary surface having two boundary components on the planes $z=z_i$, $i=0,1$. One may ask what will occur if the volume of this surface is decreased. It is expected that one or both boundary components will detach from the supporting planes and that the surface of the drop forms into one or more sessile or pendent drops or rescaled Wulff shapes.

In order for the drop to remain, it must be the case that the sessile drop with contact angle $\vartheta(\omega)$ or the entire rescaled Wulff shape with height z_1-z_0 has volume at least as large as V_0. We consider here only the case $\omega=0$. Since, for the fixed height, the entire rescaled Wulff shape contains one-fourth the volume of half of the Wulff shape, we consider the first possibility.

We assume for convenience that $z_1-z_0=1$. Theorem 2.2 (i) implies that the radius R of the least volume stable capillary cylinder satisfies

$$R=\big(\mu_1(0)/\mu_2(0)\big)^{1/2}\pi^{-1}.$$

Therefore, the minimum volume is

$$V_0(\omega=0)=\frac{\mu_1(0)}{\pi\mu_2(0)}\,.$$

Representing, the generating curve of the Wulff shape as $u = u(v)$, we have $\mu_1(0) = -u_{vv}(0)$, $\mu_2(0) = 1/u(0)$, and hence

$$V_0(\omega = 0) = \frac{u(0)|u_{vv}(0)|}{\pi} \,. \tag{39}$$

Proposition 5.1. *Assume* (W1)–(W3). *Then there holds*

$$V_2(\omega = 0) > V_0(\omega = 0) \,.$$

Proof. Let $u = u(v)$ be the generating curve of W. We claim that

$$u(t\bar{\omega}) \geq (1-t)u(0) - (u_{vv}(0)/2)\bar{\omega}^2 t(1-t) \tag{40}$$

holds. Assume this for now.

Using the inequality $(a+b)^2 \geq 4ab$ for all $a, b \geq 0$, we obtain

$$u^2(t\bar{\omega}) \geq 2u(0)|u_{vv}(0)|\bar{\omega}^2 t(1-t)^2 \,.$$

Therefore,

$$\begin{aligned}\int_0^{\bar{\omega}} u^2 \, dv &= \bar{\omega} \int_0^1 u^2(t\bar{\omega}) \, dt \\ &\geq 2u(0)|u_{vv}(0)|\bar{\omega}^3 \int_0^1 t(1-t)^2 \, dt \\ &= (1/6)u(0)|u_{vv}(0)|\bar{\omega}^3 \,.\end{aligned}$$

It follows that

$$\begin{aligned}V_2(\omega = 0) = (\pi/\bar{\omega}^3) \int_0^{\bar{\omega}} u^2 \, dv &\geq (\pi/6)u(0)|u_{vv}(0)| \\ &> (1/\pi)u(0)|u_{vv}(0)| = V_0(\omega = 0) \,.\end{aligned}$$

We now show (40). Let $H(t) = u(t\bar{\omega}) - (1-t)u(0) + (u_{vv}(0)/2)\bar{\omega}^2 t(1-t)$. Note that $H(0) = H(1) = 0$. If (40) doesn't hold, then H attains a negative minimum at some point in $(0,1)$ where $H'' \geq 0$ holds. A simple calculation shows that $H''(t) = \bar{\omega}^2(u_{vv}(t\bar{\omega}) - u_{vv}(0))$, which is negative since $-u_{vv} > -u_{vv}(1+u_v^2)^{-3/2} \geq -u_{vv}(0)$ holds on $(0, \bar{\omega})$ by assumption (W3) on the curvature of the Wulff shape. ☐

6. Appendix. Anisotropic Delaunay surfaces. We summarize important results about surfaces of revolution with constant anisotropic mean curvature for a rotationally symmetric energy functional (anisotropic Delaunay surfaces). Such surfaces were studied in detail by the authors in [4] (see also [5] and [6]).

Let

$$\chi(\sigma, \theta) = (u(\sigma)e^{i\theta}, v(\sigma))$$

be a parametrization of the Wulff shape W, where $(u(\sigma), v(\sigma))$ is the arc length parametrization of the generating curve. We have identified $\mathbf{R}^3$ with $\mathbf{C} \times \mathbf{R}$ in the formula above. We may extend $(u(\sigma), v(\sigma))$ so that it is defined for all real numbers σ. In this case, $(u(\sigma), v(\sigma))$ represents the section of W by (x_1, x_3)-plane.

Consider an anisotropic Delaunay surface Σ parameterized by

$$X(s,\theta) = (x(s)e^{i\theta}, z(s))\,,$$

where $(x(s), z(s))$ is the arc length parameterization of the generating curve, and $x(s) \geq 0$ holds for all s. The Gauss map of the surface X is given by

$$\nu = (z'(s)e^{i\theta}, -x'(s)).$$

We choose the orientation of the generating curve so that ν points "outward" from the surface. There is a natural map from the surface to the Wulff shape W defined by the requirement that the oriented tangent planes to both surfaces agree at corresponding points. Thus, at corresponding points the outward pointing unit normals must agree and we have

$$x' = u_\sigma\,, \qquad z' = v_\sigma. \tag{41}$$

In [4], we showed that the profile curve (x, z) satisfies the equation

$$2{\mu_2}^{-1}xz' + \Lambda x^2 = c\,, \tag{42}$$

where Λ is the anisotropic mean curvature and c is a real constant called the flux parameter. Also, $-\mu_2$ is the principal curvature of the Wulff shape in the θ direction. Since W is a surface of revolution, we have $\mu_2 = \mu_2(\nu_3) = \mu_2(-u_\sigma) = \mu_2(-x')$ by (41). Computing the principal curvature $-\mu_2 = -v_\sigma/u$, (42) can be expressed as

$$2ux + \Lambda x^2 = c\,. \tag{43}$$

The orientation of an anisotropic Delaunay surface may be chosen so that $\Lambda \leq 0$ holds and then the anisotropic Delaunay surfaces fall into six cases as follows:

(I-1) $\Lambda = 0$ and $c = 0$: *horizontal plane.*
(I-2) $\Lambda = 0$ and $c \neq 0$: *anisotropic catenoid.*
(II-1) $\Lambda < 0$ and $c = 0$: *Wulff shape (up to vertical translation and homothety).*
(II-2) $\Lambda < 0$ and $c = ((\mu_2|_{\nu_3=0})^2|\Lambda|)^{-1}$: *cylinder of radius* $(\mu_2|_{\nu_3=0}|\Lambda|)^{-1}$.
(II-3) $\Lambda < 0$ and $((\mu_2|_{\nu_3=0})^2|\Lambda|)^{-1} > c > 0$: *anisotropic unduloid.*
(II-4) $\Lambda < 0$ and $c < 0$: *anisotropic nodoid.*

Any surface in each case above is complete, and it has similar properties to the corresponding CMC surface in the sense of the following theorem.

THEOREM 6.1 (see [4], [5], [6]).

(i) *The generating curve $C : (x(s), z(s))$ of an anisotropic catenoid is a graph over the whole z axis, and $z'(s) \neq 0$ for all s. C is perpendicular to the horizontal line at a unique point.*
(ii) *Let $(x(s), z(s))$, $x \geq 0$, be the generating curve of an anisotropic unduloid or an anisotropic nodoid. Then there is a unique local maximum B and a unique local minimum $N > 0$ of x, which we will call a bulge and a neck, respectively.*
(iii) *The generating curve $C : (x(s), z(s))$ of an anisotropic unduloid is a graph over the z axis, and $z'(s) > 0$ for all s. C is a periodic curve with respect to the vertical translation, and the region from one neck to the next neck (and/or one bulge to the next bulge) gives one period. Therefore, C has a unique inflection point (x, z) between each neck and the next bulge, which satisfies $x = \sqrt{c/(-\Lambda)}$.*

(iv) *The curvature of the generating curve C of an anisotropic nodoid has a definite sign. C is a nonembedding periodic curve with respect to the vertical translation. The region from one neck to the next neck (and/or one bulge to the next bulge) gives one period.*

In the previous sections, we needed a representation formula for the profile curves, which is summarized in the following result from [4].

THEOREM 6.2 (see [4]). *Let W be the Wulff shape of a rotationally symmetric anisotropic surface energy $\mathcal{F}$. Let*

$$\sigma \mapsto (u(\sigma), v(\sigma)), \quad \sigma \in (-\infty, \infty),$$

be the profile curve of W, where σ is the arc length. Then

$$\mu_2^{-1} v_\sigma - u = 0$$

holds. Let $X(s,\theta) = (x(s)e^{i\theta}, z(s))$ be a surface with constant anisotropic mean curvature $\Lambda \leq 0$, and let the Gauss map of X coincide with that of W at $s = s(\sigma)$. Then X is given as follows.

(i) *When X is an anisotropic catenoid,*

$$x = c/(2u)$$

for some nonzero constant c.

(ii) *When X is an anisotropic unduloid,*

$$x = \frac{u \pm \sqrt{u^2 + \Lambda c}}{-\Lambda}$$

for some constants $c > 0$ and $\Lambda < 0$, where $x = x(u(\sigma))$ is defined in $\{\sigma | u \geq \sqrt{-\Lambda c}\}$.

(iii) *When X is an anisotropic nodoid,*

$$x = \frac{u + \sqrt{u^2 + \Lambda c}}{-\Lambda}$$

for some constants $c < 0$ and $\Lambda < 0$, where $x = x(u(\sigma))$ is defined in $\{-\infty < \sigma < \infty\}$.

In all cases above, z is given by

$$z = \int^u v_u x_u \, du. \tag{44}$$

Conversely, for a Wulff shape W defined as above, define x and z as in (i)–(iii) *and* (44). *Then $X(s,\theta) = (x(s)e^{i\theta}, z(s))$ is an anisotropic Delaunay surface which satisfies*

$$2\mu_2^{-1} z_s x + \Lambda x^2 = c,$$

where s is the arc length of (x, z), and Λ is supposed to be zero for case (i). *Moreover, X has the same regularity as that of W.*

Remark 6.1. In (ii) of Theorem 6.2, $x = (-\Lambda)^{-1}(u + \sqrt{u^2 + \Lambda c})$ gives the part of the anisotropic unduloid whose Gaussian curvature is positive (i.e., the convex part), while $x = (-\Lambda)^{-1}(u - \sqrt{u^2 + \Lambda c})$ gives the part of the anisotropic unduloid whose Gaussian curvature is negative.

Remark 6.2. In (iii) of Theorem 6.2, $u > 0$ corresponds to the part of the anisotropic nodoid whose Gaussian curvature is positive (i.e., the convex part), while $u < 0$ gives the part of the anisotropic nodoid whose Gaussian curvature is negative.

Acknowledgment. We would like to convey our sincere thanks to the referees for calling the references [3], [9], and [10] to our attention and suggesting various improvements to our paper.

REFERENCES

[1] M. Athanassenas, *A variational problem for constant mean curvature surfaces with free boundary*, J. Reine Angew. Math., 377 (1987), pp. 97–107.
[2] J. E. Brothers and F. Morgan, *The isoperimetric theorem for general integrands*, Michigan Math. J., 41 (1994), pp. 419–431.
[3] R. Finn and T. I. Vogel, *On the volume infimum for liquid bridges*, Z. Anal. Anwendungen, 11 (1992), pp. 3–23.
[4] M. Koiso and B. Palmer, *Geometry and stability of surfaces with constant anisotropic mean curvature*, Indiana Univ. Math. J., 54 (2005), pp. 1817–1852.
[5] M. Koiso and B. Palmer, *Stability of anisotropic capillary surfaces between two parallel planes*, Calc. Var. Partial Differential Equations, 25 (2006), pp. 275–298.
[6] M. Koiso and B. Palmer, *Anisotropic capillary surfaces with wetting energy*, Calc. Var. Partial Differential Equations, 29 (2007), pp. 295–345.
[7] T. I. Vogel, *Stability of a liquid drop trapped between two parallel planes*, SIAM J. Appl. Math., 47 (1987), pp. 516–525.
[8] T. I. Vogel, *Stability of a liquid drop trapped between two parallel planes* II. *General contact angles*, SIAM J. Appl. Math., 49 (1989), pp. 1009–1028.
[9] L. Zhou, *On the volume infimum for liquid bridges*, Z. Anal. Anwendungen, 12 (1993), pp. 629–642.
[10] L. Zhou, *Stability of Liquid Bridges*, Ph.D. Thesis, Stanford University, Stanford, CA, 1995.

SIAM J. MATH. ANAL.
Vol. 39, No. 3, pp. 742–800

NAVIER–STOKES EQUATIONS INTERACTING WITH A NONLINEAR ELASTIC BIOFLUID SHELL*

C. H. ARTHUR CHENG†, DANIEL COUTAND†, AND STEVE SHKOLLER‡

Abstract. We study a moving boundary value problem consisting of a viscous incompressible fluid moving and interacting with a nonlinear elastic fluid shell. The fluid motion is governed by the Navier–Stokes equations, while the fluid shell is modeled by a bending energy which extremizes the Willmore functional and a membrane energy with density given by a convex function of the local area ratio. The fluid flow and shell deformation are coupled together by continuity of displacements and tractions (stresses) along the moving surface defining the shell. We prove the existence and uniqueness of solutions in Sobolev spaces for a short time.

Key words. Navier–Stokes equations, free boundary problems, shell theory, biofluids, Willmore energy

AMS subject classifications. 74F10, 35Q30, 74K25, 35Q72, 74B20, 74H20, 74H25, 76D05

DOI. 10.1137/060656085

1. Introduction.

1.1. The problem statement and background. We are concerned with establishing the existence and uniqueness of solutions to the time-dependent incompressible Navier–Stokes equations interacting with a nonlinear elastic fluid shell (biomembrane) for a short time. Recently, there have been many experimental and analytic studies on the configurations and deformations of elastic biomembranes (see, for example, [3], [11], [13], [16], [17], [18], [19], and [21]), but the basic analysis of the coupled fluid-structure interaction remains open. The fundamental difficulties arise from the degenerate elliptic operators that arise as the shell tractions. As we detail below, the bending energy of the shell is the well-known Willmore function, the integral over the moving surface of the square of the mean curvature. The degenerate elliptic operator arising from the first variation of this functional is a fourth order nonlinear operator that smoothes only in the direction which is normal to the moving domain. Our analysis will provide a well-posedness theorem and explain the interesting interaction between the shape of the shell and the flow of the fluid.

Fluid-structure interaction problems involving moving material interfaces have been the focus of active research since the 1990s. The first problem solved in this area was for the case of a rigid body moving in a viscous fluid (see [9], [14], and the early works of [22] and [20] for a rigid body moving in a Stokes flow in the full space). The case of an elastic body moving in a viscous fluid was considerably more challenging because of some apparent regularity incompatibilities between the parabolic fluid phase and the hyperbolic solid phase. The first existence results in this area were for regularized elasticity laws, such as in [10] for a *finite* number of

*Received by the editors April 2, 2006; accepted for publication (in revised form) March 21, 2007; published electronically August 17, 2007. This research was supported by the National Science Foundation under grants DMS-0313370 and EAR-0327799.
http://www.siam.org/journals/sima/39-3/65608.html

†Department of Mathematics, University of California, Davis, CA 95616 (cchsiao@math.ucdavis.edu, coutand@math.ucdavis.edu).
‡Corresponding author. Department of Mathematics, University of California, Davis, CA 95616 (shkoller@math.ucdavis.edu).

elastic modes, or in [1], [4], and [2] for hyperviscous elasticity laws, or in [15] in which a phase-field regularization "fattens" the sharp interface via a diffuse-interface model.

The treatment of classical elasticity laws for the solid phase, without any regularizing term, was considered only recently in [7] for the three-dimensional linear St. Venant–Kirchhoff constitutive law and in [8] for quasi-linear elastodynamics coupled to the Navier–Stokes equations. Some of the basic new ideas introduced in those works concerned a functional framework that scales in a hyperbolic fashion (and is therefore driven by the solid phase), the introduction of approximate problems either penalized with respect to the divergence-free constraint in the moving fluid domain or smoothed by an appropriate parabolic artificial viscosity in the solid phase (chosen in an asymptotically convergent and consistent fashion), and the tracking of the motion of the interface by difference quotient techniques.

In our companion paper [5], we study the interaction of the Navier–Stokes equations with an elastic solid shell. Herein, we treat the case of a fluid shell or biomembrane. This is a moving boundary problem that models the motion of a viscous incompressible Newtonian fluid inside of a deformable elastic fluid structure.

Let $\Omega \subset \mathbb{R}^3$ denote an open bounded domain with boundary $\Gamma := \partial\Omega$. For each $t \in (0,T]$, we wish to find the domain $\Omega(t)$, a divergence-free velocity field $u(t,\cdot)$, a pressure function $p(t,\cdot)$ on $\Omega(t)$, and a volume-preserving transformation $\eta(t,\cdot) : \Omega \to \mathbb{R}^3$ such that

$$\Omega(t) = \eta(t,\Omega), \tag{1.1a}$$
$$\eta_t(t,x) = u(t,\eta(t,x)), \tag{1.1b}$$
$$u_t + \nabla_u u - \nu\Delta u = -\nabla p + f \quad \text{in } \Omega(t), \tag{1.1c}$$
$$\operatorname{div} u = 0 \quad \text{in } \Omega(t), \tag{1.1d}$$
$$(\nu \operatorname{Def} u - p \operatorname{Id})n = \mathfrak{t}_{shell} \quad \text{on } \Gamma(t), \tag{1.1e}$$
$$u(0,x) = u_0(x) \quad \forall\, x \in \Omega, \tag{1.1f}$$
$$\eta(0,x) = x \quad \forall\, x \in \Omega, \tag{1.1g}$$

where ν is the kinematic viscosity, $n(t,\cdot)$ is the outward pointing unit normal to $\Gamma(t)$, $\Gamma(t) := \partial\Omega(t)$ denotes the boundary of $\Omega(t)$, $\operatorname{Def} u$ is twice the rate of deformation tensor of u, given in coordinates by $u^i_{,j} + u^j_{,i}$, and $\mathfrak{t}_{shell}$ is the traction imparted onto the fluid by the elastic shell, which we describe next.

We shall consider a thin elastic shell modeled by the nonlinear Saint Venant–Kirchhoff constitutive law. With ε denoting the thickness of the shell, the hyperelastic stored energy function has the asymptotic expansion

$$E_{shell} = \varepsilon E_{mem} + \varepsilon^3 E_{ben} + \mathcal{O}(\varepsilon^4).$$

The membrane energy satisfies

$$E_{mem} = \int_\Gamma \mathcal{P}(\mathcal{J})dS, \tag{1.2}$$

where $\mathcal{J}$ is the local area ratio and $\mathcal{P}$ is a convex function attaining its minimum at $\mathcal{J} = 1$, while the bending energy E_{ben} is given by

$$E_{ben} = \int_{\Gamma(t)} (\sigma H^2 - \sigma_1 K)dS, \tag{1.3}$$

where H and K denote the mean and Gauss curvatures on $\Gamma(t)$, respectively, and where σ and σ_1 are positive constants. The traction vector

$$\mathfrak{t}_{shell} = \varepsilon \mathfrak{t}_{mem} + \varepsilon^3 \mathfrak{t}_{ben} + \mathcal{O}(\varepsilon^4)$$

is computed from the first variation of the energy function E_{shell}; the traction vector associated with the membrane energy is

$$\mathfrak{t}_{\text{mem}} = \Big[\mathcal{J}\mathcal{P}''(\mathcal{J}) + 2\mathcal{P}'(\mathcal{J})\Big]\mathcal{J}_{,\beta} g^{\alpha\beta}\eta_{,\alpha} + \Big[\mathcal{J}\mathcal{P}'(\mathcal{J}) + \mathcal{P}(\mathcal{J})\Big]Hn, \tag{1.4}$$

while the traction associated with the bending energy has a simple intrinsic characterization given by

$$\mathfrak{t}_{ben} = \sigma(\Delta_g H - 2HK + 2H^3)n, \tag{1.5}$$

where Δ_g denotes the Laplacian with respect to the induced metric g on $\Gamma(t)$:

$$\Delta_g f = \frac{1}{\sqrt{\det(g)}} \frac{\partial}{\partial x^\alpha}\left(\sqrt{\det(g)} g^{\alpha\beta} \frac{\partial f}{\partial x^\beta}\right).$$

In this paper, we ignore the inertia of the shell and focus our analysis on the difficulties associated with the degenerate elliptic operators in $\mathfrak{t}_{shell}$.

1.2. Outline of the paper. In section 2, in addition to the use of Lagrangian variables, we introduce a new coordinate system near the boundary (shell) and three new maps, η^ν, η^τ, and h, which facilitate the computation of the membrane and bending tractions $\mathfrak{t}_{\text{mem}}$ and $\mathfrak{t}_{\text{ben}}$. A key observation is the symmetry relation (2.7) which reduces the derivative count on the tangential reparameterization map η^τ.

The space of solutions (to the problem $\mathfrak{t}_{mem} = 0$) is introduced in section 3, and the main theorem is stated in section 4. Section 5 defines our notation, and section 6 provides some useful lemmas.

In section 7, we introduce the linearized and regularized problems. The regularization requires smoothing certain variables as well as the introduction of a certain artificial viscosity term on the boundary of the fluid domain. Weak solutions of this linear problem are established via a penalization scheme which approximates the incompressibility of the fluid.

In section 8, we establish a regularity theory for our weak solution using energy estimates for the mollified problem with constants that depend on the mollification parameters. In section 9, we improve these estimates so that the constants are independent of the artificial viscosity as well as other regularization parameters. This requires an elliptic estimate, arising from the boundary condition (1.1e), which provides additional regularity for the shape of the boundary.

In section 10, the Tychonoff fixed-point theorem is used to prove the existence of solutions to the original nonlinear problem (1.1). Uniqueness, following required compatibility conditions, is established in sections 4 and 10.

In section 11, we consider the inclusion of the lower order membrane traction into the problem formulation so that the full problem is solved.

The inclusion of the inertial term $\epsilon_1\eta_{tt}$ into the membrane traction $\mathfrak{t}_{\text{mem}}$ will be studied in a future publication.

2. Lagrangian formulation.

2.1. A new coordinate system near the shell. Consider the isometric immersion $\eta_0 : (\Gamma, g_0) \to (\mathbb{R}^3, \text{Id})$. Let $\mathcal{B} = \Gamma \times (-\epsilon_1, \epsilon_1)$, where ϵ_1 is chosen sufficiently small so that the map

$$B : \mathcal{B} \to \mathbb{R}^3 : (y, z) \mapsto y + zN(y)$$

is itself an immersion, defining a tubular neighborhood of Γ in $\mathbb{R}^3$. We can choose a coordinate system $\frac{\partial}{\partial y^\alpha}$, $\alpha = 1, 2$, and $\frac{\partial}{\partial z}$ on $\mathcal{B}$, where $\frac{\partial}{\partial y^\alpha}$ denotes the tangential derivative and $\frac{\partial}{\partial z}$ denotes the normal derivative.

Let $G = B^*(\mathrm{Id})$ denote the induced metric on $\mathcal{B}$ from $\mathbb{R}^3$ so that

$$G(y, z) = G_z(y) + dz \otimes dz,$$

where G_z is the metric on the surface $\Gamma \times \{z\}$; note that $G_0 = g_0$.

REMARK 1. *By assumption,* $g_{0\alpha\beta} = \frac{\partial}{\partial y^\alpha} \cdot \frac{\partial}{\partial y^\beta}$, *where* $\cdot$ *denotes the usual Cartesian inner product on* $\mathbb{R}^n$. *Let* $C_{\alpha\beta}$ *denote the covariant components of the second fundamental form of the base manifold* Γ *so that* $C_{\alpha\beta} = -N_{,\alpha} \cdot \frac{\partial}{\partial y^\beta}$. *Then* G_z *is given by*

$$(G_z)_{\alpha\beta} = (g_0)_{\alpha\beta} - 2zC_{\alpha\beta} + z^2 g_0^{\gamma\delta} C_{\alpha\gamma} C_{\beta\delta}.$$

Let $h : \Gamma \to (-\epsilon_1, \epsilon_1)$ be a smooth height function and consider the graph of h in $\mathcal{B}$, parameterized by $\phi : \Gamma \to \mathcal{B} : y \mapsto (y, h(y))$. The tangent space to graph(h), considered as a submanifold of $\mathcal{B}$, is spanned at a point $\phi(x)$ by the vectors

$$\phi_* \left(\frac{\partial}{\partial y^\alpha} \right) = \frac{\partial \phi}{\partial y^\alpha} = \frac{\partial}{\partial y^\alpha} + \frac{\partial h}{\partial y^\alpha} \frac{\partial}{\partial z},$$

and the normal to graph(h) is given by

$$n(y) = J_h^{-1}(y) \left(-G_{h(y)}^{\alpha\beta} \frac{\partial h}{\partial y^\alpha} \frac{\partial}{\partial y^\beta} + \frac{\partial}{\partial z} \right), \tag{2.1}$$

where $J_h = (1 + h_{,\alpha} G_{h(y)}^{\alpha\beta} h_{,\beta})^{1/2}$. The mean curvature H of graph(h) is defined to be the trace of ∇n, where

$$(\nabla n)_{ij} = G\left(\nabla^{\mathcal{B}}_{\frac{\partial}{\partial w^i}} n, \frac{\partial}{\partial w^j} \right) \quad \text{for } i, j = 1, 2, 3,$$

where $\frac{\partial}{\partial w^\alpha} = \frac{\partial}{\partial y^\alpha}$ for $\alpha = 1, 2$ and $\frac{\partial}{\partial w^3} = \frac{\partial}{\partial z}$, and $\nabla^{\mathcal{B}}$ denotes the covariant derivative. Using (2.1),

$$\begin{aligned}
(\nabla n)_{\alpha\beta} &= G\left(\nabla^{\mathcal{B}}_{\frac{\partial}{\partial y^\alpha}} \left[-J_h^{-1} G_h^{\gamma\delta} h_{,\gamma} \frac{\partial}{\partial y^\delta} + J_h^{-1} \frac{\partial}{\partial z} \right], \frac{\partial}{\partial y^\beta} \right) \\
&= -(G_h)_{\delta\beta} \Big[(J_h^{-1} G_h^{\gamma\delta} h_{,\gamma})_{,\alpha} + J_h^{-1} (-G_h^{\gamma\sigma} h_{,\gamma} \Gamma^\delta_{\alpha\sigma} + \Gamma^\delta_{\alpha 3}) \Big]; \\
(\nabla n)_{33} &= G\left(\nabla^{\mathcal{B}}_{\frac{\partial}{\partial z}} \left[-J_h^{-1} G_h^{\gamma\delta} h_{,\gamma} \frac{\partial}{\partial y^\delta} + J_h^{-1} \frac{\partial}{\partial z} \right], \frac{\partial}{\partial z} \right) \\
&= J_h^{-1} (-G_h^{\gamma\delta} h_{,\gamma} \Gamma^3_{3\delta} + \Gamma^3_{33}),
\end{aligned}$$

where Γ^k_{ij} denotes the Christoffel symbols with respect to the metric G. It follows that the curvature of graph(h) (in the divergence form) is

$$H = -(J_h^{-1} G_h^{\gamma\delta} h_{,\gamma})_{,\delta} + J_h^{-1} (-G_h^{\gamma\delta} h_{,\gamma} \Gamma^j_{j\delta} + \Gamma^j_{j3}), \tag{2.2}$$

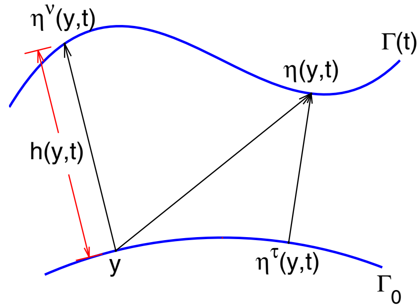


FIG. 1. *The maps η^τ and η^ν.*

or (in the quasi-linear form)

$$H = -J_h^{-1} G_h^{\alpha\beta} \Big[\delta_{\beta\gamma} - J_h^{-2} G_h^{\gamma\delta} h_{,\beta} h_{,\delta} \Big] h_{,\alpha\gamma} + G_h^{\alpha\beta} F_{\alpha\beta}(y, h, \nabla h), \tag{2.3}$$

where $F_{\alpha\beta}$ denotes a smooth generic function of y, h, and ∇h.

REMARK 2. *Note that G_h denotes the metric $G_{z=h(y)}$ and not the metric on the submanifold graph(h).*

REMARK 3. *If the initial height function is zero, i.e., $h(0) = 0$, then $H(0) = \Gamma^j_{j3}(0)$ which is the mean curvature of the base manifold Γ as required.*

2.2. Tangential reparameterization symmetry. Let $\mathcal{N}$ denote the normal bundle to Γ so that for each $y \in \Gamma$ we have the Whitney sum $\mathbb{R}^3 = T_y\Gamma \oplus \mathcal{N}_y$.

Given a signed height function $h : \Gamma \times [0, T) \to \mathbb{R}$, for each $t \in [0, T)$, define the *normal* map (see Figure 1)

$$\eta^\nu : \Gamma \times [0, T) \to \Gamma(t), \quad (y, t) \mapsto y + h(y, t)N(y), \quad N(y) \in \mathcal{N}_y.$$

Then there exists a unique *tangential* map $\eta^\tau : \Gamma \times [0, T) \to \Gamma$ (a diffeomorphism as long as h remains a graph) such that the diffeomorphism $\eta(t)$ has the decomposition

$$\eta(\cdot, t) = \eta^\nu(\cdot, t) \circ \eta^\tau(\cdot, t), \quad \eta(y, t) = \eta^\tau(y, t) + h(\eta^\tau(y, t), t)N(\eta^\tau(y, t)).$$

The tangent vector $\eta_{,\alpha}$ to $\Gamma(t)$ can be decomposed with respect to the Whitney sum as $\eta_{,\alpha}(y, t) = \eta^\kappa_{,\alpha}(y, t)\frac{\partial}{\partial y^\kappa} + h_{,\kappa}(\eta^\tau(y, t), t)\eta^\kappa_{,\alpha}\frac{\partial}{\partial z}$, and hence the induced metric $g_{\alpha\beta} = \eta_{,\alpha} \cdot \eta_{,\beta}$ may be expressed as

$$g_{\alpha\beta} = \Big[\Big((G_h)_{\kappa\sigma} + h_{,\kappa} h_{,\sigma} \Big) \circ \eta^\tau \Big] \eta^\kappa_{,\alpha} \eta^\sigma_{,\beta} := \Big[\mathcal{G}_{\kappa\sigma} \circ \eta^\tau \Big] \eta^\kappa_{,\alpha} \eta^\sigma_{,\beta}. \tag{2.4}$$

Note that $\mathcal{G}_{\kappa\sigma}$ is the induced metric with respect to the *normal* map η^ν. Furthermore, we have the following useful relationship between the determinant of the two induced metrics:

$$\det(g) = \det(\nabla_0 \eta^\tau)^2 \Big[\det(G_h) J_h^2 \Big] \circ \eta^\tau = \det(\nabla_0 \eta^\tau)^2 \Big[\det(\mathcal{G}) \Big] \circ \eta^\tau, \tag{2.5}$$

where ∇_0 denotes the surface gradient.

REMARK 4. *The identity* (2.4) *can also be read as $(\eta^\tau)^* g = \mathcal{G}$.*

Let y and $\tilde{y} = \varphi(y)$ denote two different coordinate systems on Γ with associated metrics

$$g_{\alpha\beta} = \frac{\partial \eta^i}{\partial y^\alpha}\frac{\partial \eta^i}{\partial y^\beta}, \qquad \tilde{g}_{\alpha\beta} = \frac{\partial \eta^i}{\partial \tilde{y}^\alpha}\frac{\partial \eta^i}{\partial \tilde{y}^\beta}.$$

It follows that $\varphi^*\tilde{g} = g$. Let H, $\tilde{H}$, K, $\tilde{K}$, n, and $\tilde{n}$ denote the mean curvature, Gauss curvature, and the unit normal vector computed with respect to y and $\tilde{y}$, respectively. Since H, K, and n depend only on the shape of $\Gamma(t)$, these geometric quantities are invariant to tangential reparameterization; thus, we have the identity

$$\tilde{H} = H \circ \varphi, \quad \tilde{K} = K \circ \varphi, \quad \tilde{n} = n \circ \varphi. \tag{2.6}$$

Similarly, computing the first variation of $\int_{\Gamma(t)} H^2 dS$ in our two coordinate systems yields

$$\Big[\Big(\Delta_g H + H(H^2 - K)\Big)n\Big](y) = \Big[\Big(\Delta_{\tilde{g}}\tilde{H} + \tilde{H}(\tilde{H}^2 - \tilde{K})\Big)\tilde{n}\Big](\tilde{y}) \qquad \forall\ \tilde{y} = \varphi(y).$$

By (2.6), we have the following important identity:

$$\Big[\Delta_{\varphi^*\tilde{g}} H\Big](y) = \Big[\Delta_{\tilde{g}}(H \circ \varphi)\Big](\tilde{y}) \qquad \forall\ \tilde{y} = \varphi(y), \tag{2.7}$$

and hence

$$[\Delta_{\mathcal{G}}(H \circ \eta^{-\tau})] \circ \eta^\tau = \Delta_g H, \tag{2.8}$$

where by (2.3),

$$H \circ \eta^{-\tau} = -J_h^{-1} G_h^{\alpha\beta}\Big[\delta_{\beta\gamma} - J_h^{-2} G_h^{\gamma\delta} h_{,\beta} h_{,\delta}\Big] h_{,\alpha\gamma} + G_h^{\alpha\beta} F_{\alpha\beta}(y, h, \nabla h). \tag{2.9}$$

2.3. Bounds on $\boldsymbol{\eta^\tau}$. Let u^τ denote the tangential velocity defined by $\eta^\tau_t = u^\tau \circ \eta^\tau$. Time differentiating the relation $\eta = \eta^\nu \circ \eta^\tau$ and using the definition of η^ν, we find that

$$u^\tau = (\nabla_0 \eta^\nu)^{-1}\Big[u \circ \eta^\nu - h_t \frac{\partial}{\partial z}\Big]. \tag{2.10}$$

From the trace theorem, it follows that

$$\|u^\tau\|_{H^{2.5}(\Gamma)} \le C\mathcal{P}(\|h\|_{H^{3.5}(\Gamma)}, \|\eta\|_{H^3(\Omega)})\Big[\|v\|_{H^3(\Omega)} + \|h_t\|_{H^{2.5}(\Gamma)}\Big] \tag{2.11}$$

for some polynomial $\mathcal{P}$. Since $\eta^\tau(y,t) = y + \int_0^t (u^\tau \circ \eta^\tau)(y,s)ds$, it follows that

$$\|\nabla_0 \eta^\tau(y,t)\|_{H^{1.5}(\Gamma)} \le C\Big[1 + \int_0^t \|u^\tau\|_{H^{2.5}(\Gamma)}\Big(1 + \|\nabla_0\eta^\tau\|_{H^{1.5}(\Gamma)}\Big)^4 ds\Big],$$

and hence by Gronwall's inequality,

$$\|\nabla_0 \eta^\tau(y,t)\|_{H^{1.5}(\Gamma)} \le C\Big[1 + \int_0^t \|u^\tau\|_{H^{2.5}(\Gamma)} ds\Big] \tag{2.12}$$

for $t \in [0,T]$ sufficiently small. Furthermore, we also have

$$\|\eta^\tau_t(y,t)\|_{H^{2.5}(\Gamma)} \le C\|u^\tau\|_{H^{2.5}(\Gamma)}\Big[1 + \|\nabla_0\eta^\tau\|_{H^{1.5}(\Gamma)}\Big]^4. \tag{2.13}$$

2.4. An expression for $\mathbf{t}_{\mathbf{ben}}$ in terms of h and η^τ. Now we can compute $\mathbf{t}_{ben}$ in terms of h and η^τ: the highest order term of $\Delta_g H$ is

$$\left\{\frac{1}{\sqrt{\det(\mathcal{G})}}\frac{\partial}{\partial y^\gamma}\left[\sqrt{\det(\mathcal{G})}\mathcal{G}^{\gamma\delta}\frac{\partial}{\partial y^\delta}\Big(J_h^{-1}(G_h^{\alpha\beta}-J_h^{-2}G_h^{\alpha\kappa}G_h^{\beta\sigma}h_{,\kappa}h_{,\sigma})h_{,\alpha\beta}\Big)\right]\right\}\circ\eta^\tau.$$

Since $\mathcal{G}_{\alpha\beta}=(G_h)_{\alpha\beta}+h_{,\alpha}h_{,\beta}$, the inverse of $\mathcal{G}_{\gamma\delta}$ is

$$\frac{1}{\det(\mathcal{G})}\begin{bmatrix}(G_h)_{22}+h_{,2}^2 & -(G_h)_{12}-h_{,1}h_{,2}\\ -(G_h)_{12}-h_{,1}h_{,2} & (G_h)_{11}+h_{,1}^2\end{bmatrix},$$

which can also be written as

$$\mathcal{G}^{\alpha\beta}=J_h^{-2}\Big[G_h^{\alpha\beta}-(-1)^{\kappa+\sigma}\det(G_h)^{-1}(1-\delta_{\alpha\kappa})(1-\delta_{\beta\sigma})h_{,\kappa}h_{,\sigma}\Big].$$

Therefore, the highest order term of $\Delta_g H$ can be written as

$$\frac{1}{\sqrt{\det(g_0)}}\Big[\sqrt{\det(g_0)}A^{\alpha\beta\gamma\delta}h_{,\alpha\beta}\Big]_{,\gamma\delta}\circ\eta^\tau,$$

where

$$\begin{aligned}A^{\alpha\beta\gamma\delta}=J_h^{-3}&\Big[G_h^{\alpha\gamma}-(-1)^{\kappa+\sigma}\det(G_h)^{-1}(1-\delta_{\alpha\kappa})(1-\delta_{\gamma\sigma})h_{,\kappa}h_{,\sigma}\Big]\\ &\times(G_h^{\beta\delta}-J_h^{-2}G_h^{\beta\kappa}G_h^{\delta\sigma}h_{,\kappa}h_{,\sigma})\end{aligned}\tag{2.14}$$

is a fourth-rank tensor.

2.5. Lagrangian formulation of the problem. Let $\eta(t,x)=x+\int_0^t u(s,x)ds$ denote the Lagrangian particle placement field, a volume-preserving embedding of Ω onto $\Omega(t)\subset\mathbb{R}^3$, and denote the cofactor matrix of $\nabla\eta(x,t)$ by

$$a(x,t)=[\nabla\eta(x,t)]^{-1}.\tag{2.15}$$

Let $v=u\circ\eta$ denote the Lagrangian or material velocity field, $q=p\circ\eta$ the Lagrangian pressure function, and $F=f\circ\eta$ the forcing function in the material frame. In the following discussion, we also set $\varepsilon=1$. Then the system (1.1) can be reformulated as

$$\eta_t=v \quad \text{in } (0,T)\times\Omega,\tag{2.16a}$$

$$v_t^i-\nu(a_\ell^j D_\eta(v)_\ell^i)_{,j}=-(a_i^k q)_{,k}+F^i \quad \text{in } (0,T)\times\Omega,\tag{2.16b}$$

$$a_i^k v_{,k}^i=0 \quad \text{in } (0,T)\times\Omega,\tag{2.16c}$$

$$(\nu D_\eta(v)_\ell^i-q\delta_\ell^i)a_\ell^j N_j=\sigma\Theta\Big[L(h)B_*(-G_h^{\alpha\beta}h_{,\alpha},1)\Big]\circ\eta^\tau \quad \text{on } (0,T)\times\Gamma,\tag{2.16d}$$

$$h_t=B_*((-G_h^{\alpha\beta}h_{,\alpha},1))\cdot(v\circ\eta^{-\tau}) \quad \text{on } (0,T)\times\Gamma,\tag{2.16e}$$

$$v=u_0 \quad \text{on } \{t=0\}\times\Omega,\tag{2.16f}$$

$$h=0 \quad \text{on } \{t=0\}\times\Gamma,\tag{2.16g}$$

$$\eta=\mathrm{Id} \quad \text{on } \{t=0\}\times\Omega,\tag{2.16h}$$

where $D_\eta(v)_\ell^i:=(a_\ell^k v_{,k}^i+a_i^k v_{,k}^\ell)$, N denotes the outward-pointing unit normal to Γ, Θ is defined in Remark 5, and B_* is the pushforward of B defined as

$$B_*(\gamma'(0))=(B\circ\gamma)'(0)\qquad\forall\ \gamma(t)\subset\Gamma.$$

$L(h)$ is the representation of $\mathfrak{t}_{shell} \cdot n$ using the height function h. It is defined as follows:

$$L(h) = \frac{1}{\sqrt{\det(g_0)}} \Big[\sqrt{\det(g_0)} A^{\alpha\beta\gamma\delta} h_{,\alpha\beta} \Big]_{,\gamma\delta} + L_1^{\alpha\beta\gamma}(y, h, Dh, D^2 h) h_{,\alpha\beta\gamma} \\ + L_2(y, h, Dh, D^2 h),$$

where L_1 and L_2 are polynomials of their variables with $L_1(y, 0) = 0$, and g_0 is the metric tensor on Γ. Note that $\mathfrak{t}_{mem}$ is included in L_2, since it is a second order operator of h.

REMARK 5. *For a point $\eta(y,t) \in \Gamma(t)$, there are two ways of defining the unit normal n to $\Gamma(t)$:*

1. *Let $n = \sqrt{g}^{-1} a^T N$, where N is the unit normal to Γ.*
2. *Let $n = [J_h^{-1}(-G_h^{\alpha\beta} h_{,\alpha} \frac{\partial}{\partial y^\beta} + \frac{\partial}{\partial z})] \circ \eta^\tau$ (denoted by $[J_h^{-1}(-\nabla_0 h, 1)] \circ \eta^\tau$).*

The function Θ is defined by

$$\Theta(-\nabla_0 h \circ \eta^\tau, 1) = a^T N.$$

Equating the modulus of both sides, by (2.5) *we must have*

$$\Theta = \sqrt{\det(g)} [(J_h^{-1}) \circ \eta^\tau] = \det(\nabla_0 \eta^\tau) \sqrt{\det(G_h) \circ \eta^\tau}.$$

REMARK 6. *An equivalent form of* (2.16e) *is given by*

$$h_t = -h_{,\alpha} (v \circ \eta^{-\tau})_\alpha + (v \circ \eta^{-\tau})_z.$$

This equation states that the shape of the boundary moves with the normal velocity of the fluid.

REMARK 7. *For many of the nonlinear estimates that appear later, it is important that $L(h)$ is linear in the third derivative $h_{,\alpha\beta\gamma}$.*

REMARK 8. *Without using the symmetry* (2.8), *we can still compute $\Delta_g H$ in terms of h and η^τ by using* (2.4) *and* (2.5); *however, L_1 would then depend on $\nabla_0^2 \eta^\tau$ and thus lose one derivative of regularity, preventing the closure of our energy estimate.*

3. Notation and conventions. For $T > 0$, we set

$$\mathcal{V}^1(T) = \Big\{ v \in L^2(0,T; H^1(\Omega)) \;\Big|\; v_t \in L^2(0,T; H^1(\Omega)') \Big\};$$
$$\mathcal{V}^2(T) = \Big\{ v \in L^2(0,T; H^2(\Omega)) \;\Big|\; v_t \in L^2(0,T; L^2(\Omega)) \Big\};$$
$$\mathcal{V}^k(T) = \Big\{ v \in L^2(0,T; H^k(\Omega)) \;\Big|\; v_t \in L^2(0,T; H^{k-2}(\Omega)) \Big\} \quad \text{for } k \geq 3;$$
$$\mathcal{H}(T) = \Big\{ h \in L^2(0,T; H^{5.5}(\Gamma)) \;\Big|\; h_t \in L^2(0,T; H^{2.5}(\Gamma)),\, h_{tt} \in L^2(0,T; H^{0.5}(\Gamma)) \Big\}$$

with norms

$$\|v\|^2_{\mathcal{V}^1(T)} = \|v\|^2_{L^2(0,T;H^1(\Omega))} + \|v_t\|^2_{L^2(0,T;H^1(\Omega)')};$$
$$\|v\|^2_{\mathcal{V}^2(T)} = \|v\|^2_{L^2(0,T;H^2(\Omega))} + \|v_t\|^2_{L^2(0,T;L^2(\Omega))};$$
$$\|v\|^2_{\mathcal{V}^k(T)} = \|v\|^2_{L^2(0,T;H^k(\Omega))} + \|v_t\|^2_{L^2(0,T;H^{k-2}(\Omega))} \quad \text{for } k \geq 3;$$
$$\|h\|^2_{\mathcal{H}(T)} = \|h\|^2_{L^2(0,T;H^{5.5}(\Gamma))} + \|h_t\|^2_{L^2(0,T;H^{2.5}(\Gamma))} + \|h_{tt}\|^2_{L^2(0,T;H^{0.5}(\Gamma))}.$$

We then introduce the space (of "divergence-free" vector fields)

$$\mathcal{V}_v = \Big\{ w \in H^1(\Omega) \,\Big|\, a_i^j(t) w^i_{,j} = 0 \ \forall\, t \in [0,T] \Big\}$$

and

$$\mathcal{V}_v(T) = \Big\{ w \in L^2(0,T;H^1(\Omega)) \,\Big|\, a_i^j(t) w^i_{,j} = 0 \ \forall\, t \in [0,T] \Big\},$$

where the cofactor matrix a is defined by (2.15). We use X_T to denote the space $\mathcal{V}^3(T) \times \mathcal{H}(T)$ with norm

$$\|(v,h)\|^2_{X_T} = \|v\|^2_{\mathcal{V}^3(T)} + \|h\|^2_{\mathcal{H}(T)}$$

and use Y_T, a subspace of X_T, to denote the space

$$Y_T = \Big\{ (v,h) \in \mathcal{V}^3(T) \times \mathcal{H}(T) \,\Big|\, h_t \in L^\infty(0,T;H^2(\Gamma)) \Big\}$$

with norm

$$\begin{aligned}\|(v,h)\|^2_{Y_T} = \|(v,h)\|^2_{X_T} &+ \|v\|^2_{L^\infty(0,T;H^2(\Omega))} + \|h\|^2_{L^\infty(0,T;H^4(\Gamma))} \\ &+ \|h_t\|^2_{L^\infty(0,T;H^2(\Gamma))}.\end{aligned}$$

We will solve (2.16) by a fixed-point method in an appropriate subset of Y_T.

4. The main theorem. Before stating the main theorem, we define the following quantities. Let q_0 be defined by

$$\Delta q_0 = -\nabla u_0 : (\nabla u_0)^T + \nu [a^k_\ell D_\eta(u_0)^i_\ell]_{,ki}(0) + \operatorname{div} F(0) \quad \text{in} \quad \Omega, \tag{4.1a}$$

$$q_0 = \nu(\operatorname{Def} u_0 \cdot N)\cdot N - \sigma L(0) \quad \text{on} \quad \Gamma \tag{4.1b}$$

and

$$u_1 = \nu \Delta u_0 - \nabla q_0 + F(0). \tag{4.2}$$

We also define the projection operator $\mathcal{P}_{ij}(x) : \mathbb{R}^3 \to T_{\eta(x,t)}\Gamma(t)$ by

$$\mathcal{P}_{ij}(x) = [\delta_{ij} - (J_h^{-2} \circ \eta^\tau) a^k_i a^\ell_j N_k(x) N_\ell(x)] = \left[\delta_{ij} - \frac{a^k_i N_k(x)}{|a^k_i N_k(x)|} \frac{a^\ell_j N_\ell(x)}{|a^\ell_j N_\ell(x)|} \right].$$

Theorem 4.1. *Let $\nu > 0$, $\sigma > 0$ be given, and*

$$F \in L^2(0,T;H^2(\Omega)), \quad F_t \in L^2(0,T;L^2(\Omega)), \quad F(0) \in H^1(\Omega).$$

Suppose that the shell traction satisfies the compatibility condition

$$[\operatorname{Def} u_0 \cdot N]_{tan} = 0. \tag{4.3}$$

There exists $T > 0$ depending on u_0 and F such that there exists a solution $(v,h) \in Y_T$ of problem (2.16). Moreover, if $u_0 \in H^{5.5}(\Omega) \cap H^{7.5}(\Gamma)$ and the associated u_1, q_0 also satisfy the compatibility condition

$$\begin{aligned} CP := &\Big[g_0^{ki} u^j_{0,k} N_j N_\ell + g_0^{k\ell} u^j_{0,k} N_j N_i\Big]\Big[\nu (\operatorname{Def} u_0)^j_i - q_0 \delta^j_i\Big] N_j \\ &+ \nu(\delta_{i\ell} - N_i N_\ell)\Big[(\operatorname{Def} u_1)^j_i - \Big((\nabla u_0 \nabla u_0) + (\nabla u_0 \nabla u_0)^T\Big)^j_i\Big] N_j \\ &- (\delta_{i\ell} - N_i N_\ell)\Big[\nu(\operatorname{Def} u_0)^j_i - q_0 \delta^j_i\Big] u^k_{0,j} N_k = 0, \end{aligned} \tag{4.4}$$

then the solution $(v,h) \in Y_T$ is unique.

5. A bounded convex closed set of Y_T.

DEFINITION 5.1. *Given $M > 0$, let $C_T(M)$ denote the subset of Y_T consisting of elements of (v,h) in Y_T such that*

$$\|(v,h)\|^2_{Y_T} \le M \tag{5.1}$$

and such that $v(0) = u_0$, $h(0) = 0$, and $h_t(0) = (B_0)_*((0,1)) \cdot u_0$.

REMARK 9. *For $(v,h) \in C_T(M)$, define u^τ by (2.10) and let η^τ be the associated flow map. Also define v^τ as $u^\tau \circ \eta^\tau$. By (2.12) and (2.13), we have*

$$\sup_{t\in[0,T]} \|\nabla_0 \eta^\tau(t)\|_{H^{1.5}(\Gamma)} + \|v^\tau\|^2_{L^2(0,T;H^{2.5}(\Gamma))} \le C(M) \tag{5.2}$$

for some constant $C(M)$.

We will make use of the following lemmas (proved in [7]).

LEMMA 5.2. *There exists $T_0 \in (0,T)$ such that for all $T \in (0,T_0)$ and for all $v \in C_T(M)$, the matrix a is well defined (by (2.15)) with the estimate (independent of $v \in C_T(M)$)*

$$\begin{aligned} &\|a\|_{L^\infty(0,T;H^2(\Omega))} + \|a_t\|_{L^\infty(0,T;H^1(\Omega))} + \|a_t\|_{L^2(0,T;H^2(\Omega))} \\ &+ \|a_{tt}\|_{L^\infty(0,T;L^2(\Omega))} + \|a_{tt}\|_{L^2(0,T;H^1(\Omega))} \le C(M). \end{aligned} \tag{5.3}$$

LEMMA 5.3. *There exist $T_1 \in (0,T)$ and a constant C (independent of M) such that for all $T \in (0,T_1)$ and $v \in C_T(M)$, for all $\phi \in H^1(\Omega)$ and $t \in [0,T]$*

$$C\|\phi\|^2_{H^1(\Omega)} \le \int_\Omega \Big[|v|^2 + |D_\eta(v)|^2\Big] dx, \tag{5.4}$$

where

$$|D_\eta(v)|^2 := D_\eta(v)^i_j D_\eta(v)^i_j = (a^k_j v^i_{,k} + a^k_j v^i_{,k})(a^\ell_j v^i_{,\ell} + a^\ell_i v^j_{,\ell}).$$

In the remainder of the paper, we will assume that

$$0 < T < \min\{T_0, T_1, \bar{T}\}$$

for some fixed $\bar{T}$ where the forcing F is defined on the time interval $[0,\bar{T}]$.

6. Preliminary results.

6.1. Pressure as a Lagrange multiplier. In the following discussion, we use $H^{1;2}(\Omega;\Gamma)$ to denote the space $H^1(\Omega) \cap H^2(\Gamma)$ with norm

$$\|u\|^2_{H^{1;2}(\Omega;\Gamma)} = \|u\|^2_{H^1(\Omega)} + \|u\|^2_{H^2(\Gamma)}$$

and $\bar{\mathcal{V}}_{\bar{v}}$ ($\bar{\mathcal{V}}_{\bar{v}}(T)$) to denote the space

$$\Big\{ v \in \mathcal{V}_{\bar{v}} \;\Big|\; v \in H^2(\Gamma) \Big\} \Big(\Big\{ v \in \mathcal{V}_{\bar{v}}(T) \;\Big|\; v \in L^2(0,T;H^2(\Gamma)) \Big\} \Big).$$

LEMMA 6.1. *For all $p \in L^2(\Omega)$, $t \in [0,T]$, there exist a constant $C > 0$ and $\phi \in H^{1;2}(\Omega;\Gamma)$ such that $a^j_i(t)\phi^i_{,j} = p$ and*

$$\|\phi\|_{H^{1;2}(\Omega;\Gamma)} \le C\|p\|_{L^2(\Omega)}. \tag{6.1}$$

Proof. We solve the following problem on the time-dependent domain $\Omega(t)$:

$$\operatorname{div}(\phi \circ \eta(t)^{-1}) = p \circ \eta(t)^{-1} \quad \text{in } \eta(t,\Omega) := \Omega(t).$$

The solution to this problem can be written as the sum of the solutions to the following two problems:

$$\operatorname{div}(\phi \circ \eta(t)^{-1}) = p \circ \eta(t)^{-1} - \bar{p}(t) \qquad \text{in} \quad \eta(t,\Omega), \tag{6.2}$$

$$\operatorname{div}(\phi \circ \eta(t)^{-1}) = \bar{p}(t) \qquad \text{in} \quad \eta(t,\Omega), \tag{6.3}$$

where $\bar{p}(t) = \frac{1}{|\Omega|}\int_\Omega p(t,x)dx$. The existence of the solution to problem (6.2) with zero boundary condition is standard (see, for example, [12, Chapter 3]), and the solution to problem (6.3) can be chosen as a linear function (linear in x), for example, $\bar{p}(t)x_1$. The estimate (6.1) follows from the estimates of the solutions to (6.2). ☐

Define the linear functional on $H^{1;2}(\Omega;\Gamma)$ by $(p, a_i^j(t)\varphi^i_{,j})_{L^2(\Omega)}$, where $\varphi \in H^{1;2}(\Omega;\Gamma)$. By the Riesz representation theorem, there is a bounded linear operator $Q(t) : L^2(\Omega) \to H^{1;2}(\Omega;\Gamma)$ such that for all $\varphi \in H^{1;2}(\Omega;\Gamma)$,

$$(p, a_i^j(t)\varphi^i_{,j})_{L^2(\Omega)} = (Q(t)p, \varphi)_{H^{1;2}(\Omega;\Gamma)} := (Q(t)p,\varphi)_{H^1(\Omega)} + (Q(t)p,\varphi)_{H^2(\Gamma)}.$$

Letting $\varphi = Q(t)p$ shows that

$$\|Q(t)p\|_{H^{1;2}(\Omega;\Gamma)} \le C\|p\|_{L^2(\Omega)}$$

for some constant $C > 0$. By Lemma 6.1,

$$\|p\|^2_{L^2(\Omega)} \le \|Q(t)p\|_{H^{1;2}(\Omega;\Gamma)}\|\varphi\|_{H^{1;2}(\Omega;\Gamma)} \le C\|Q(t)p\|_{H^{1;2}(\Omega;\Gamma)}\|p\|_{L^2(\Omega)},$$

which shows that $R(Q(t))$ is closed in $H^{1;2}(\Omega;\Gamma)$. Since $\bar{\mathcal{V}}_v(t) \subset R(Q(t))^\perp$ and $R(Q(t))^\perp \subset \bar{\mathcal{V}}_v(t)$, it follows that

$$H^{1;2}(\Omega;\Gamma)(t) = R(Q(t)) \oplus_{H^{1;2}(\Omega;\Gamma)} \bar{\mathcal{V}}_v(t). \tag{6.4}$$

We can now introduce our Lagrange multiplier.

LEMMA 6.2. *Let $\mathcal{L}(t) \in H^{1;2}(\Omega;\Gamma)'$ be such that $\mathcal{L}(t)\varphi = 0$ for any $\varphi \in \bar{\mathcal{V}}_v(t)$. Then there exists a unique $q(t) \in L^2(\Omega)$, which is termed the pressure function, satisfying*

$$\forall\ \varphi \in H^{1;2}(\Omega;\Gamma), \qquad \mathcal{L}(t)(\varphi) = (q(t), a_i^j(t)\varphi^i_{,j})_{L^2(\Omega)}.$$

Moreover, there is a $C > 0$ (which does not depend on $t \in [0,T]$ and ϵ_1 and on the choice of $v \in C_T(M)$) such that

$$\|q(t)\|_{L^2(\Omega)} \le C\|\mathcal{L}(t)\|_{H^{1;2}(\Omega;\Gamma)'}.$$

Proof. By the decomposition (6.4), for given $\tilde{a}$, let $\varphi = v_1 + v_2$, where $v_1 \in \mathcal{V}_v(t)$ and $v_2 \in R(Q(t)$. It follows that

$$\mathcal{L}(t)(\varphi) = \mathcal{L}(t)(v_2) = (\psi(t), v_2)_{H^{1;2}(\Omega;\Gamma)} = (\psi(t),\varphi)_{H^{1;2}(\Omega;\Gamma)}$$

for a unique $\psi(t) \in R(Q(t))$.

From the definition of $Q(t)$ we then get the existence of a unique $q(t) \in L^2(\Omega)$ such that

$$\forall\ \varphi \in H^{1;2}(\Omega;\Gamma), \qquad \mathcal{L}(t)(\varphi) = (q(t), a_i^j(t)\varphi^i_{,j})_{L^2(\Omega)}.$$

The estimate stated in the lemma is then a simple consequence of (6.1). ☐

6.2. Estimates for a and h. We make use of near-identity transformations. The following lemmas can be found in [7].

LEMMA 6.3. *There exist $K > 0$ and $T_0 > 0$ such that if $0 < t \leq T_0$, then, for any $(\tilde{v}, \tilde{h}) \in C_{T_0}(M)$,*

$$\|\tilde{a}^T - Id\|_{L^\infty(0,T;\mathcal{C}^0(\overline{\Omega}_0))} \leq K\sqrt{t}, \tag{6.5a}$$

$$\|\tilde{a} - Id\|_{L^\infty(0,T;H^2(\Omega))} \leq K\sqrt{t}, \tag{6.5b}$$

$$\|\tilde{a}_t - \tilde{a}_t(0)\|_{L^\infty(0,T;H^1(\Omega))} \leq C(M)t, \tag{6.5c}$$

$$\|\tilde{a}_t\|_{L^\infty(0,T;H^1(\Omega))} \leq K. \tag{6.5d}$$

We also need the following lemma.

LEMMA 6.4. *For any $(\tilde{v}, \tilde{h}) \in C_{T_0}(M)$,*

$$\|\tilde{h}\|_{H^{3.5}(\Gamma)} \leq CMt^{1/4} \tag{6.6}$$

for all $0 < t \leq T_0$.

Proof. For $(\tilde{v}, \tilde{h}) \in C_T(M)$, $\|\tilde{h}\|^2_{H^4(\Gamma)} + \|\tilde{h}_t\|^2_{H^2(\Gamma)} \leq M$. By $\tilde{h}(0) = 0$,

$$\|\tilde{h}(t)\|_{H^2(\Gamma)} \leq \int_0^t \|\tilde{h}_t\|_{H^2(\Gamma)} ds \leq \sqrt{M}t.$$

Finally, the interpolation inequality

$$\|\nabla_0^2 f(t)\|_{H^{1.5}(\Gamma)} \leq C\|\nabla_0^4 f\|^{3/4}_{L^2(\Gamma)} \|\nabla_0^2 f\|^{1/4}_{L^2(\Gamma)} \tag{6.7}$$

implies

$$\|\tilde{h}\|_{H^{3.5}(\Gamma)} \leq C\|\tilde{h}\|^{3/4}_{H^4(\Gamma)} \|\tilde{h}\|^{1/4}_{H^2(\Gamma)} \leq CMt^{1/4}. \qquad \square$$

COROLLARY 6.5. *$\|L_1(t)\|_{H^{1.5}(\Gamma)}$ and $\|L_2(t)\|_{H^{1.5}(\Gamma)}$ converge to zero as $t \to 0$, uniformly in $(v, h) \in C_{T_0}(M)$. Furthermore, for $t \leq 1$,*

$$\|L_1(t)\|_{H^{1.5}(\Gamma)} + \|L_2(t)\|_{H^{1.5}(\Gamma)} \leq C(M)t^{1/4}.$$

By the fact that $\|\tilde{h}_t\|^2_{H^2(\Gamma)} \leq M$ and $\|\tilde{h}_{tt}\|^2_{L^2(0,T;H^{0.5}(\Gamma))} \leq M$ if $(\tilde{v}, \tilde{h}) \in C_T(M)$, similar computations lead to the following lemma.

LEMMA 6.6. *For all $(\tilde{v}, \tilde{h}) \in C_T(M)$,*

$$\|\tilde{h}_t(t)\|_{H^{1.5}(\Gamma)} \leq CMt^{1/8} \tag{6.8}$$

for all $0 < t \leq T$.

7. The linearized problem. Suppose that $(\tilde{v}, \tilde{h}) \in C_T(M)$ is given. Let $\tilde{\eta}(t) = \text{Id} + \int_0^t \tilde{v}(s)ds$ and $\tilde{a} = (\nabla\tilde{\eta})^{-1}$. We are concerned with the following time-dependent linear problem, whose fixed point $v = \tilde{v}$ provides a solution to (2.16):

$$v^i_t - \nu[\tilde{a}^k_\ell D_{\tilde{\eta}}(v)^i_\ell]_{,k} = -(\tilde{a}^k_i q)_{,k} + F^i \quad \text{in } (0,T)\times\Omega, \tag{7.1a}$$

$$\tilde{a}^j_i v^i_{,j} = 0 \quad \text{in } (0,T)\times\Omega, \tag{7.1b}$$

$$\begin{aligned}[\nu D_{\tilde{\eta}}(v)^j_i - q\delta^j_i]\tilde{a}^\ell_j N_\ell = {} & \sigma\tilde{\Theta}\Big[\mathcal{L}_{\tilde{h}}(h)(-\nabla_0\tilde{h}, 1)\Big]\circ\tilde{\eta}^\tau \\ & + \sigma\tilde{\Theta}\Big[[\mathcal{M}(\tilde{h})(-\nabla_0\tilde{h}, 1)]\circ\tilde{\eta}^\tau\Big]\end{aligned} \quad \text{on } (0,T)\times\Gamma, \tag{7.1c}$$

$$h_t\circ\tilde{\eta}^\tau = [\tilde{h}_{,\alpha}\circ\tilde{\eta}^\tau]v_\alpha - v_z \quad \text{on } (0,T)\times\Gamma, \tag{7.1d}$$

$$v = u_0 \quad \text{on } \{t=0\}\times\Omega, \tag{7.1e}$$

$$h = 0 \quad \text{on } \{t=0\}\times\Gamma, \tag{7.1f}$$

where $D_{\tilde\eta}(v)^j_i = \tilde a^k_i v^j_{,k} + \tilde a^k_j v^i_{,k}$, $\tilde\Theta = \det(\nabla_0 \tilde\eta^\tau)$, and

$$\mathcal{L}_{\tilde h}(h) = \frac{1}{\sqrt{\det(g_0)}} \Big[\sqrt{\det(g_0)} \tilde A^{\alpha\beta\gamma\delta} h_{,\alpha\beta} \Big]_{,\gamma\delta}$$

with

$$\begin{aligned} \tilde A^{\alpha\beta\gamma\delta} &= J^{-3}_{\tilde h} \sqrt{\det(G_{\tilde h})} \Big[G^{\alpha\gamma}_{\tilde h} - (-1)^{\kappa+\sigma} \det(G_{\tilde h})^{-1} (1-\delta_{\alpha\kappa})(1-\delta_{\gamma\sigma}) \tilde h_{,\kappa} \tilde h_{,\sigma} \Big] \\ &\quad \times (G^{\beta\delta}_{\tilde h} - J^{-2}_{\tilde h} G^{\beta\mu}_{\tilde h} G^{\delta\nu}_{\tilde h} \tilde h_{,\mu} \tilde h_{,\nu}) \end{aligned}$$

and

$$\mathcal{M}(\tilde h) = \sqrt{\det(G_{\tilde h}) \circ \tilde\eta^\tau} \Big[L_1^{\alpha\beta\gamma}(y, \tilde h, D\tilde h, D^2 \tilde h) \tilde h_{,\alpha\beta\gamma} + L_2(y, \tilde h, D\tilde h, D^2\tilde h) \Big].$$

Here the thickness ϵ_1 is assumed to be 1.

We will also use $L_{\tilde h}(h)$ to denote $\mathcal{L}_{\tilde h}(h) + \mathcal{M}(\tilde h)$.

Remark 10. *$\mathcal{L}_{\tilde h}$ is a coercive fourth order operator for small $\tilde h \le \delta$. Actually, it is easy to see that $\mathcal{L}_{\tilde h}$ is coercive at time $t = 0$, and the coercivity of $\mathcal{L}_{\tilde h}$ for $t > 0$ (but sufficiently small) follows from the continuity of $\tilde h$ in time into the space $H^2(\Gamma)$. Moreover, by Lemma* 6.4*, we have the following corollary.*

Corollary 7.1. *There exist $\nu_1 > 0$ and $0 < T \le T_0$ such that for all $0 < t \le T$,*

$$\nu_1 \|\nabla_0^2 f(t)\|^2_{L^2(\Gamma)} \le \int_\Gamma \tilde A^{\alpha\beta\gamma\delta} f_{,\alpha\beta}(t) f_{,\gamma\delta}(t) dS$$

for all $0 < t \le T$. Later we will denote the right-hand side quantity of this inequality by $E_{\bar h}(f)$, where the subscript $\bar h$ indicates that $\bar A$ is a function of $\bar h$.

Remark 11. *Given $(\tilde v, \tilde h) \in \mathcal{V}^3(T) \times \mathcal{H}(T)$, for the corresponding $\tilde\eta^\tau$, we have*

$$\|\tilde\eta^\tau\|^2_{L^\infty(0,T;H^{2.5}(\Omega))} + \|\tilde\eta^\tau_t\|^2_{L^2(0,T;H^{2.5}(\Gamma))} \le C(M),$$

where (2.13) *and* (2.12) *are used to obtain this estimate.*

The solution of (7.1) is found as a weak limit of a sequence of regularized problems.

Definition 7.2 (mollifiers on Γ). *For $\epsilon_1 > 0$, let*

$$K^p_{\epsilon_1} := (1 - \epsilon_1 \Delta_0)^{-\frac{p}{2}} : H^s(\Gamma) \to H^{s+p}(\Gamma)$$

denote the usual self-adjoint Friedrich mollifier on the compact manifold Γ, where Δ_0 is the surface Laplacian defined on Γ.

By the Sobolev extension theorem, there exist bounded extension operators

$$E_s : H^s(\Omega) \to H^s(\mathbb{R}^n), \qquad s \ge 1.$$

For fixed (but small) ϵ_1 and $\epsilon_{11} > 0$, let ρ_{ϵ_1} be a (positive) smooth mollifier on $\mathbb{R}^n$. Set $\bar v = \rho_{\epsilon_1} * E_1(\tilde v)$, $\tilde F = \rho_{\epsilon_1} * E_2(F)$, $\tilde u_0 = \rho_{\epsilon_1} * E_3(u_0)$, where $*$ denotes the convolution in space, and $\bar h = K^m_{\epsilon_1}(\tilde h)$ for large enough m. Define $\bar\eta$ and $\bar a$ in the same fashion as $\tilde\eta$ and $\tilde a$. Note that $\bar v \to \tilde v \in V(T)$, $\tilde F \to F$ in $\mathcal{V}^2(T)$, $\tilde{u_0} \to u_0$ in $H^{2.5}(\Omega)$, and $\bar h \to \tilde h$ in $\mathcal{H}(T)$ as $\epsilon_1 \to 0$.

The regularized problem takes the form

$$v_t^i - \nu[\bar{a}_\ell^k D_{\bar{\eta}}(v)_\ell^i]_{,k} = -(\bar{a}_i^k q)_{,k} + \tilde{F}^i \quad \text{in } (0,T)\times\Omega, \tag{7.2a}$$

$$\bar{a}_i^j v^i_{,j} = 0 \quad \text{in } (0,T)\times\Omega, \tag{7.2b}$$

$$\begin{aligned}[\nu D_{\bar{\eta}}(v)_i^j - q\delta_i^j]\bar{a}_j^\ell N_\ell &= \sigma\mathcal{L}_{\bar{h}}^{\epsilon_2}(h^{\epsilon_2})(-\nabla_0\bar{h}\circ\bar{\eta}^\tau, 1)\\ &\quad + \sigma\mathcal{M}_{\bar{h}}^{\epsilon_2}(-\nabla_0\bar{h}\circ\bar{\eta}^\tau, 1) + \kappa\Delta_0^2 v \quad \text{on } (0,T)\times\Gamma,\end{aligned} \tag{7.2c}$$

$$h_t\circ\bar{\eta}^\tau = [(\bar{h}_{,\alpha})\circ\bar{\eta}^\tau]v_\alpha - v_z \quad \text{on } (0,T)\times\Gamma, \tag{7.2d}$$

$$v = \tilde{u}_0 \quad \text{on } \{t=0\}\times\Omega, \tag{7.2e}$$

$$h = 0 \quad \text{on } \{t=0\}\times\Gamma, \tag{7.2f}$$

where

$$\begin{aligned}\bar{\mathcal{L}}_{\bar{h}}^{\epsilon_2}(f) &= \frac{\bar{\Theta}}{\sqrt{\det(g_0)}}\Big[\Big(\sqrt{\det(g_0)}\bar{A}^{\alpha\beta\gamma\delta}f_{,\alpha\beta}\Big)_{,\gamma\delta}\Big]^{\epsilon_2}\circ\bar{\eta}^\tau,\\ \bar{\mathcal{M}}_{\bar{h}}^{\epsilon_2} &= \bar{\Theta}\Big[\Big(L_1^{\alpha\beta\gamma}(\cdot,\bar{h},D\bar{h},D^2\bar{h})\bar{h}_{,\alpha\beta\gamma} + L_2(\cdot,\bar{h},D\bar{h})\Big)^{\epsilon_2}\Big]^{\epsilon_2}\circ\bar{\eta}^\tau(y,t).\end{aligned}$$

Note that $\bar{\mathcal{L}}_{\bar{h}}^{\epsilon_2}(f) + \bar{\mathcal{M}}_{\bar{h}}^{\epsilon_2} = \bar{\Theta}\big[L_{\bar{h}}(f)\big]^{\epsilon_2}\circ\bar{\eta}^\tau$.

7.1. Weak solutions.

DEFINITION 7.3. *A vector $v\in\bar{\mathcal{V}}_{\bar{v}}(T)$ with $v_t\in\bar{\mathcal{V}}_{\bar{v}}(T)'$ for almost all (a.a.) $t\in(0,T)$ is a weak solution of* (7.2)*, provided that*

$$\begin{aligned}&\text{(i) } \langle v_t,\varphi\rangle + \frac{\nu}{2}\int_\Omega D_{\bar{\eta}}v : D_{\bar{\eta}}\varphi dx + \sigma\int_\Gamma \bar{A}^{\alpha\beta\gamma\delta}h^{\epsilon_2}_{,\alpha\beta}\Big[-\bar{h}_{,\sigma}(\varphi^\sigma\circ\bar{\eta}^{-\tau})\\ &\qquad + (\varphi^z\circ\bar{\eta}^{-\tau})\Big]^{\epsilon_2}_{,\gamma\delta}dS + \kappa\int_\Gamma \Delta_0 v\cdot\Delta_0\varphi dS = \langle\tilde{F},\varphi\rangle - \sigma\langle\mathcal{M}_{\bar{h}}^{\epsilon_2},\varphi\rangle_\Gamma,\end{aligned} \tag{7.3a}$$

$$\text{(ii) } v(0,\cdot) = \tilde{u}_0 \tag{7.3b}$$

for a.a. $t\in[0,T]$, where $\langle\cdot,\cdot\rangle$ denotes the duality product between $\bar{\mathcal{V}}_v(t)$ and its dual $\bar{\mathcal{V}}_v(t)'$, and h is given by the evolution equation (7.2d) *and the initial condition* (7.2f)*:*

$$h(y,t) = \int_0^t \Big[-\bar{h}_{,\alpha}(y,s)v^\alpha(\bar{\eta}^{-\tau}(y,s),0,s) + v^z(\bar{\eta}^{-\tau}(y,s),0,s)\Big]ds. \tag{7.4}$$

7.2. Penalized problems. Letting $\theta>0$ denote the penalized parameter, we define w_θ (also with ϵ_1 and ϵ_{11} dependence in mind) to be the "unique" solution of the problem (whose existence can be obtained via a modified Galerkin method which will be presented in the following sections):

$$\begin{aligned}&\text{(i)}\quad \langle w_{\theta t},\varphi\rangle + \frac{\nu}{2}\int_\Omega D_{\bar{\eta}}w_\theta : D_{\bar{\eta}}\varphi dx + \sigma\int_\Gamma \bar{A}^{\alpha\beta\gamma\delta}h^{\epsilon_2}_{,\alpha\beta}\Big[-\bar{h}_{,\sigma}(\varphi^\sigma\circ\bar{\eta}^{-\tau})\\ &\qquad + (\varphi^z\circ\bar{\eta}^{-\tau})\Big]^{\epsilon_2}_{,\gamma\delta}dS + \kappa\int_\Gamma\Delta_0 v\cdot\Delta_0\varphi dS + \left(\frac{1}{\theta}\bar{a}_i^j v^i_{,j}, \bar{a}_k^\ell\varphi^k_{,\ell}\right)_{L^2(\Omega)}\\ &\quad = \langle\tilde{F},\varphi\rangle - \sigma\langle\bar{\mathcal{M}}_{\bar{h}}^{\epsilon_2}(-\nabla_0\bar{h}\circ\bar{\eta}^\tau,1),\varphi\rangle_\Gamma,\end{aligned} \tag{7.5a}$$

$$\text{(ii) } v(0,\cdot) = \tilde{u}_0, \tag{7.5b}$$

where $\langle\cdot,\cdot\rangle$ denotes the pairing between $H^1(\Omega)$ and its dual, and h in (7.5a) satisfies (7.4) with v replaced by w_θ.

7.3. Weak solutions for the penalized problem. The goal of this section is to establish the existence of v to the problem (7.2) (or the weak formulation (7.3)), as well as the energy inequality satisfied by v and v_t. Before proceeding, we introduce variables $\tilde{q}_0$ and $\tilde{w}_1$ as follows: let $\tilde{q}_0$ be the solution of the Laplace equation

$$\Delta\tilde{q}_0 = \nabla\tilde{u}_0 : (\nabla\tilde{u}_0)^t - \operatorname{div}\tilde{F}(0) \qquad \text{in } \Omega, \tag{7.6a}$$

$$\tilde{q}_0 = \nu(\operatorname{Def}\tilde{u}_0)^j_i N_i N_j - \sigma\mathcal{M}_0^{\epsilon_2}(0) + \kappa\Delta_0^2\tilde{u}_0\cdot N \qquad \text{on } \Gamma \tag{7.6b}$$

and $\tilde{w}_1$ be defined by

$$\tilde{w}_1 = \nu\Delta\tilde{u}_0 - \nabla\tilde{q}_0 + \tilde{F}(0). \tag{7.7}$$

By elliptic regularity,

$$\begin{aligned}\|\tilde{q}_0\|^2_{H^1(\Omega)} &\leq C\Big[\|\tilde{u}_0\|^2_{H^2(\Omega)} + \|\tilde{F}(0)\|^2_{L^2(\Omega)} + \|\mathcal{M}_0^{\epsilon_2}(0)\|^2_{H^{0.5}(\Gamma)} + \|\Delta_0^2\tilde{u}_0\|^2_{H^{0.5}(\Gamma)}\Big]\\ &\leq C(M)\Big[\|\tilde{u}_0\|^2_{H^2(\Omega)} + \|\tilde{u}_0\|^2_{H^{4.5}(\Gamma)} + \|\tilde{F}(0)\|^2_{L^2(\Omega)} + 1\Big],\end{aligned}$$

and hence

$$\|\tilde{w}_1\|^2_{L^2(\Omega)} \leq C(M)\Big[\|\tilde{u}_0\|^2_{H^2(\Omega)} + \|\tilde{u}_0\|^2_{H^{4.5}(\Gamma)} + \|\tilde{F}(0)\|^2_{L^2(\Omega)} + 1\Big].$$

REMARK 12. *By* (6.6)*, the constant $C(M)$ in the estimates above can also be refined as a constant independent of M if T is chosen small enough.*

By introducing a (smooth) basis $(e_\ell)_{\ell=1}^\infty$ of $H^{1;2}(\Omega;\Gamma)$, taking the approximation at rank $m \geq 2$ under the form $w_\ell(t,x) = \sum_{k=1}^{\ell} d_k(t)e_k(x)$ with

$$h_\ell(y,t) = \int_0^t \Big[-\bar{h}_{,\alpha}(y,s)w_\ell^\alpha(\bar{\eta}^{-\tau}(y,s),0,s) + w_\ell^z(\bar{\eta}^{-\tau}(y,s),0,s)\Big]ds, \tag{7.8}$$

and satisfying on $[0,T]$,

(7.9a)

$$\begin{aligned}&\text{(i)}\quad (w_{\ell tt},\varphi)_{L^2(\Omega)} + \nu(\bar{a}^j_i w_{\ell t,j}, \bar{a}^k_i\varphi_{,k})_{L^2(\Omega)} + \nu((\bar{a}^j_i\bar{a}^k_i)_t w_\ell,\varphi_{,k})_{L^2(\Omega)}\\ &+ \nu\int_\Omega \Big[\bar{a}^j_r\bar{a}^k_i w^i_{\ell t,j} + (\bar{a}^j_r\bar{a}^k_i)_t w^i_{\ell,j}\Big]\varphi^r_{,k}dx + \kappa\int_\Gamma \Delta_0 w_{\ell t}\cdot\Delta_0\varphi dS - ((\bar{a}^j_i q_\ell)_t,\varphi^i_{,j})_{L^2(\Omega)}\\ &+ \sigma\int_\Gamma \bar{A}^{\alpha\beta\gamma\delta}[-\bar{h}_{,\sigma}(w_\ell^\sigma\circ\bar{\eta}^{-\tau}) + w_\ell^z\circ\bar{\eta}^{-\tau}]^{\epsilon_2}_{,\alpha\beta}[-\bar{h}_{,\sigma}(\varphi^\sigma\circ\bar{\eta}^{-\tau}) + \varphi^z\circ\bar{\eta}^{-\tau}]^{\epsilon_2}_{,\gamma\delta}dS\\ &+ \sigma\int_\Gamma (\bar{A}^{\alpha\beta\gamma\delta})_t h^{\epsilon_2}_{\ell,\alpha\beta}[-\bar{h}_{,\sigma}(\varphi^\sigma\circ\bar{\eta}^{-\tau}) + \varphi^z\circ\bar{\eta}^{-\tau}]^{\epsilon_2}_{,\gamma\delta}dS\\ &+ \sigma\int_\Gamma \bar{A}^{\alpha\beta\gamma\delta}h^{\epsilon_2}_{\ell,\alpha\beta}[-\bar{h}_{t,\sigma}(\varphi^\sigma\circ\bar{\eta}^{-\tau}) + \bar{h}_{,\sigma}\bar{v}^\kappa(\varphi^\sigma_{,\kappa}\circ\bar{\eta}^{-\tau}) + \bar{v}^\kappa(\varphi^z_{,\kappa}\circ\bar{\eta}^{-\tau})]^{\epsilon_2}_{,\gamma\delta}dS\\ &= \langle\tilde{F}_t,\varphi\rangle - \sigma\int_\Gamma \Big[L_1^{\alpha\beta\gamma}\bar{h}_{,\alpha\beta\gamma} + L_2\Big]^{\epsilon_2}_t\Big[\bar{h}_{,\sigma}(\varphi^\sigma\circ\bar{\eta}^{-\tau}) - \varphi^z\circ\bar{\eta}^{-\tau}\Big]^{\epsilon_2}dS\\ &- \sigma\int_\Gamma \Big[L_1^{\alpha\beta\gamma}\bar{h}_{,\alpha\beta\gamma} + L_2\Big]^{\epsilon_2}\Big[\bar{h}_{t,\sigma}(\varphi^\sigma\circ\bar{\eta}^{-\tau}) - \bar{h}_{,\sigma}\bar{v}^\kappa(\varphi^\sigma_{,\kappa}\circ\bar{\eta}^{-\tau}) - \bar{v}^\kappa(\varphi^z_{,\kappa}\circ\bar{\eta}^{-\tau})\Big]^{\epsilon_2}dS\\ &\forall\ \varphi\in\operatorname{span}(e_1,\dots,e_\ell),\end{aligned}$$

(7.9b)

$$\text{(ii)}\quad w_{\ell t}(0) = (w_1)_\ell,\ w_\ell(0) = (u_0)_\ell \quad \text{in } \Omega,$$

where $q_\ell = \tilde{q}_0 - \frac{1}{\theta}\bar{a}_i^j w^i_{\ell,j}$, and $(\tilde{u}_0)_\ell$ denotes the respective $H^{1;2}(\Omega;\Gamma)$ projections of u_0 on $\text{span}(e_1, e_2, \dots, e_\ell)$.

REMARK 13. *The existence of w_k follows from the solution of*

$$d_k''(t) + d_\ell'(t)A_{k\ell}(t) + d_\ell(t)B_{k\ell}(t) + \int_0^t d_\ell(s)C_{k\ell}(s,t)ds = F(t)$$

for functions A, B, C, and F; however, the existence of the solution d_k does not immediately follow from the fundamental theorem of ODE due to the presence of the time integral. A straightforward fixed-point argument can be implemented, whose details we leave to the interested reader.

The use of the test function $\varphi = w_{\ell t}$ in this system of ODE gives us, in turn, the energy law

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|w_{\ell t}\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|D_{\bar\eta}(w_{\ell t})\|^2_{L^2(\Omega)} + \frac{\sigma}{2}\frac{d}{dt}E_{\bar h}(h^{\epsilon_2}_{\ell t,\alpha\beta}) + \theta\|q_{\ell t}\|^2_{L^2(\Omega)}\\
&+ \nu((\bar{a}_i^j\bar{a}_i^k)_t w_{\ell,j}, w_{\ell t,k})_{L^2(\Omega)} + \nu\int_\Omega(\bar{a}_r^j\bar{a}_i^k)_t w^i_{\ell,j}w^r_{\ell t,k}dx + \kappa\|\Delta_0 w_{\ell t}\|^2_{L^2(\Gamma)}\\
&+ (q_{\ell t}, \bar{a}^j_{it}w^i_{\ell,j})_{L^2(\Omega)} - (q_\ell, \bar{a}^j_{it}w^i_{\ell t,j})_{L^2(\Omega)} - \frac{\sigma}{2}\int_\Gamma(\bar{A}^{\alpha\beta\gamma\delta})_t h^{\epsilon_2}_{\ell t,\alpha\beta}h^{\epsilon_2}_{\ell t,\gamma\delta}dS\\
&- \sigma\int_\Gamma(\bar{A}^{\alpha\beta\gamma\delta})_t h^{\epsilon_2}_{\ell,\alpha\beta}\Big[h_{\ell tt} + \bar{h}_{t,\sigma}(w^\sigma_{\ell t}\circ\bar\eta^{-\tau})\Big]^{\epsilon_2}_{,\gamma\delta}dS + \sigma\int_\Gamma\bar{A}^{\alpha\beta\gamma\delta}h^{\epsilon_2}_{\ell,\alpha\beta}\\
&\times\Big[-\bar{h}_{t,\sigma}(w^\sigma_{\ell t}\circ\bar\eta^{-\tau}) + \bar{h}_{,\sigma}\bar{v}^\kappa(w^\sigma_{\ell t,\kappa}\circ\bar\eta^{-\tau}) + \bar{v}^\kappa(w^z_{\ell t,\kappa}\circ\bar\eta^{-\tau})\Big]^{\epsilon_2}_{,\gamma\delta}dS\\
=\;&\langle\tilde{F}_t, w_{\ell t}\rangle - \sigma\int_\Gamma\Big[(L_1^{\alpha\beta\gamma}\bar{h}_{,\alpha\beta\gamma} + L_2)(-\nabla_0\bar{h},1)\Big]_t\cdot(w_{\ell t}\circ\bar\eta^{-\tau})dS\\
&- \sigma\int_\Gamma(L_1^{\alpha\beta\gamma}\bar{h}_{,\alpha\beta\gamma} + L_2)\bar{v}^\kappa\Big[-\bar{h}_{,\sigma}(w^\sigma_{\ell t,\kappa}\circ\bar\eta^{-\tau}) + (w^z_{\ell t,\kappa}\circ\bar\eta^{-\tau})\Big]dS.
\end{aligned}\tag{7.10}$$

For the tenth term (the integral with $\frac{\sigma}{2}$ as its coefficient), we have

$$\left|\int_\Gamma(\bar{A}^{\alpha\beta\gamma\delta})_t h^{\epsilon_2}_{\ell t,\alpha\beta}h^{\epsilon_2}_{\ell t,\gamma\delta}dS\right| \le C(M)\|\bar{h}_t\|_{H^{2.5}(\Gamma)}\|\nabla_0^2 h_{\ell t}\|^2_{L^2(\Gamma)}.$$

By ϵ_2-regularization and the identity

$$\begin{aligned}
\int_\Gamma(\bar{A}^{\alpha\beta\gamma\delta})_t h^{\epsilon_2}_{\ell,\alpha\beta}h^{\epsilon_2}_{\ell tt,\gamma\delta}dS &= \int_\Gamma\frac{1}{\sqrt{\det(g_0)}}\Big[\sqrt{\det(g_0)}(\bar{A}^{\alpha\beta\gamma\delta})_t\Big]_{,\gamma\delta}h^{\epsilon_2}_{\ell,\alpha\beta}h^{\epsilon_2}_{\ell tt}dS\\
&+ \int_\Gamma\frac{2}{\sqrt{\det(g_0)}}\Big[\sqrt{\det(g_0)}(\bar{A}^{\alpha\beta\gamma\delta})_t\Big]_{,\gamma}h^{\epsilon_2}_{\ell,\alpha\beta\delta}h^{\epsilon_2}_{\ell tt}dS\\
&+ \int_\Gamma(\bar{A}^{\alpha\beta\gamma\delta})_t h^{\epsilon_2}_{\ell,\alpha\beta\gamma\delta}h^{\epsilon_2}_{\ell tt}dS,
\end{aligned}$$

we find that

$$\begin{aligned}
&\left|\int_\Gamma(\bar{A}^{\alpha\beta\gamma\delta})_t h^{\epsilon_2}_{\ell,\alpha\beta}h^{\epsilon_2}_{\ell tt,\gamma\delta}dS\right|\\
&\le C(\epsilon_2)\Big[1 + \|\bar{h}_t\|_{H^{2.5}(\Gamma)}\Big]\|\nabla_0^2 h_\ell\|_{L^2(\Gamma)}\Big[\|w_\ell\|_{H^1(\Omega)} + \|w_{\ell t}\|_{H^1(\Omega)}\Big].
\end{aligned}$$

Similarly, the second part of the eleventh term and the last term of the left-hand side can be bounded by

$$C(\epsilon_2)\|\bar{h}_t\|_{H^{2.5}(\Gamma)}\|\nabla_0^2 h_\ell\|_{L^2(\Gamma)}\|w_{\ell t}\|_{H^1(\Omega)},$$

where we also use the ϵ_2-regularization to control $\nabla_0^3 w_{\ell t}$. It also follows that the last two terms on the right-hand side can be bounded by

$$C(M)\Big[1+\|\bar{h}_t\|_{H^{2.5}(\Gamma)}\Big]\|w_{\ell t}\|_{H^1(\Omega)}.$$

With positive θ, the fourth term of the left-hand side involving the square of $q_{\ell t}$ acts as a viscous energy term. Integrating (7.10) in time from 0 to t, we then get

(7.11)

$$\begin{aligned}
&\|w_{\ell t}\|^2_{L^2(\Omega)}+\|\nabla_0^2 h_{\ell t}\|^2_{L^2(\Gamma)}+\int_0^t\Big[\|\nabla w_{\ell t}\|^2_{L^2(\Omega)}+\kappa\|w_{\ell t}\|^2_{H^2(\Gamma)}+\theta\|q_{\ell t}\|^2_{L^2(\Omega)}\Big]ds\\
&\le C(M)\Big[\|w_{\ell t}(0)\|^2_{L^2(\Omega)}+\|w_\ell(0)\|^2_{H^1(\Omega)}+\|q_\ell(0)\|^2_{H^{0.5}(\Omega)}\Big]\\
&\quad+C(\epsilon_2)\int_0^t\Big[1+\|\bar{h}_t(s)\|^2_{H^{2.5}(\Gamma)}\Big]\|\nabla_0^2 h_{\ell t}(s)\|^2_{L^2(\Gamma)}ds\\
&\quad+C(\theta)\int_0^t\|\bar{v}(t')\|^2_{H^3(\Omega)}\int_0^{t'}\Big[\|\nabla w_{\ell t}(s)\|^2_{L^2(\Omega)}+\|q_{\ell t}(s)\|^2_{L^2(\Omega)}\Big]ds\,dt',
\end{aligned}$$

where $C(\epsilon_2), C(\theta)\to\infty$ as $\epsilon_2,\theta\to 0$, and we use

$$\|f(t)\|_X\le\|f(0)\|_X+\int_0^t\|f_t(s)\|_X ds\le\|f(0)\|_X+\sqrt{t}\int_0^t\|f_t(s)\|_X^2 ds$$

for $f=w_\ell$, $f=h_\ell$, and $f=g_\ell$ to obtain (7.11).

REMARK 14. *The θ-dependence follows from estimating the terms $(q_{\ell t},\bar{a}^j_{it}w^i_{\ell,j})_{L^2(\Omega)}$:*

$$\begin{aligned}
\Big|(q_{\ell t},\bar{a}^j_{it}w^i_{\ell,j})_{L^2(\Omega)}\Big|&\le\frac{\theta}{2}\|q_{\ell t}\|^2_{L^2(\Omega)}+\frac{1}{2\theta}\|\bar{a}^j_{it}\|^2_{L^\infty(\Omega)}\|w^i_{\ell,j}\|^2_{L^2(\Omega)}\\
&\le\frac{\theta}{2}\|q_{\ell t}\|^2_{L^2(\Omega)}+\frac{C(M)}{\theta}\Big[\|\nabla w_\ell(0)\|^2_{L^2(\Omega)}+t\int_0^t\|\nabla w_{\ell t}\|^2_{L^2(\Omega)}(s)ds\Big].
\end{aligned}$$

By the Gronwall inequality, for $0\le t\le T$,

$$\begin{aligned}
&\|w_{\ell t}(t)\|^2_{L^2(\Omega)}+\|\nabla_0^2 h_{\ell t}(t)\|^2_{L^2(\Gamma)}\\
&+\int_0^t\Big[\|\nabla w_{\ell t}\|^2_{L^2(\Omega)}+\kappa\|w_{\ell t}\|^2_{H^2(\Gamma)}+\theta\|q_{\ell t}\|^2_{L^2(\Omega)}\Big]ds\le C(\epsilon_2,\theta)N_0(u_0,F),
\end{aligned}\tag{7.12}$$

where

$$N_0(u_0,F):=\|u_0\|^2_{H^{2.5}(\Omega)}+\|u_0\|^2_{H^{4.5}(\Gamma)}+\|F_t\|^2_{L^2(0,T;H^1(\Omega)')}+\|F(0)\|^2_{H^{0.5}(\Omega)}+1.$$

We can then infer that w_ℓ is defined on $[0,T]$, and that there is a subsequence, still denoted with the subscript ℓ, satisfying

$$w_\ell\rightharpoonup w_\theta\quad\text{in } L^2(0,T;H^{1;2}(\Omega;\Gamma)),\tag{7.13a}$$

$$w_{\ell t}\rightharpoonup w_{\theta t}\quad\text{in } L^2(0,T;H^{1;2}(\Omega;\Gamma)),\tag{7.13b}$$

$$\nabla_0^2 h_\ell\rightharpoonup\nabla_0^2 h_\theta\quad\text{in } L^2(0,T;L^2(\Gamma)),\tag{7.13c}$$

$$\nabla_0^2 h_{\ell t}\rightharpoonup\nabla_0^2 h_{\theta t}\quad\text{in } L^2(0,T;L^2(\Gamma)),\tag{7.13d}$$

$$q_{\ell t}\rightharpoonup q_{\theta t}\quad\text{in } L^2(0,T;L^2(\Omega)),\tag{7.13e}$$

where

$$q_\theta = \tilde{q}_0 - \frac{1}{\theta}\bar{a}_i^j w^i_{\theta,j}.$$

From the standard procedure for weak solutions, we can now infer from these weak convergences and the definition of w_ℓ that $w_{\ell tt} \in L^2(0,T;H^1(\Omega)')$. In turn, $w_{\ell t} \in \mathcal{C}^0([0,T];H^1(\Omega)')$, $w_\ell \in \mathcal{C}^0([0,T];L^2(\Omega))$ with $w_\theta(0) = u_0$, $w_{\theta t}(0) = w_1$.

Moreover, (7.13) implies that w_θ satisfies

(7.14a)

$$\begin{aligned}
\text{(i)}\quad & \int_0^T \Big[(w_{\theta tt},\varphi)_{L^2(\Omega)} + \nu(\bar{a}_i^j w_{\theta t,j}, \bar{a}_i^k \varphi_{,k})_{L^2(\Omega)} + \nu((\bar{a}_i^j\bar{a}_i^k)_t w_\theta, \varphi_{,k})_{L^2(\Omega)}\Big]dt \\
& + \nu\int_0^T \left[\int_\Omega \bar{a}_r^j\bar{a}_i^k w^i_{\theta t,j}\varphi^r_{,k}dx + \nu\int_\Omega (\bar{a}_r^j\bar{a}_i^k)_t w^i_{\theta,j}\varphi^r_{,k}dx\right]dt + \sigma\int_0^T\int_\Gamma \bar{A}^{\alpha\beta\gamma\delta} \\
& \qquad \times [-\bar{h}_{,\sigma}(w^\sigma_\theta\circ\bar{\eta}^{-\tau}) + w^z_\theta\circ\bar{\eta}^{-\tau}]^{\epsilon_2}_{,\alpha\beta}[-\bar{h}_{,\sigma}(\varphi^\sigma\circ\bar{\eta}^{-\tau}) + \varphi^z\circ\bar{\eta}^{-\tau}]^{\epsilon_2}_{,\gamma\delta}dSdt \\
& + \sigma\int_0^T\int_\Gamma (\bar{A}^{\alpha\beta\gamma\delta})_t h^{\epsilon_2}_{\theta,\alpha\beta}[-\bar{h}_{,\sigma}(\varphi^\sigma\circ\bar{\eta}^{-\tau}) + \varphi^z\circ\bar{\eta}^{-\tau}]^{\epsilon_2}_{,\gamma\delta}dSdt \\
& + \sigma\int_0^T\int_\Gamma \bar{A}^{\alpha\beta\gamma\delta}h^{\epsilon_2}_{\theta,\alpha\beta}[-\bar{h}_{t,\sigma}(\varphi^\sigma\circ\bar{\eta}^{-\tau}) + \bar{h}_{,\sigma}\bar{v}^\kappa(\varphi^\sigma_{,\kappa}\circ\bar{\eta}^{-\tau}) + \bar{v}^\kappa(\varphi^z_{,\kappa}\circ\bar{\eta}^{-\tau})]^{\epsilon_2}_{,\gamma\delta}dSdt \\
& + \kappa\int_0^T\int_\Gamma \Delta_0 w_{\theta t}\cdot\Delta_0\varphi dSdt - \int_0^T ((\bar{a}_i^j q_\theta)_t, \varphi^i_{,j})_{L^2(\Omega)}dt \\
& = \int_0^T \Bigg\{\langle \tilde{F}_t,\varphi\rangle - \sigma\int_\Gamma \Big[L_1^{\alpha\beta\gamma}\bar{h}_{,\alpha\beta\gamma} + L_2\Big]^{\epsilon_2}_t \Big[\bar{h}_{,\sigma}(\varphi^\sigma\circ\bar{\eta}^{-\tau}) - \varphi^z\circ\bar{\eta}^{-\tau}\Big]^{\epsilon_2}dS \\
& \qquad - \sigma\int_\Gamma \Big[L_1^{\alpha\beta\gamma}\bar{h}_{,\alpha\beta\gamma} + L_2\Big]^{\epsilon_2}\Big[\bar{h}_{t,\sigma}(\varphi^\sigma\circ\bar{\eta}^{-\tau}) - \bar{h}_{,\sigma}\bar{v}^\kappa(\varphi^\sigma_{,\kappa}\circ\bar{\eta}^{-\tau}) \\
& \qquad\qquad - \bar{v}^\kappa(\varphi^z_{,\kappa}\circ\bar{\eta}^{-\tau})\Big]^{\epsilon_2}dS\Bigg\}dt,
\end{aligned}$$

(7.14b)

(ii) $w_{\theta t}(0) = \tilde{w}_1$, $w_\theta(0) = \tilde{u}_0$ in Ω

for all $\varphi \in L^2(0,T;H^{1;2}(\Omega;\Gamma))$. Choosing φ to be independent of time, we find that for all $t\in[0,T]$,

$$\begin{aligned}
& (w_{\theta t},\varphi)_{L^2(\Omega)} + \frac{\nu}{2}\int_\Omega D_{\bar{\eta}}(w_\theta) : D_{\bar{\eta}}(\varphi)dx + \kappa\int_\Gamma \Delta_0 w_\theta\cdot\Delta_0\varphi dS \\
& + \sigma\int_\Gamma \bar{A}^{\alpha\beta\gamma\delta}h^{\epsilon_2}_{\theta,\alpha\beta}[-\bar{h}_{,\sigma}(\varphi^\sigma\circ\bar{\eta}^{-\tau}) + \varphi^z\circ\bar{\eta}^{-\tau}]^{\epsilon_2}_{,\gamma\delta}dS - (\bar{a}_i^j q_\theta, \varphi^i_{,j})_{L^2(\Omega)} \\
= {} & \langle\tilde{F},\varphi\rangle + \sigma\int_\Gamma \Big[L_1^{\alpha\beta\gamma\delta}\bar{h}_{,\alpha\beta\gamma} + L_2\Big]^{\epsilon_2}\Big[-\bar{h}_{,\sigma}\varphi^\sigma\circ\bar{\eta}^{-\tau} + \varphi^z\circ\bar{\eta}^{-\tau}\Big]^{\epsilon_2}dS + c(\varphi)
\end{aligned}$$

for all $\varphi\in H^{1;2}(\Omega;\Gamma)$, where $c(\varphi)\in\mathbb{R}$ is given by

$$\begin{aligned}
c(\varphi) = {} & (\tilde{w}_1,\varphi)_{L^2(\Omega)} + \frac{\nu}{2}\int_\Omega \mathrm{Def}(\tilde{u}_0) : \mathrm{Def}\,\varphi dx - \left(\tilde{q}_0 - \frac{1}{\theta}\operatorname{div}\tilde{u}_0, \operatorname{div}\varphi\right)_{L^2(\Omega)} \\
& - (\tilde{F}(0),\varphi)_{L^2(\Omega)} - \sigma(\bar{\mathcal{M}}_0^{\epsilon_2}(0)(0,1),\varphi)_{L^2(\Gamma)} + \kappa(\Delta_0\tilde{u}_0,\Delta_0\varphi)_{L^2(\Gamma)}.
\end{aligned}$$

By compatibility conditions (7.6) and (7.7), $c(\varphi) = 0$. Therefore, the weak limit (w_θ, h_θ) satisfies, for all $t \in [0,T]$,

$$\begin{aligned}(7.15)\qquad &(w_{\theta t},\varphi)_{L^2(\Omega)} + \frac{\nu}{2}\int_\Omega D_{\bar\eta}(w_\theta) : D_{\bar\eta}(\varphi)dx + \kappa\int_\Gamma \Delta_0 w_\theta \cdot \Delta_0\varphi dS \\ &- (\bar a_i^j q_\theta, \varphi^i_{,j})_{L^2(\Omega)} + \sigma\int_\Gamma \bar A^{\alpha\beta\gamma\delta} h^{\epsilon_2}_{\theta,\alpha\beta}[-\bar h_{,\sigma}(\varphi^\sigma\circ\bar\eta^{-\tau}) + \varphi^z\circ\bar\eta^{-\tau}]^{\epsilon_2}_{,\gamma\delta} dS \\ &= \langle \tilde F, \varphi\rangle - \sigma\int_\Gamma \left[L_1^{\alpha\beta\gamma\delta}\bar h_{,\alpha\beta\gamma} + L_2\right]^{\epsilon_2}\left[-\bar h_{,\sigma}\varphi^\sigma\circ\bar\eta^{-\tau} + \varphi^z\circ\bar\eta^{-\tau}\right]^{\epsilon_2} dS\end{aligned}$$

for all $\varphi \in H^{1;2}(\Omega;\Gamma)$.

Since $w_\theta \in L^2(0,T;H^{1;2}(\Omega;\Gamma))$, we can use it as a test function in (7.15) and obtain (after time integration)

$$\begin{aligned}(7.16)\qquad &\frac12\|w_\theta\|^2_{L^2(\Omega)} + \frac{\sigma}{2}E_{\bar h}(h_\theta^{\epsilon_2}) + \int_0^t \left[\frac{\nu}{2}\|D_{\bar\eta}w_\theta\|^2_{L^2(\Omega)} + \kappa\|\Delta_0 w_\theta\|^2_{L^2(\Gamma)}\right. \\ &\left. + \theta\|q_\theta\|^2_{L^2(\Omega)}\right]ds - \theta\int_0^t (q_\theta,\tilde q_0)dt - \frac{\sigma}{2}\int_0^t\int_\Gamma (\bar A^{\alpha\beta\gamma\delta})_t h^{\epsilon_2}_{\theta,\alpha\beta}h^{\epsilon_2}_{\theta,\gamma\delta}dSds \\ &= \frac12\|\tilde u_0\|^2_{L^2(\Omega)} + \int_0^t \langle\tilde F,\varphi\rangle + \sigma\langle \bar{\mathcal M}^{\epsilon_2}_{\bar h}(-\nabla_0\bar h\circ\bar\eta^\tau, 1),\varphi\rangle_\Gamma dt.\end{aligned}$$

Consequently,

$$\begin{aligned}&\left[\|w_\theta(t)\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_\theta^{\epsilon_2}(t)\|^2_{L^2(\Gamma)}\right] + \int_0^t \|\nabla w_\theta\|^2_{L^2(\Omega)}ds + \kappa\int_0^t \|w_\theta\|^2_{H^2(\Gamma)}ds \\ &+ \theta\int_0^t \|q_\theta\|^2_{L^2(\Omega)}ds \\ &\le C(M)\left[\|\tilde u_0\|^2_{L^2(\Omega)} + \theta\|\tilde q_0\|^2_{L^2(\Omega)} + \|\tilde F\|^2_{H^1(\Omega)'} + \|\bar{\mathcal M}^{\epsilon_2}_{\bar h}(-\nabla_0\bar h\circ\bar\eta^\tau,1)\|^2_{L^2(\Gamma)}\right] \\ &+ C(M)\int_0^t \|\bar h_t\|_{H^{2.5}(\Gamma)}\|\nabla_0^2 h_\theta^{\epsilon_2}\|^2_{L^2(\Gamma)}ds \\ &\le C(M)\left[N_1(u_0,F) + \int_0^t \|\bar h_t\|_{H^{2.5}(\Gamma)}\|\nabla_0^2 h_\theta^{\epsilon_2}\|^2_{L^2(\Gamma)}ds\right],\end{aligned}$$

where

$$\begin{aligned}N_1(u_0,F) = {}&\|u_0\|^2_{H^2(\Omega)} + \|u_0\|^2_{H^{4.5}(\Gamma)} + \|F\|^2_{L^2(0,T;H^1(\Omega)')} + \|F_t\|^2_{L^2(0,T;H^1(\Omega)')} \\ &+ \|F(0)\|^2_{H^1(\Omega)} + 1.\end{aligned}$$

By the Gronwall inequality,

$$\begin{aligned}(7.17)\qquad &\sup_{0\le t\le T}\left[\|w_\theta(t)\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_\theta^{\epsilon_2}(t)\|^2_{L^2(\Gamma)}\right] + \int_0^T \left[\|\nabla w_\theta\|^2_{L^2(\Omega)} + \theta\|q_\theta\|^2_{L^2(\Omega)}\right]ds \\ &\le C(M)N_1(u_0,F).\end{aligned}$$

7.4. Improved pressure estimates. By ϵ_2-regularization, we can rewrite (7.15) as, for a.a. $t \in [0,T]$,

$$\begin{aligned}&(w_{\theta t},\varphi)_{L^2(\Omega)} + \frac{\nu}{2}\int_\Omega D_{\bar\eta}(w_\theta) : D_{\bar\eta}(\varphi)dx + \kappa(\Delta_0 w_\theta, \Delta_0\varphi)_{L^2(\Gamma)} - (\bar a_i^j q_\theta, \varphi^i_{,j})_{L^2(\Omega)} \\ &+ \sigma\int_\Gamma \bar{\mathcal L}^{\epsilon_2}_{\bar h}(h_\theta^{\epsilon_2})\left[-\bar h_{,\sigma}\circ\bar\eta^\tau\varphi^\sigma + \varphi^z\right]dS = \langle\tilde F,\varphi\rangle + \sigma\langle\bar{\mathcal M}^{\epsilon_2}_{\bar h}(-\nabla_0\bar h\circ\bar\eta^\tau,1),\varphi\rangle_\Gamma.\end{aligned}$$

Therefore, by the Lagrange multiplier lemma, we conclude that

$$\|q_\theta\|^2_{L^2(\Omega)} \le C(M)\Big[\|w_{\theta t}\|^2_{H^1(\Omega)'} + \|\nabla w_\theta\|^2_{L^2(\Omega)} + \|\tilde F\|^2_{H^1(\Omega)'} + \kappa\|\Delta_0^2 w_\theta\|^2_{H^{-2}(\Gamma)} + \|[\bar{\mathcal{L}}^{\epsilon_2}_{\bar h}(h^{\epsilon_2}_\theta) + \bar{\mathcal{M}}^{\epsilon_2}_{\bar h}](-\nabla_0 \bar h \circ \bar\eta^\tau, 1)\|^2_{H^{-2}(\Gamma)}\Big],$$

and hence

$$\|q_\theta\|^2_{L^2(\Omega)} \le C(M)\Big[\|w_{\theta t}\|^2_{L^2(\Omega)} + \|\nabla w_\theta\|^2_{L^2(\Omega)} + \kappa\|w_\theta\|^2_{H^2(\Gamma)} + \|\nabla_0^2 h_\theta\|^2_{L^2(\Gamma)} + \|F\|^2_{H^1(\Omega)'} + 1\Big]. \tag{7.18}$$

7.5. Weak limits as $\theta \to 0$. Since $w_{\theta t} \in L^2(0,T;H^{1;2}(\Omega;\Gamma))$, we can use it as a test function in (7.14). Similar to the way we obtain (7.11), we find that

$$\begin{aligned}
&\frac12\|w_{\theta t}\|^2_{L^2(\Omega)} + \frac{\nu}{2}\int_0^t \|D_{\bar\eta} w_{\theta t}\|^2_{L^2(\Omega)} ds + \frac{\sigma}{2} E_{\bar h}(h^{\epsilon_2}_{\theta t}) + \kappa\int_0^t \|\Delta_0^2 w_{\theta t}\|^2_{L^2(\Gamma)} ds \\
&\quad + \theta\int_0^t \|q_{\theta t}\|^2_{L^2(\Omega)} ds + \int_0^t (q_{\theta t}, \bar a^j_{it} w^i_{\theta,j})_{L^2(\Omega)} ds - \int_0^t (q_\theta, \bar a^j_i w^i_{\theta t,j}) ds \\
&\le C(M) N_0(u_0,F) + C(M)\int_0^t \|\bar v(t')\|^2_{H^3(\Omega)} \int_0^{t'} \|\nabla w_{\theta t}(s)\|^2_{L^2(\Omega)} ds dt' \\
&\quad + C(\epsilon_2)\int_0^t \Big[1 + \|\bar h_t\|_{H^{2.5}(\Gamma)}\Big] \|\nabla_0^2 h^{\epsilon_2}_{\theta t}\|^2_{L^2(\Gamma)} ds.
\end{aligned}$$

By (7.18),

$$\begin{aligned}
\left|\int_0^t (q_\theta, \bar a^j_i w^i_{\theta t,j}) ds\right| &\le C(M,\delta)\int_0^t \|q_\theta\|^2_{L^2(\Omega)} ds + \delta\int_0^t \|\nabla w_{\theta t}\|^2_{L^2(\Omega)} ds \\
&\le C(M)\Big[N_1(u_0,F) + \int_0^t \Big(\|w_{\theta t}\|^2_{L^2(\Omega)} + \kappa\|w_\theta\|^2_{H^2(\Gamma)} + \|\nabla_0^2 h_\theta\|^2_{L^2(\Gamma)}\Big) ds\Big] \\
&\quad + \delta\int_0^t \|\nabla w_{\theta t}\|^2_{L^2(\Omega)} ds,
\end{aligned} \tag{7.19}$$

where (7.17) is used to bound $\|\nabla w_\theta\|^2_{L^2(0,T;L^2(\Omega))}$.

Integrating by parts,

$$\begin{aligned}
\int_0^t (q_{\theta t}, \bar a^j_{it} w^i_{\theta,j})_{L^2(\Omega)} ds &= (q_\theta, \bar a^j_{it} w^i_{\theta,j})_{L^2(\Omega)}(t) + (\tilde q_0, \tilde u^j_{0,i} \tilde u^i_{0,j})_{L^2(\Omega)} \\
&\quad - \int_0^t (q_\theta, \bar a^j_{itt} w^i_{\theta,j})_{L^2(\Omega)} ds - \int_0^t (q_\theta, \bar a^j_{it} w^i_{\theta t,j})_{L^2(\Omega)} ds.
\end{aligned}$$

By ϵ_1-regularization, the last two terms can be bounded by

$$C(M)\int_0^t \|q_\theta\|_{L^2(\Omega)}\Big[C(\epsilon_1)\|\nabla w_\theta\|_{L^2(\Omega)} + \|\nabla w_{\theta t}\|_{L^2(\Omega)}\Big] ds,$$

and hence

$$\left|\int_0^t (q_\theta, \bar{a}^j_{itt} w^i_{\theta,j})_{L^2(\Omega)} ds\right| + \left|\int_0^t (q_\theta, \bar{a}^j_{it} w^i_{\theta t,j})_{L^2(\Omega)} ds\right|$$

$$\leq C(M,\delta)\int_0^t \|q_\theta\|^2_{L^2(\Omega)} ds + C(\epsilon_1)\int_0^t \|\nabla w_\theta\|^2_{L^2(\Omega)} ds + \delta\int_0^t \|\nabla w_{\theta t}\|^2_{L^2(\Omega)} ds$$

$$\leq C(\epsilon_1,\delta)N_1(u_0,F) + C(M,\delta)\int_0^t \|w_{\theta t}\|^2_{L^2(\Omega)} ds + C(\epsilon_2)\int_0^t \|\nabla_0^2 h_\theta\|^2_{L^2(\Gamma)} ds$$

$$+\,\delta\int_0^t \|\nabla w_{\theta t}\|^2_{L^2(\Omega)} ds. \tag{7.20}$$

For $(q_\theta, \bar{a}^j_{it} w^i_{\theta,j})_{L^2(\Omega)}(t)$, it is easy to see that

$$\left|(q_\theta, \bar{a}^j_{it} w^i_{\theta,j})_{L^2(\Omega)}(t)\right| \leq \delta_1 \|w_{\theta t}\|^2_{L^2(\Omega)} + C(\epsilon_1,\delta_1)\|\nabla w_\theta\|^2_{L^2(\Omega)}$$

$$\leq C(\epsilon_1,\delta_1)\|\nabla w_\theta\|^2_{L^2(\Omega)} + \delta_1 C(\epsilon_2)\|\nabla_0^2 h_\theta\|^2_{L^2(\Gamma)} + \delta_1\Big[\|w_{\theta t}\|^2_{L^2(\Omega)} + \|F\|_{L^2(\Omega)} + 1\Big],$$

while for $(\tilde{q}_0, \tilde{u}^j_{0,i}\tilde{u}^i_{0,j})_{L^2(\Omega)}$, it is bounded by $C(M)N_1(u_0,F)$. Combining (7.19), (7.20), and the estimates above, by choosing $\delta > 0$ and $\delta_1 > 0$ small enough,

$$\|w_{\theta t}\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_{\theta t}\|^2_{L^2(\Gamma)} + \int_0^t \Big[\|\nabla w_{\theta t}\|^2_{L^2(\Omega)} + \kappa\|w_{\theta t}\|^2_{H^2(\Gamma)} + \theta\|q_{\theta t}\|^2_{L^2(\Omega)}\Big] ds$$

$$\leq C(\epsilon_2,\epsilon_1)\Bigg[N_2(u_0,F) + \int_0^t \Bigg(\|w_{\theta t}\|^2_{L^2(\Omega)} + (1+\|\bar{h}_t\|_{H^{2.5}(\Gamma)})\|\nabla_0^2 h_{\theta t}\|^2_{L^2(\Gamma)}$$

$$+\,\|\bar{v}\|^2_{H^3(\Omega)}\int_0^s \|\nabla w_{\theta t}\|^2_{L^2(\Omega)} dt'\Bigg) ds\Bigg] + C_1(\epsilon_2,\epsilon_1)\|\nabla w_\theta\|^2_{L^2(\Omega)},$$

where $N_2(u_0,F) = N_1(u_0,F) + \|F\|^2_{L^\infty(0,T;L^2(\Omega))}$. By the Gronwall inequality,

$$\|w_{\theta t}\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_{\theta t}\|^2_{L^2(\Gamma)} + \int_0^t \Big[\|\nabla w_{\theta t}\|^2_{L^2(\Omega)} + \kappa\|w_{\theta t}\|^2_{H^2(\Gamma)}\Big] ds$$

$$\leq C(\epsilon_2,\epsilon_1)N_2(u_0,F) + C_1(\epsilon_2,\epsilon_1)\|\nabla w_\theta\|^2_{L^2(\Omega)}. \tag{7.21}$$

By using $w_\theta(t) = \tilde{u}_0 + \int_0^t w_{\theta t} ds$, we see that

$$\|w_{\theta t}\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_{\theta t}\|^2_{L^2(\Gamma)} + \int_0^t \Big[\|\nabla w_{\theta t}\|^2_{L^2(\Omega)} + \kappa\|w_{\theta t}\|^2_{H^2(\Gamma)}\Big] ds$$

$$\leq C(\epsilon_2,\epsilon_1)N_2(u_0,F) + C_1(\epsilon_2,\epsilon_1)t\int_0^t \|\nabla w_{\theta t}\|^2_{L^2(\Omega)} ds.$$

Therefore, for any $0 \leq t \leq t_1 = \min\{T, \frac{1}{2C_1}\}$, we have

$$\|w_{\theta t}\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_{\theta t}\|^2_{L^2(\Gamma)} + \frac{1}{2}\int_0^t \Big[\|\nabla w_{\theta t}\|^2_{L^2(\Omega)} + \kappa\|w_{\theta t}\|^2_{H^2(\Gamma)}\Big] ds$$

$$\leq C(\epsilon_2,\epsilon_1)N_2(u_0,F).$$

By $w_\theta(t_1) = \tilde{u}_0 + \int_0^{t_1} w_{\theta t} ds$, we also have

$$\|\nabla w_\theta(t_1)\|^2_{L^2(\Omega)} \leq C(\epsilon_2,\epsilon_1)N_2(u_0,F). \tag{7.22}$$

For $t \geq t_1$, since $w_\theta(t) = w_\theta(t_1) + \int_{t_1}^t w_{\theta t} ds$, we have from (7.21) and (7.22) that

$$\begin{aligned}
&\|w_{\theta t}\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_{\theta t}\|^2_{L^2(\Gamma)} + \int_0^t \left[\|\nabla w_{\theta t}\|^2_{L^2(\Omega)} + \kappa \|w_{\theta t}\|^2_{H^2(\Gamma)}\right] ds \\
&\leq C(\epsilon_2, \epsilon_1) N_2(u_0, F) + C_1(\epsilon_2, \epsilon_1)\left[\|w_\theta(t_1)\|^2_{L^2(\Omega)} + (t - t_1)\int_{t_1}^t \|\nabla_0 w_{\theta t}\|^2_{L^2(\Omega)} ds\right] \\
&\leq C(\epsilon_2, \epsilon_1) N_2(u_0, F) + C_1(\epsilon_2, \epsilon_1)(t - t_1)\int_{t_1}^t \|\nabla_0 w_{\theta t}\|^2_{L^2(\Omega)} ds.
\end{aligned}$$

Therefore, for any $t_1 \leq t \leq 2t_1$, we also have

$$\begin{aligned}
&\|w_{\theta t}\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_{\theta t}\|^2_{L^2(\Gamma)} + \frac{1}{2}\int_0^t \left[\|\nabla w_{\theta t}\|^2_{L^2(\Omega)} + \kappa \|w_{\theta t}\|^2_{H^2(\Gamma)}\right] ds \\
&\leq C(\epsilon_2, \epsilon_1) N_2(u_0, F),
\end{aligned}$$

which with $w_\theta(2t_1) = \tilde{u}_0 + \int_0^{2t_1} w_{\theta t} ds$ gives

$$\|\nabla w_\theta(2t_1)\|^2_{L^2(\Omega)} \leq C(\epsilon_2, \epsilon_1) N_2(u_0, F).$$

By induction, for any $t \in [0, T]$,

$$\begin{aligned}
&\|w_{\theta t}\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_{\theta t}\|^2_{L^2(\Gamma)} + \frac{1}{2}\int_0^t \left[\|\nabla w_{\theta t}\|^2_{L^2(\Omega)} + \kappa \|w_{\theta t}\|^2_{H^2(\Gamma)}\right] ds \\
&\leq C(\epsilon_2, \epsilon_1) N_2(u_0, F).
\end{aligned} \tag{7.23}$$

We also get a θ-independent bound for $\|q_\theta\|^2_{L^2(0,T;L^2(\Omega))}$ by (7.18):

$$\|q_\theta\|^2_{L^2(0,T;L^2(\Omega))} \leq C(\epsilon_2, \epsilon_1) N_2(u_0, F). \tag{7.24}$$

Let $\theta = \frac{1}{m}$. Energy inequalities (7.17), (7.23), and (7.24) show that there exists a subsequence $w_{\frac{1}{m_\ell}}$ such that

$$w_{\frac{1}{m_\ell}} \rightharpoonup \mathfrak{v} \quad \text{in} \quad L^2(0, T; H^{1;2}(\Omega; \Gamma)), \tag{7.25a}$$

$$w_{\frac{1}{m_\ell} t} \rightharpoonup \mathfrak{v}_t \quad \text{in} \quad L^2(0, T; H^{1;2}(\Omega; \Gamma)), \tag{7.25b}$$

$$\nabla_0^2 h_{\frac{1}{m_\ell}} \rightharpoonup \nabla_0^2 \mathfrak{h} \quad \text{in} \quad L^2(0, T; L^2(\Omega)), \tag{7.25c}$$

$$\nabla_0^2 h_{\frac{1}{m_\ell} t} \rightharpoonup \nabla_0^2 \mathfrak{h}_t \quad \text{in} \quad L^2(0, T; L^2(\Omega)), \tag{7.25d}$$

$$q_{\frac{1}{m_\ell}} \rightharpoonup \mathfrak{q} \quad \text{in} \quad L^2(0, T; L^2(\Omega)). \tag{7.25e}$$

Moreover, (7.17) also shows that $\|\bar{a}_i^j w^i_{\frac{1}{m},j}\|_{L^2(0,T;L^2(\Omega))} \to 0$ as $m \to \infty$. Therefore, the weak limit $\mathfrak{v}$ satisfies the "divergence-free" condition (7.2b), i.e.,

$$\mathfrak{v} \in \mathcal{V}_{\bar{v}}(T). \tag{7.26}$$

Since (7.17) is independent of θ and ϵ_2, by the property of lower semicontinuity of norms,

$$\begin{aligned}
&\sup_{0 \leq t \leq T}\left[\|\mathfrak{v}(t)\|^2_{L^2(\Omega)} + \|\nabla_0^2 \mathfrak{h}(t)\|^2_{L^2(\Gamma)}\right] + \|\nabla \mathfrak{v}\|^2_{L^2(0,T;L^2(\Omega))} + \kappa \|\mathfrak{v}\|^2_{H^2(\Gamma)} \\
&\leq C(M) N_1(u_0, F).
\end{aligned} \tag{7.27}$$

By (7.25) and ϵ_2-regularization, the weak limit $(\mathfrak{v}, \mathfrak{h}, \mathfrak{q})$ satisfies, for all $\varphi \in L^2(0,T; H^{1;2}(\Omega;\Gamma))$,

$$\begin{aligned}
&\int_0^T (\mathfrak{v}_t, \varphi)_{L^2(\Omega)}dt + \frac{\nu}{2}\int_0^T\int_\Omega D_{\bar\eta}(\mathfrak{v}) : D_{\bar\eta}(\varphi)dxdt + \kappa\int_0^T\int_\Gamma \Delta_0\mathfrak{v}\cdot\Delta_0\varphi dSdt \\
&-\int_0^T (\bar a_i^j \mathfrak{q}, \varphi^i_{,j})_{L^2(\Omega)}dt + \sigma\int_0^T\int_\Gamma \bar A^{\alpha\beta\gamma\delta}\mathfrak{h}^{\epsilon_2}_{,\alpha\beta}[-\bar h_{,\sigma}(\varphi^\sigma\circ\bar\eta^{-\tau}) + \varphi^z\circ\bar\eta^{-\tau}]^{\epsilon_2}_{,\gamma\delta}dSdt \\
= &\int_0^T \left\{\langle \tilde F, \varphi\rangle - \sigma\int_\Gamma \left[L_1^{\alpha\beta\gamma\delta}\bar h_{,\alpha\beta\gamma} + L_2\right]^{\epsilon_2}\left[-\bar h_{,\sigma}\varphi^\sigma\circ\bar\eta^{-\tau} + \varphi^z\circ\bar\eta^{-\tau}\right]^{\epsilon_2} dS\right\}dt.
\end{aligned}$$

By the density argument, we find that for a.a. $t\in[0,T]$, $\varphi\in H^{1;2}(\Omega;\Gamma)$,

$$\begin{aligned}
&(\mathfrak{v}_t, \varphi)_{L^2(\Omega)} + \frac{\nu}{2}\int_\Omega D_{\bar\eta}(\mathfrak{v}) : D_{\bar\eta}(\varphi)dx + \kappa\int_\Gamma \Delta_0\mathfrak{v}\cdot\Delta_0\varphi dS - (\bar a_i^j\mathfrak{q}, \varphi^i_{,j})_{L^2(\Omega)} \\
&+\sigma\int_\Gamma \bar A^{\alpha\beta\gamma\delta}\mathfrak{h}^{\epsilon_2}_{,\alpha\beta}[-\bar h_{,\sigma}(\varphi^\sigma\circ\bar\eta^{-\tau}) + \varphi^z\circ\bar\eta^{-\tau}]^{\epsilon_2}_{,\gamma\delta}dS \\
&= \langle\tilde F, \varphi\rangle - \sigma\int_\Gamma \left[L_1^{\alpha\beta\gamma\delta}\bar h_{,\alpha\beta\gamma} + L_2\right]^{\epsilon_2}\left[-\bar h_{,\sigma}\varphi^\sigma\circ\bar\eta^{-\tau} + \varphi^z\circ\bar\eta^{-\tau}\right]^{\epsilon_2}dS,
\end{aligned} \tag{7.28}$$

or after a change of variable $y' = \bar\eta^\tau(y,t)$,

$$\begin{aligned}
&(\mathfrak{v}_t, \varphi)_{L^2(\Omega)} + \frac{\nu}{2}(D_{\bar\eta}\mathfrak{v}, D_{\bar\eta}\varphi)_{L^2(\Omega)} + \kappa\int_\Gamma \Delta_0\mathfrak{v}\cdot\Delta_0\varphi dS - (\bar a_i^j\mathfrak{q}, \varphi^i_{,j})_{L^2(\Omega)} \\
&+\sigma\int_\Gamma \mathcal{L}^{\epsilon_2}_{\bar h}(\mathfrak{h})(-\nabla_0\bar h\circ\bar\eta^\tau, 1)\cdot\varphi dS = \langle\tilde F,\varphi\rangle - \sigma\int_\Gamma \bar{\mathcal{M}}^{\epsilon_2}_{\bar h}(-\nabla_0\bar h\circ\bar\eta^\tau, 1)\cdot\varphi dS.
\end{aligned} \tag{7.29}$$

Furthermore, if $\varphi\in\mathcal{V}_{\bar v}$, then

$$\begin{aligned}
&(\mathfrak{v}_t, \varphi)_{L^2(\Omega)} + \frac{\nu}{2}(D_{\bar\eta}\mathfrak{v}, D_{\bar\eta}\varphi)_{L^2(\Omega)} + \kappa\int_\Gamma \Delta_0\mathfrak{v}\cdot\Delta_0\varphi dS \\
&+\sigma\int_\Gamma \mathcal{L}^{\epsilon_2}_{\bar h}(\mathfrak{h})(-\nabla_0\bar h\circ\bar\eta^\tau, 1)\cdot\varphi dS = \langle\tilde F,\varphi\rangle - \sigma\int_\Gamma \bar{\mathcal{M}}^{\epsilon_2}_{\bar h}(-\nabla_0\bar h\circ\bar\eta^\tau, 1)\cdot\varphi^{\epsilon_2} dS
\end{aligned}$$

for a.a. $t\in[0,T]$. In other words, $(\mathfrak{v},\mathfrak{h},\mathfrak{q})$ is a weak solution of (7.2).

8. Estimates independent of ϵ_2.

8.1. Partition of unity. Since Ω is compact, by partition of unity, we can choose two nonnegative smooth functions ζ_0 and ζ_1 so that

$$\begin{aligned}
&\zeta_0 + \zeta_1 = 1 \quad \text{in } \Omega, \\
&\operatorname{supp}(\zeta_0) \subset\subset \Omega, \\
&\operatorname{supp}(\zeta_1) \subset\subset \Gamma\times(-\epsilon_1, \epsilon_1) := \Omega_1.
\end{aligned}$$

We will assume that $\zeta_1 = 1$ inside the region $\Omega_1' \subset \Omega_1$ and $\zeta_0 = 1$ inside the region $\Omega'\subset\Omega$. Note that then $\zeta_1 = 1$, while $\zeta_0 = 0$ on Γ.

8.2. Higher regularity.

8.2.1. ϵ_2-independent bounds for $\mathfrak{q}$. Similar to (7.18), we have

$$\begin{aligned}
\|\mathfrak{q}\|^2_{L^2(\Omega)} \le C(M)\Big[&\|\mathfrak{v}_t\|^2_{L^2(\Omega)} + \|\nabla\mathfrak{v}\|^2_{L^2(\Omega)} + \kappa\|\mathfrak{v}\|^2_{H^2(\Gamma)} + \|\nabla_0^2\mathfrak{h}^{\epsilon_2}\|^2_{L^2(\Gamma)} \\
&+ \|F\|^2_{L^2(\Omega)} + 1\Big].
\end{aligned} \tag{8.1}$$

8.2.2. Interior regularity. Converting the fluid equation (7.2) into Eulerian variables by composing with $\bar{\eta}^{-1}$, we obtain a Stokes problem in the domain $\bar{\eta}(\Omega)$:

$$-\nu\Delta\mathfrak{u} + \nabla\mathfrak{p} = \tilde{F}\circ\bar{\eta}^{-1} - \mathfrak{v}_t\circ\bar{\eta}^{-1} + \nu\bar{a}^j_{\ell,j}\circ\bar{\eta}^{-1}\mathfrak{u}_{,\ell} - \mathfrak{p}\bar{a}^j_{i,j}\circ\bar{\eta}^{-1}, \tag{8.2a}$$

$$\operatorname{div}\mathfrak{u} = 0, \tag{8.2b}$$

where $\mathfrak{u} = \mathfrak{v}\circ\bar{\eta}^{-1}$ and $\mathfrak{p} = \mathfrak{q}\circ\bar{\eta}^{-1}$. By the regularity results for the Stokes problem,

$$\begin{aligned} &\|\mathfrak{u}\|^2_{H^2(\bar{\eta}(\Omega))} + \|\mathfrak{p}\|^2_{H^1(\bar{\eta}(\Omega))} \\ &\le C\Big[\|\tilde{F}\circ\bar{\eta}^{-1}\|^2_{L^2(\bar{\eta}(\Omega))} + \|\mathfrak{v}_t\circ\bar{\eta}^{-1}\|^2_{L^2(\bar{\eta}(\Omega))} + \|\nabla\mathfrak{u}\|^2_{L^2(\bar{\eta}(\Omega))} + \|\mathfrak{p}\|^2_{L^2(\bar{\eta}(\Omega))} \\ &\qquad + \|\mathfrak{u}\|^2_{H^{1.5}(\Gamma)}\Big] \end{aligned}$$

or

$$\begin{aligned} \|\mathfrak{v}\|^2_{H^2(\Omega)} + \|\mathfrak{q}\|^2_{H^1(\Omega)} \le C\Big[\|F\|^2_{L^2(\Omega)} + \|\mathfrak{v}_t\|^2_{L^2(\Omega)} + \|\mathfrak{v}\|^2_{H^{1.5}(\Gamma)}\Big] \\ + C(M)\Big[\|\nabla\mathfrak{v}\|^2_{L^2(\Omega)} + \|\mathfrak{q}\|^2_{L^2(\Omega)}\Big] \end{aligned}$$

for some constant C independent of M and ϵ_1. By (8.1),

$$\begin{aligned} \|\mathfrak{v}\|^2_{H^2(\Omega)} + \|\mathfrak{q}\|^2_{H^1(\Omega)} \le C(M)\Big[\|\mathfrak{v}_t\|^2_{L^2(\Omega)} + \|\nabla\mathfrak{v}\|^2_{L^2(\Omega)} + \|\mathfrak{v}\|^2_{H^2(\Gamma)} \\ + \|\nabla_0^2\mathfrak{h}^{\epsilon_2}\|^2_{L^2(\Gamma)} + \|F\|^2_{L^2(\Omega)} + 1\Big]. \end{aligned} \tag{8.3}$$

Similarly,

$$\begin{aligned} \|\mathfrak{v}\|^2_{H^3(\Omega)} + \|\mathfrak{q}\|^2_{H^2(\Omega)} \le C\Big[\|F\|^2_{H^1(\Omega)} + \|\mathfrak{v}_t\|^2_{H^1(\Omega)} + \|\mathfrak{v}\|^2_{H^{2.5}(\Gamma)}\Big] \\ + C(M)\Big[\|\nabla\mathfrak{v}\|^2_{H^1(\Omega)} + \|\mathfrak{q}\|^2_{H^1(\Omega)}\Big], \end{aligned}$$

and therefore by (8.1) and (8.3),

$$\begin{aligned} \|\mathfrak{v}\|^2_{H^3(\Omega)} + \|\mathfrak{q}\|^2_{H^2(\Omega)} \le C(M)\Big[\|\mathfrak{v}_t\|^2_{H^1(\Omega)} + \|\nabla\mathfrak{v}\|^2_{L^2(\Omega)} + \|\nabla_0^2\mathfrak{v}\|^2_{H^1(\Omega_1)} \\ + \|\nabla_0^2\mathfrak{h}^{\epsilon_2}\|^2_{L^2(\Gamma)} + \|F\|^2_{H^1(\Omega)} + 1\Big]. \end{aligned} \tag{8.4}$$

For the regularized problem, because the ϵ_1-regularization ensures that the forcing and the initial data are smooth, while the ϵ_2-regularization ensures that the right-hand side of (7.2c) is smooth, by the standard difference quotient technique, it is also easy to see that

$$\nabla_0^k\mathfrak{v} \in L^2(0,T;H^1(\Omega_1)\cap H^2(\Gamma)) \quad \text{for } k = 1,2,3,4. \tag{8.5}$$

Since (7.25b) implies that $\mathfrak{v}_t \in L^2(0,T;H^1(\Omega))$, by ϵ_2-regularization and (8.4) we conclude that

$$\mathfrak{v} \in L^2(0,T;H^3(\Omega)), \qquad \mathfrak{q} \in L^2(0,T;H^2(\Omega)). \tag{8.6}$$

8.3. Estimates for $\mathfrak{v}_t(0)$ and $\mathfrak{q}(0)$. By (8.6) and ϵ_2-regularization, $(\mathfrak{v},\mathfrak{h},\mathfrak{q})$ satisfies the strong form (7.2). Taking the "divergence" of (7.2a) and then making use of condition (7.2b), we find that

$$-\bar{a}^k_{it}\mathfrak{v}^i_{,k} - \nu\bar{a}^k_i[\bar{a}^j_\ell D_{\bar{\eta}}(\mathfrak{v})^i_\ell]_{,jk} = -\bar{a}^k_i(\bar{a}^j_i\mathfrak{q})_{,jk} + \bar{a}^k_i\tilde{F}^i_{,k}. \tag{8.7}$$

Let $t=0$; by the identity $\bar{a}^\ell_{kt} = -\bar{a}^i_k\bar{v}^j_{,i}\bar{a}^\ell_j$,

$$\Delta\mathfrak{q}(0) = \nabla\tilde{u}_0 : (\nabla\tilde{u}_0)^T - \operatorname{div}(\tilde{F}(0)) \quad \text{in } \Omega$$

with

$$\mathfrak{q}(0) = \nu(\operatorname{Def}\tilde{u}_0)^j_i N_i N_j - \sigma\mathcal{M}^{\epsilon_2}_0(0) + \kappa\Delta_0^2\tilde{u}_0 \quad \text{on } \Gamma,$$

while (7.2a) gives us

$$\mathfrak{v}_t(0) = \nu\Delta\tilde{u}_0 - \nabla\mathfrak{q}(0) + \tilde{F}(0) \quad \text{in } \Omega.$$

By standard elliptic regularity result,

$$\|\mathfrak{v}_t(0)\|^2_{L^2(\Omega)} + \|\mathfrak{q}(0)\|^2_{H^1(\Omega)} \le CN_0(u_0,F) \tag{8.8}$$

for some constant independent of M, ϵ_1, and ϵ_2.

8.4. $L^2_tL^2_x$-estimates for $\mathfrak{v}_t$. Since $\mathfrak{v}_t \in L^2(0,T;H^1(\Omega))$, we can use it as a test function in (7.29). By (7.26), we find that

$$\begin{aligned}
&\|\mathfrak{v}_t\|^2_{L^2(\Omega)} + \frac{\nu}{4}\frac{d}{dt}\int_\Omega |D_{\bar{\eta}}\mathfrak{v}|^2dx - \frac{\nu}{2}\int_\Omega (D_{\bar{\eta}}\mathfrak{v})^j_i\bar{a}^k_{jt}\mathfrak{v}^i_{,k}dx + \kappa\int_\Gamma \Delta_0\mathfrak{v}\cdot\Delta_0\varphi dS\\
&\quad + \int_\Omega \mathfrak{q}\bar{a}^\ell_{kt}\mathfrak{v}^k_{,\ell}dx + \sigma\int_\Gamma \mathcal{L}^{\epsilon_2}_{\bar{h}}(\mathfrak{h})(-\nabla_0\bar{h}\circ\bar{\eta}^\tau,1)\cdot\mathfrak{v}_t dS\\
&= \langle\tilde{F},\mathfrak{v}_t\rangle - \sigma\int_\Gamma \mathcal{M}^{\epsilon_2}_{\bar{h}}(-\nabla_0\bar{h}\circ\bar{\eta}^\tau,1)\cdot\mathfrak{v}_t dS.
\end{aligned}$$

By (5.3),

$$\int_\Omega (D_{\bar{\eta}}\mathfrak{v})^j_i\bar{a}^k_{jt}\mathfrak{v}^i_{,k}dx \le C(M)C(\delta)\|\nabla\mathfrak{v}\|^2_{L^2(\Omega)} + \delta\|\mathfrak{v}\|^2_{H^2(\Omega)},$$

and by (8.1) and the interpolation inequality,

$$\begin{aligned}
\left|\int_\Omega \mathfrak{q}\bar{a}^\ell_{kt}\mathfrak{v}^k_{,\ell}dx\right| &\le C(M)C(\delta)\Big[\|\nabla\mathfrak{v}\|^2_{L^2(\Omega)} + \|\nabla_0^4\mathfrak{h}^{\epsilon_2}\|^2_{L^2(\Gamma)} + \|F\|^2_{L^2(\Omega)} + 1\Big]\\
&\quad + \delta\|\mathfrak{v}\|^2_{H^2(\Omega)} + \frac{1}{2}\|\mathfrak{v}_t\|^2_{L^2(\Omega)}
\end{aligned}$$

for some $C(\delta)$. Also, the last term on the left-hand side is bounded by

$$\begin{aligned}
&C(M)\Big[\|\nabla_0^4\mathfrak{h}^{\epsilon_2}\|_{L^2(\Gamma)} + 1\Big]\|\mathfrak{v}_t\|_{H^1(\Omega)}\\
&\le C(M)C(\delta_1)\Big[\|\nabla_0^4\mathfrak{h}^{\epsilon_2}\|^2_{L^2(\Gamma)} + 1\Big] + \delta_1\|\mathfrak{v}_t\|^2_{H^1(\Omega)}.
\end{aligned}$$

Combining all the estimates above,

$$\begin{aligned}&\frac{1}{2}\|\mathfrak{v}_t\|^2_{L^2(\Omega)}+\frac{\nu}{4}\frac{d}{dt}\int_\Omega |D_{\bar\eta}\mathfrak{v}|^2dx+\frac{\kappa}{2}\frac{d}{dt}\int_\Gamma |\Delta_0\mathfrak{v}|^2dS\\&\le C\Big[\|\nabla\mathfrak{v}\|^2_{L^2(\Omega)}+\|\nabla_0^4\mathfrak{h}^{\epsilon_2}\|^2_{L^2(\Gamma)}+\|F\|^2_{L^2(\Omega)}+1\Big]+\delta\|\mathfrak{v}\|^2_{H^2(\Omega)}+\delta_1\|\mathfrak{v}_t\|^2_{H^1(\Omega)}\end{aligned}$$

for some constant C depending on M, δ, and δ_1. Therefore, by (7.27),

$$\begin{aligned}&\int_0^t\|\mathfrak{v}_t\|^2_{L^2(\Omega)}ds+\|\nabla\mathfrak{v}(t)\|^2_{L^2(\Omega)}+\kappa\|\mathfrak{v}\|^2_{H^2(\Gamma)}\\&\le C\Big[N_2(u_0,F)+\int_0^t\|\nabla_0^4\mathfrak{h}^{\epsilon_2}\|^2_{L^2(\Gamma)}ds\Big]+\delta\int_0^t\|\mathfrak{v}\|^2_{H^2(\Omega)}ds+\delta_1\int_0^t\|\mathfrak{v}_t\|^2_{H^1(\Omega)}ds.\end{aligned}\tag{8.9}$$

8.5. Energy estimates for $\nabla_0^2 v$ near the boundary. Because of (8.5), $\nabla_0^2(\zeta_1^2\nabla_0^2\mathfrak{v})$ in (7.28) can be used as a test function in (7.29). It follows that

$$\begin{aligned}&\left|\int_\Gamma\Big[\bar{\mathcal{L}}^{\epsilon_2}_{\bar h}(\mathfrak{h}^{\epsilon_2})+\bar{\mathcal{M}}^{\epsilon_2}_{\bar h}\Big](-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot\nabla_0^4\mathfrak{v}dS\right|\\&\le C(M)\Big[\|\nabla_0^2\mathfrak{h}^{\epsilon_2}\|_{H^2(\Gamma)}+1\Big]\|\mathfrak{v}\|_{H^4(\Gamma)}\\&\le C(M,\delta_3)\Big[1+\|\mathfrak{h}\|^2_{H^4(\Gamma)}\Big]+\delta_3\|\mathfrak{v}\|^2_{H^4(\Gamma)}.\end{aligned}$$

By (7.4), we find that

$$\|\mathfrak{h}\|^2_{H^4(\Gamma)}\le C(\epsilon_1)\Big[\int_0^t\|\bar h\|_{H^5(\Gamma)}\|\mathfrak{v}\|_{H^4(\Gamma)}ds\Big]^2\le C(\epsilon_1)\int_0^t\|\mathfrak{v}\|^2_{H^4(\Gamma)}ds,$$

and hence

$$\begin{aligned}&\left|\int_\Gamma\Big[\bar{\mathcal{L}}^{\epsilon_2}_{\bar h}(\mathfrak{h}^{\epsilon_2})+\bar{\mathcal{M}}^{\epsilon_2}_{\bar h}\Big](-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot\nabla_0^4\mathfrak{v}dS\right|\\&\le\bar C\Big[1+\int_0^t\|\mathfrak{v}\|^2_{H^4(\Gamma)}\Big]+\delta_3\|\mathfrak{v}\|^2_{H^4(\Gamma)}\end{aligned}$$

for some constant $\bar C$ depending on M, ϵ_1, and δ_3. Since

$$\Delta_0 f=\frac{1}{\sqrt{\det(g_0)}}\frac{\partial}{\partial y^\alpha}\Big[\sqrt{\det(g_0)}g_0^{\alpha\beta}\frac{\partial}{\partial y^\beta}f\Big],$$

by the regularity on Γ (and hence on g_0),

$$\begin{aligned}\int_\Gamma|\Delta_0\nabla_0^2\mathfrak{v}|^2dS&\le\int_\Gamma\Delta_0^2\mathfrak{v}\cdot(\nabla_0^4v)dS+C\|\mathfrak{v}\|_{H^3(\Gamma)}\|\mathfrak{v}\|_{H^4(\Gamma)}\\&\le\int_\Gamma\Delta_0^2v\cdot(\nabla_0^4v)dS+C(\delta)\|\mathfrak{v}\|^2_{H^1(\Omega)}+\delta\|\mathfrak{v}\|^2_{H^4(\Gamma)},\end{aligned}$$

which implies, by choosing $\delta>0$ small enough, that

$$\nu_2\|v\|^2_{H^4(\Gamma)}\le\int_\Gamma\Delta_0^2v\cdot(\nabla_0^4v)dS+C\|v\|^2_{H^1(\Omega)}.$$

By the identity

$$
\begin{aligned}
&(\mathfrak{q}, \bar{a}^\ell_k \nabla_0^2(\zeta_1^2 \nabla_0^2 \mathfrak{v}^k)_{,\ell}) \\
&= (\mathfrak{q}, \nabla_0^2 \bar{a}^\ell_k (\zeta_1^2 \nabla_0^2 \mathfrak{v}^k)_{,\ell}) + 4(\zeta_1 \nabla_0 \mathfrak{q}, \nabla_0 \bar{a}^\ell_k \zeta_{1,\ell} \nabla_0^2 \mathfrak{v}^k) + 2(\nabla_0 \mathfrak{q}, \zeta_1^2 \nabla_0 \bar{a}^\ell_k \nabla_0^2 \mathfrak{v}^k_{,\ell}) \\
&\quad - 2(\zeta_1 \nabla_0 \mathfrak{q}, \nabla_0(\bar{a}^\ell_k \zeta_{1,\ell} \nabla_0^2 \mathfrak{v}^k)) + 2(\mathfrak{q}, \nabla_0(\bar{a}^\ell_k \zeta_{1,\ell} \nabla_0 \zeta_1 \nabla_0^2 \mathfrak{v}^k)) \\
&\quad + (\nabla_0 \mathfrak{q}, \nabla_0(\zeta_1^2 \nabla_0 \bar{a}^\ell_k \nabla_0 \mathfrak{v}^k_{,\ell})),
\end{aligned}
\tag{8.10}
$$

(5.3) and (8.3) imply that

$$
\begin{aligned}
&(\mathfrak{q}, \bar{a}^\ell_k \nabla_0'^2(\zeta_1^2 \nabla_0^2 \mathfrak{v}^k)_{,\ell}) \le C(M) \|\mathfrak{q}\|_{H^1(\Omega)} \|\mathfrak{v}\|_{H^3(\Omega)} \\
&\le C(M)C(\delta)\Big[\|\mathfrak{v}_t\|^2_{L^2(\Omega)} + \|\nabla \mathfrak{v}\|^2_{L^2(\Omega)} + \|\nabla\nabla_0 \mathfrak{v}\|^2_{L^2(\Omega_1)} + \kappa \|\mathfrak{v}\|^2_{H^2(\Gamma)} \\
&\qquad + \|\nabla_0^2 \mathfrak{h}^{\epsilon_2}\|^2_{L^2(\Gamma)} + \|F\|^2_{L^2(\Omega)} + 1\Big] + \delta \|\mathfrak{v}\|^2_{H^3(\Omega)}.
\end{aligned}
$$

For the viscosity term,

$$
\begin{aligned}
&\int_\Omega D_{\bar\eta} \mathfrak{v} : D_{\bar\eta}(\nabla_0^2(\zeta_1^2 \nabla_0^2 \mathfrak{v}))dx \\
&= \|\zeta_1 D_{\bar\eta} \nabla_0^2 \mathfrak{v}\|^2_{L^2(\Omega)} + \frac{1}{2}\int_\Omega \Big[\nabla_0^2(\bar{a}^k_i \bar{a}^\ell_i)\mathfrak{v}^j_{,\ell} + \nabla_0^2(\bar{a}^k_i \bar{a}^\ell_j)\mathfrak{v}^i_{,\ell}\Big](\zeta_1^2 \nabla_0^2 \mathfrak{v}^j)_{,k} dx \\
&\quad + \int_\Omega \Big[\nabla_0(\bar{a}^k_i \bar{a}^\ell_i)\nabla_0 \mathfrak{v}^j_{,\ell} + \nabla_0(\bar{a}^k_i \bar{a}^\ell_j)\nabla_0 \mathfrak{v}^i_{,\ell}\Big](\zeta_1^2 \nabla_0^2 \mathfrak{v}^j)_{,k} dx \\
&\quad + \int_\Omega D_{\bar\eta}(\nabla_0^2 \mathfrak{v})^j_i \bar{a}^k_i \zeta_1 \zeta_{1,k} \nabla_0^2 \mathfrak{v}^j dx,
\end{aligned}
$$

and hence by interpolation

$$
\begin{aligned}
\frac{1}{2}\|\zeta_1 D_{\bar\eta} \nabla_0'^2 \mathfrak{v}\|^2_{L^2(\Omega)} &\le \int_\Omega D_{\bar\eta} \mathfrak{v} : D_{\bar\eta}(\nabla_0^2(\zeta_1^2 \nabla_0^2 \mathfrak{v}))dx \\
&\quad + C(M)C(\delta)\Big[\|\nabla \mathfrak{v}\|^2_{L^2(\Omega)} + \|\nabla\nabla_0 \mathfrak{v}\|^2_{L^2(\Omega_1')}\Big] + \delta \|\mathfrak{v}\|^2_{H^3(\Omega)}.
\end{aligned}
$$

Summing all the estimates, by letting $\delta_3 = \frac{\nu_2 \kappa}{2}$, we conclude that

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\zeta_1 \nabla_0^2 \mathfrak{v}\|^2_{L^2(\Omega)} + \frac{\nu}{4}\|\zeta_1 D_{\bar\eta} \nabla_0^2 \mathfrak{v}\|^2_{L^2(\Omega)} + \frac{\nu_2 \kappa}{2}\|\mathfrak{v}\|^2_{H^4(\Gamma)} \\
&\le \bar{C}\Big[\|\mathfrak{v}_t\|^2_{L^2(\Omega)} + \|\mathfrak{v}\|^2_{H^1(\Omega)} + \|\nabla\nabla_0 \mathfrak{v}\|^2_{L^2(\Omega_1')} + \|\mathfrak{v}\|^2_{H^2(\Gamma)} + \|\nabla_0^2 \mathfrak{h}^{\epsilon_2}\|^2_{L^2(\Gamma)} \\
&\qquad + \|F\|^2_{H^1(\Omega)} + 1\Big] + \bar{C}\int_0^t \|\mathfrak{v}\|^2_{H^4(\Gamma)} ds + \delta \|\mathfrak{v}\|^2_{H^3(\Omega)}
\end{aligned}
$$

for some constant $\bar{C}$ depending on M, κ, ϵ_1, and δ. Integrating the inequality above in time from 0 to t, by (7.27) we find that

$$
\begin{aligned}
&\|\nabla_0^2 \mathfrak{v}(t)\|^2_{L^2(\Omega_1)} + \int_0^t \Big[\|\nabla\nabla_0^2 \mathfrak{v}\|^2_{L^2(\Omega_1)} + \kappa \|\mathfrak{v}\|^2_{H^4(\Gamma)}\Big] ds \\
&\le \bar{C} N_2(u_0, F) + \bar{C}\int_0^t \Big[\|\mathfrak{v}_t\|^2_{L^2(\Omega)} + \|\nabla\nabla_0 \mathfrak{v}\|^2_{L^2(\Omega_1')} + \|\mathfrak{v}\|^2_{H^2(\Gamma)}\Big] ds \\
&\quad + \bar{C}\int_0^t \int_0^s \|\mathfrak{v}(r)\|^2_{H^4(\Gamma)} dr + \delta \int_0^t \|\mathfrak{v}\|^2_{H^3(\Omega)} ds.
\end{aligned}
\tag{8.11}
$$

By using $\nabla_0(\zeta_1^2\nabla_0\mathfrak{v})$ as a testing function in (7.29), similar computations lead to

$$
\begin{aligned}
&\|\nabla_0\mathfrak{v}(t)\|^2_{L^2(\Omega_1)} + \int_0^t \Big[\|\nabla\nabla_0\mathfrak{v}\|^2_{L^2(\Omega_1)} + \kappa\|\mathfrak{v}\|^2_{H^3(\Gamma)}\Big]ds \\
&\le C(M)N_2(u_0,F) + C(M,\delta)\int_0^t \Big[\|\mathfrak{v}_t\|^2_{L^2(\Omega)} + \kappa\|\mathfrak{v}\|^2_{H^2(\Gamma)}\Big]ds \\
&\quad + C(M)\int_0^t\int_0^s \|\mathfrak{v}(r)\|^2_{H^4(\Gamma)}drds + \delta\int_0^t \|\mathfrak{v}\|^2_{H^3(\Omega)}ds.
\end{aligned}
\tag{8.12}
$$

8.6. Energy estimates for v_t: $L^2_tH^1_x$-estimates. In this section, we time differentiate (7.29) and then use $\mathfrak{v}_t$ as a test function to obtain

$$
\begin{aligned}
&\langle \mathfrak{v}_{tt},\mathfrak{v}_t\rangle + \nu\int_\Omega \Big[\bar a^k_\ell (D_{\bar\eta}\mathfrak{v})^i_{\ell,k}\Big]_t \mathfrak{v}^i_t dx + \sigma\int_\Gamma \Big[\bar{\mathcal{L}}^{\epsilon_2}_{\bar h}(\mathfrak{h}^{\epsilon_2})(-\nabla_0\bar h\circ\bar\eta^\tau,1)\Big]_t\cdot\mathfrak{v}_t dS \\
&+\kappa\int_\Gamma |\Delta_0\mathfrak{v}_t|^2 dS - \int_\Omega (\bar a^\ell_k\mathfrak{q})_t\mathfrak{v}^k_{t,\ell}dx = \langle F_t,\mathfrak{v}_t\rangle - \sigma\int_\Gamma \Big[\bar{\mathcal{M}}^{\epsilon_2}_{\bar h}(-\nabla_0\bar h\circ\bar\eta^\tau,1)\Big]_t\cdot\mathfrak{v}_t dS.
\end{aligned}
$$

By the chain rule,

$$
\begin{aligned}
&\int_\Gamma \Big[(\bar{\mathcal{L}}^{\epsilon_2}_{\bar h}(\mathfrak{h}^{\epsilon_2}) + \bar{\mathcal{M}}^{\epsilon_2}_{\bar h})(-\nabla_0\bar h\circ\bar\eta^\tau,1)\Big]_t\cdot\mathfrak{v}_t dS \\
&= \int_\Gamma \bar\Theta_t\Big[L_{\bar h}(\mathfrak{h}^{\epsilon_2})\Big]^{\epsilon_2}\circ\bar\eta^\tau(-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot\mathfrak{v}_t dS \\
&\quad + \int_\Gamma \bar\Theta\bar\eta^\tau_t\cdot\Big[\nabla_0[L_{\bar h}(\mathfrak{h}^{\epsilon_2})]^{\epsilon_2}(-\nabla_0\bar h,1)\Big]\circ\bar\eta^\tau\cdot\mathfrak{v}_t dS \\
&\quad + \int_\Gamma \bar\Theta\Big[[L_{\bar h}(\mathfrak{h}^{\epsilon_2})]^{\epsilon_2}(\nabla_0\bar h,-1)\Big]_t\circ\bar\eta^\tau\cdot\mathfrak{v}_t dS.
\end{aligned}
$$

By using the $H^2(\Gamma)$-$H^{-2}(\Gamma)$ duality pairing with ϵ_1-regularization on $\bar\Theta$ and $\bar v$, it follows that

$$
\begin{aligned}
&\left|\int_\Gamma \Big[(\bar{\mathcal{L}}^{\epsilon_2}_{\bar h}(\mathfrak{h}^{\epsilon_2}) + \bar{\mathcal{M}}^{\epsilon_2}_{\bar h})(-\nabla_0\bar h\circ\bar\eta^\tau,1)\Big]_t\cdot\mathfrak{v}_t dS\right| \\
&\le C(\epsilon_1)\Big[\|\nabla_0^3\mathfrak{h}\|_{L^2(\Gamma)} + \|\nabla_0^2\mathfrak{h}_t\|_{L^2(\Gamma)} + 1\Big]\|\mathfrak{v}_t\|_{H^2(\Gamma)} \\
&\le C(\epsilon_1,\delta_3)\left[\int_0^t \|\mathfrak{v}\|^2_{H^4(\Gamma)}ds + \|\mathfrak{v}\|^2_{H^2(\Gamma)} + 1\right] + \delta_3\|\mathfrak{v}_t\|^2_{H^2(\Gamma)} \\
&\le \bar C\left[\int_0^t \|\mathfrak{v}\|^2_{H^4(\Gamma)}ds + \|\mathfrak{v}\|^2_{H^1(\Omega)} + 1\right] + \delta\|\mathfrak{v}\|^2_{H^3(\Omega)} + \delta_3\|\mathfrak{v}_t\|^2_{H^2(\Gamma)}
\end{aligned}
$$

for some constant $\bar C$ depending on M, ϵ_1, δ, and δ_3, where we estimate $\|\mathfrak{v}\|^2_{H^2(\Gamma)}$ by interpolation.

Also by interpolation,

$$
\begin{aligned}
\int_\Omega |D_{\bar\eta}\mathfrak{v}_t|^2 dx &= 2\int_\Omega \Big[\bar a^k_i D_{\bar\eta}(\mathfrak{v})^j_i\Big]_t \mathfrak{v}^j_{t,k}dx - 2\int_\Omega \Big[(\bar a^k_i\bar a^\ell_i)_t\mathfrak{v}^j_{,\ell} + (\bar a^k_i\bar a^\ell_j)_t\mathfrak{v}^i_{,\ell}\Big]\mathfrak{v}^j_{t,k}dx \\
&\le 2\int_\Omega \Big[\bar a^k_i D_{\bar\eta}(\mathfrak{v})^j_i\Big]_t \mathfrak{v}^j_{t,k}dx + C(M)C(\delta,\delta_1)\|\nabla\mathfrak{v}\|^2_{L^2(\Omega)} \\
&\quad - \int_\Omega (\bar a^\ell_k\mathfrak{q})_t\mathfrak{v}^k_{t,\ell}dx + \delta\|\mathfrak{v}\|^2_{H^2(\Omega)} + \delta_1\|\mathfrak{v}_t\|^2_{H^1(\Omega)}.
\end{aligned}
$$

Note that

$$\langle F_t, \mathfrak{v}_t\rangle \le C\|F_t\|_{H^1(\Omega)'}\|\mathfrak{v}_t\|_{H^1(\Omega)} \le C(\delta_1)\|F_t\|^2_{H^1(\Omega)'} + \delta_1\|\mathfrak{v}_t\|^2_{H^1(\Omega)}.$$

Summing all the estimates above,

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\mathfrak{v}_t\|^2_{L^2(\Omega)} + \frac{\nu}{4}\|\nabla\mathfrak{v}_t\|^2_{L^2(\Omega)} + \kappa\|\Delta_0\mathfrak{v}_t\|^2_{L^2(\Gamma)}\\
&\le \bar{C}\left[\int_0^t \|\mathfrak{v}\|^2_{H^4(\Gamma)}ds + \|\mathfrak{v}\|^2_{H^1(\Omega)} + 1\right] + C(\delta_1)\|F_t\|^2_{H^1(\Omega)'}\\
&\quad + \delta\|\mathfrak{v}\|^2_{H^3(\Omega)} + \delta_1\|\mathfrak{v}_t\|^2_{H^1(\Omega)} + \delta_3\|\mathfrak{v}_t\|^2_{H^2(\Gamma)} + \int_\Omega (\bar{a}^\ell_k\mathfrak{q})_t\mathfrak{v}^k_{t,\ell}dx
\end{aligned} \tag{8.13}$$

for some constant $\bar{C}$ depending on M, κ, δ, and δ_1. As in [7] and [8], the integral involving the pressure q has the following estimate:

$$\begin{aligned}
\int_0^t\int_\Omega (\bar{a}^\ell_k\mathfrak{q})_t\mathfrak{v}^k_{t,\ell}dxds \le C(M)C(\delta,\delta_1)N_3(u_0,F) + \delta\int_0^t \|\mathfrak{v}\|^2_{H^3(\Omega)}ds\\
+ \delta_1\int_0^t \|\mathfrak{v}_t\|^2_{H^1(\Omega)}ds,
\end{aligned}$$

where

$$\begin{aligned}
N_3(u_0,F) := \|u_0\|^2_{H^{2.5}(\Omega)} + \|u_0\|^2_{H^{4.5}(\Gamma)} + \|F\|^2_{L^2(0,T;H^1(\Omega))}\\
+ \|F_t\|^2_{L^2(0,T;H^1(\Omega)')} + \|F(0)\|^2_{H^1(\Omega)} + 1.
\end{aligned}$$

Integrating (8.13) in time from 0 to t and choosing $\delta_1, \delta_3 > 0$ small enough, (7.27) and (8.9) imply that, for all $t \in [0,T]$,

$$\begin{aligned}
&\|\mathfrak{v}_t(t)\|^2_{L^2(\Omega)} + \int_0^t \left[\|\nabla\mathfrak{v}_t\|^2_{L^2(\Omega)} + \kappa\|\mathfrak{v}_t\|^2_{H^2(\Gamma)}\right]ds\\
&\le \bar{C}N_3(u_0,F) + \bar{C}\int_0^t\int_0^s \|\mathfrak{v}(r)\|^2_{H^4(\Gamma)}drds + \delta\int_0^t \|\mathfrak{v}\|^2_{H^3(\Omega)}ds
\end{aligned} \tag{8.14}$$

for some constant $\bar{C}$ depending on M, κ, δ, and δ_2. In (8.14), (8.8) is used to bound $\|v_t(0)\|^2_{L^2(\Omega)}$.

8.7. ϵ_2-independent estimates. Integrating (8.3) in time from 0 to t, (7.27), (8.9), and (8.12) imply that

$$\begin{aligned}
&\int_0^t \left[\|\mathfrak{v}\|^2_{H^2(\Omega)} + \|\mathfrak{q}\|^2_{H^1(\Omega)}\right]ds\\
&\le C(M)N_1(u_0,F) + \int_0^t \left[\|\mathfrak{v}_t\|^2_{L^2(\Omega)} + \|\mathfrak{v}\|^2_{H^2(\Gamma)}\right]ds\\
&\le \bar{C}N_3(u_0,F) + \bar{C}\int_0^t\int_0^s \|\mathfrak{v}(r)\|^2_{H^4(\Gamma)}drds + \delta\int_0^t \|\mathfrak{v}\|^2_{H^3(\Omega)}ds
\end{aligned} \tag{8.15}$$

for some constant $\bar{C}$ depending on M, κ, and δ. Integrating (8.4) in time from 0 to t, making use of (8.11), (8.12), (8.14), and (8.15), and then choosing $\delta > 0$ small enough and T even smaller, we find that

$$\int_0^t \left[\|\mathfrak{v}\|^2_{H^3(\Omega)} + \|\mathfrak{q}\|^2_{H^2(\Omega)}\right]ds \le \bar{C}N_3(u_0,F) + \bar{C}\int_0^t\int_0^s \|\mathfrak{v}(r)\|^2_{H^4(\Gamma)}drds \tag{8.16}$$

for some constant $\bar{C}$ depending on M, κ, and ϵ_1.

Having (8.16), by choosing $\delta_2 > 0$ small enough, the estimates (8.11) can be rewritten as

$$\|\nabla_0^2 \mathfrak{v}(t)\|^2_{L^2(\Omega_1)} + \int_0^t \Big[\|\nabla\nabla_0^2 \mathfrak{v}\|^2_{L^2(\Omega_1)} + \kappa\|\mathfrak{v}\|^2_{H^4(\Gamma)}\Big] ds$$

$$\le \bar{C} N_3(u_0, F) + \bar{C}\int_0^t\int_0^s \|\mathfrak{v}(r)\|^2_{H^4(\Gamma)} dr ds \tag{8.17}$$

for some constant $\bar{C}$ depending on M, κ, and ϵ_1. Therefore,

$$X(t) \le \bar{C}\Big[\int_0^t X(s)ds + N_3(u_0, F)\Big],$$

where

$$X(t) = \int_0^t \|\mathfrak{v}\|^2_{H^4(\Gamma)} ds.$$

By the Gronwall inequality,

$$\int_0^t\int_0^s \|\mathfrak{v}(r)\|^2_{H^4(\Gamma)} dr ds \le \bar{C} N_3(u_0, F) \tag{8.18}$$

for all $t \in [0, T]$ for some constant $\bar{C}$ depending on M, κ, and ϵ_1. Having (8.18), estimates (8.9), (8.14), (8.16), and (8.17) along with the standard embedding theorem lead to

$$\sup_{0\le t\le T}\Big[\|\mathfrak{v}(t)\|^2_{H^2(\Omega)} + \|\mathfrak{v}_t(t)\|^2_{L^2(\Omega)}\Big] + \|\mathfrak{v}\|^2_{\mathcal{V}^3(T)} + \|\mathfrak{q}\|^2_{L^2(0,T;H^2(\Omega))}$$

$$+ \kappa\|\mathfrak{v}\|^2_{L^2(0,T;H^4(\Gamma))} \le \bar{C} N_3(u_0, F) \tag{8.19}$$

for some constant $\bar{C}$ depending on M, κ, and ϵ_1.

8.8. Weak limits as $\epsilon_2 \to 0$. Since the estimate (8.19) is independent of ϵ_2, the weak limit as $\epsilon_2 \to 0$ of the sequence $(\mathfrak{v}, \mathfrak{h}, \mathfrak{q})$ exists. We will denote the weak limit of $(\mathfrak{v}, \mathfrak{h}, \mathfrak{q})$ by $(v_\kappa, h_\kappa, q_\kappa)$. By lower semicontinuity, (8.8) and thus (8.19) hold for the weak limit $(v_\kappa, h_\kappa, q_\kappa)$. Furthermore,

$$\langle v_{\kappa t}, \varphi\rangle + \frac{\nu}{2}\int_\Omega D_{\bar{\eta}} v_\kappa : D_{\bar{\eta}}\varphi dx + \sigma\int_\Gamma \bar{\Theta}\Big[[\mathcal{L}_{\bar{h}}(h_\kappa)(-\nabla_0 \bar{h}, 1)] \circ \bar{\eta}^\tau\Big] \cdot \varphi dS$$

$$+ \kappa\int_\Gamma \Delta_0 v_\kappa \cdot \Delta_0 \varphi dS - (q_\kappa, \bar{a}^\ell_k \varphi^k_{,\ell})_{L^2(\Omega)} \tag{8.20}$$

$$= \langle F, \varphi\rangle - \sigma\int_\Gamma \bar{\Theta}\Big[[\mathcal{M}(\bar{h})(-\nabla_0 \bar{h}, 1)] \circ \bar{\eta}^\tau\Big] \cdot \varphi dS$$

for all $\varphi \in H^{1;2}(\Omega; \Gamma)$ and a.a. $t \in [0, T]$.

9. Estimates independent of κ and ϵ_1.

9.1. Energy estimates which are independent of κ. Although (8.19) does not imply that $h_\kappa \in H^4(\Gamma)$, h_κ is indeed in $H^4(\Gamma)$ by (7.4). Therefore, we have that

$(v_\kappa, h_\kappa, q_\kappa)$ satisfies

$$v_{\kappa}{}_t^i - \nu[\bar{a}_\ell^k D_{\bar\eta}(v_\kappa)_\ell^i]_{,k} = -(\bar{a}_i^k q_\kappa)_{,k} + \tilde{F}^i \quad \text{in } (0,T)\times\Omega, \tag{9.1a}$$

$$\bar{a}_i^j v_{\kappa}{}_{,j}^i = 0 \quad \text{in } (0,T)\times\Omega, \tag{9.1b}$$

$$[\nu D_{\bar\eta}(v_\kappa)_i^j - q_\kappa \delta_i^j]\bar{a}_j^\ell N_\ell = \sigma\bar\Theta[\mathcal{L}_{\bar h}(h_\kappa)(-\nabla_0 \bar h, 1)]\circ\bar\eta^\tau + \sigma\bar\Theta[\mathcal{M}_{\bar h}(-\nabla_0\bar h, 1)]\circ\bar\eta^\tau + \kappa\Delta_0^2 v_\kappa \quad \text{on } (0,T)\times\Gamma, \tag{9.1c}$$

$$h_t\circ\bar\eta^\tau = [(\bar h_{,\alpha})\circ\bar\eta^\tau]v_\alpha - v_z \quad \text{on } (0,T)\times\Gamma, \tag{9.1d}$$

$$v = \tilde{u}_0 \quad \text{on } \{t=0\}\times\Omega, \tag{9.1e}$$

$$h = 0 \quad \text{on } \{t=0\}\times\Gamma. \tag{9.1f}$$

Having (9.1c), (A.7) in Appendix A implies that h_κ is in $H^5(\Gamma)$ for a.a. $t\in[0,T]$ with estimate

$$\int_0^t \|\nabla_0^2 h_\kappa\|_{H^3(\Gamma)}^2 ds \le C(\epsilon_1)\int_0^t \Big[\|\nabla_0^4 h_\kappa\|_{L^2(\Gamma)}^2 + \|v_\kappa\|_{H^3(\Omega)}^2 + \|q_\kappa\|_{H^2(\Omega)}^2 + 1\Big]ds,$$

where the forcing f in (A.7) is given by

$$[\nu D_{\bar\eta}(v_\kappa)_i^j - q_\kappa\delta_i^j]\bar{a}_j^\ell N_\ell - \sigma\bar\Theta[\mathcal{M}_{\bar h}(-\nabla_0\bar h, 1)]\circ\bar\eta^\tau.$$

By the same argument, (7.18) holds with all θ replaced by κ. Therefore, by (8.4) (which follows from (7.18)),

$$\int_0^t \|\nabla_0^2 h_\kappa\|_{H^3(\Gamma)}^2 ds \le C(\epsilon_1)\int_0^t \Big[\|v_{\kappa t}\|_{H^1(\Omega)}^2 + \|\nabla_0^4 h_\kappa\|_{L^2(\Gamma)}^2 + \|\nabla_0^2 v_\kappa\|_{H^1(\Omega_1)}^2\Big]ds + C(\epsilon_1)N_2(u_0,F). \tag{9.2}$$

With this extra regularity of h_κ, the energy estimate (8.19) can be made independent of κ. In section B.2 in Appendix B, we prove that

$$\begin{aligned}\frac{\nu_1}{2}\|\nabla_0^4 h_\kappa(t)\|_{L^2(\Gamma)}^2 &\le \int_0^t\int_\Gamma \bar\Theta\Big[[L_{\bar h}(h_\kappa)(-\nabla_0\bar h,1)]\circ\bar\eta^\tau\Big]\cdot\nabla_0^2(\zeta_1^2\nabla_0^2 v_\kappa)dSds\\ &+ C'\int_0^t\Big[1+\|\tilde v\|_{H^3(\Omega)}^2 + \|\tilde h_t\|_{H^{2.5}(\Gamma)}^2 + \|\tilde h\|_{H^5(\Gamma)}^2\Big]\|\nabla_0^4 h_\kappa\|_{L^2(\Gamma)}^2 ds\\ &+ C'\int_0^t\Big[\|\tilde h\|_{H^5(\Gamma)}^2 + 1\Big]ds + \delta\int_0^t \|v_\kappa\|_{H^3(\Omega)}^2 ds + \delta_1\int_0^t \|\nabla_0^2 h_\kappa\|_{H^3(\Gamma)}^2 ds\end{aligned}$$

for some constant C' depending on M, ϵ_1, δ, and δ_1. By (9.2),

$$\begin{aligned}\frac{\nu_1}{2}\|\nabla_0^4 h_\kappa(t)\|_{L^2(\Gamma)}^2 &\le \int_0^t\int_\Gamma \bar\Theta\Big[[L_{\bar h}(h_\kappa)(-\nabla_0\bar h,1)]\circ\bar\eta^\tau\Big]\cdot\nabla_0^2(\zeta_1^2\nabla_0^2 v_\kappa)dSds\\ &+ C'N_2(u_0,F) + C'\int_0^t\Big[\|\nabla_0^2 v_\kappa\|_{H^1(\Omega_1)}^2 + K(s)\|\nabla_0^4 h_\kappa\|_{L^2(\Gamma)}^2\Big]ds\\ &+ \delta\int_0^t \|v_\kappa\|_{H^3(\Omega)}^2 ds + \delta_1\int_0^t \|v_{\kappa t}\|_{H^1(\Omega)}^2 ds,\end{aligned} \tag{9.3}$$

where

$$K(s) := 1 + \|\tilde v\|_{H^3(\Omega)}^2 + \|\tilde h_t\|_{H^{2.5}(\Gamma)}^2 + \|\tilde h\|_{H^5(\Gamma)}^2.$$

With (9.3), (8.11) now is replaced by

$$
\begin{aligned}
&\Big[\|\nabla_0^2 v_\kappa(t)\|^2_{L^2(\Omega_1)} + \|\nabla_0^4 h_\kappa(t)\|^2_{L^2(\Gamma)}\Big] + \int_0^t \Big[\|\nabla\nabla_0^2 v_\kappa\|^2_{L^2(\Omega_1)} + \kappa\|v_\kappa\|^2_{H^4(\Gamma)}\Big] ds \\
&\le C' N_2(u_0,F) + C'\int_0^t \Big[\|v_{\kappa t}\|^2_{L^2(\Omega)} + \|\nabla_0^2 v_\kappa\|^2_{H^1(\Omega_1)} + K(s)\|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)}\Big] ds \\
&\quad + \delta\int_0^t \|v_\kappa\|^2_{H^3(\Omega)} ds + \delta_1 \int_0^t \|v_{\kappa t}\|^2_{H^1(\Omega)} ds
\end{aligned}
\tag{9.4}
$$

for some C' depending on M, ϵ_1, δ, and δ_1, where (A.5) is applied to bound $\kappa\|v_\kappa\|^2_{H^3(\Gamma)}$ (this is where $\|v_{\kappa t}\|^2_{L^2(\Omega)}$ comes from). Similar computations lead to

$$
\begin{aligned}
&\Big[\|\nabla_0 v_\kappa(t)\|^2_{L^2(\Omega_1)} + \|\nabla_0^3 h_\kappa(t)\|^2_{L^2(\Gamma)}\Big] + \int_0^t \Big[\|\nabla\nabla_0 v_\kappa\|^2_{L^2(\Omega_1)} + \kappa\|v_\kappa\|^2_{H^3(\Gamma)}\Big] ds \\
&\le C N_2(u_0,F) + C\int_0^t \|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)} ds + \delta \int_0^t \|v_\kappa\|^2_{H^3(\Omega)} ds
\end{aligned}
\tag{9.5}
$$

for some constant C depending on M and δ.

In Appendix C, we establish the following κ- and ϵ_1-independent inequality for the time-differentiated problem:

$$
\begin{aligned}
&\int_0^t \|\nabla_0^2 h_{\kappa t}\|^2_{L^2(\Gamma)} ds \le \int_0^t \int_\Gamma \Big[[L_{\bar h}(h_\kappa)(\nabla_0 \bar h, -1)] \circ \bar\eta^\tau\Big]_t \cdot v_{\kappa t} dS \\
&+ C N_3(u_0,F) + C\int_0^t K(s)\Big[\|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)} + \|\nabla_0^2 h_{\kappa t}\|^2_{L^2(\Gamma)}\Big] ds \\
&+ (\delta + C t^{1/2}) \int_0^t \|v_\kappa\|^2_{H^3(\Omega)} ds + (\delta_1 + C t^{1/2}) \int_0^t \|v_{\kappa t}\|^2_{H^1(\Omega)} ds + \delta_2 \|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)}
\end{aligned}
$$

for some constant C depending on M, δ, δ_1, and δ_2. Therefore, (8.14) can be replaced by the following estimate:

$$
\begin{aligned}
&\Big[\|v_{\kappa t}\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_{\kappa t}\|^2_{L^2(\Gamma)}\Big] + \int_0^t \Big[\|\nabla v_{\kappa t}\|^2_{L^2(\Omega)} + \kappa\|\Delta_0 v_{\kappa t}\|^2_{L^2(\Gamma)}\Big] ds \\
&\le C N_3(u_0,F) + C\int_0^t K(s)\Big[\|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)} + \|\nabla_0^2 h_{\kappa t}\|^2_{L^2(\Gamma)}\Big] ds \\
&\quad + (\delta + C t^{1/2}) \int_0^t \|v_\kappa\|^2_{H^3(\Omega)} ds + (\delta_1 + C t^{1/2}) \int_0^t \|v_{\kappa t}\|^2_{H^1(\Omega)} ds + \delta_2 \|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)}.
\end{aligned}
\tag{9.6}
$$

9.2. κ-independent estimates. Just as in section 8.7, we find that

$$
\begin{aligned}
&\int_0^t \Big[\|v_\kappa\|^2_{H^3(\Omega)} + \|q_\kappa\|^2_{H^2(\Omega)}\Big] ds \\
&\le C(M) N_2(u_0,F) + C(M)\int_0^t \Big[\|v_{\kappa t}\|^2_{H^1(\Omega)} + \|\nabla_0^2 v_\kappa\|^2_{H^1(\Omega_1)}\Big] ds.
\end{aligned}
\tag{9.7}
$$

By choosing $\delta = \delta_1 = \delta_2 = 1/8$ and $T > 0$ so that $CT^{1/2} < 1/8$ in (9.6), we find that

$$
\begin{aligned}
&\int_0^t \Big[\|v_\kappa\|^2_{H^3(\Omega)} + \|q_\kappa\|^2_{H^2(\Omega)}\Big] ds \le CN_3(u_0,F) + \frac{1}{8}\|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)} \\
&\quad + C(M)\int_0^t \Big[\|\nabla_0^2 v_\kappa\|^2_{H^1(\Omega_1)} + K(s)\Big(\|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)} + \|\nabla_0^2 h_{\kappa t}\|^2_{L^2(\Gamma)}\Big)\Big] ds.
\end{aligned}
\tag{9.8}
$$

Combining the estimates (7.27), (8.9), (9.4), and (9.5) with (9.6),

$$
\begin{aligned}
&\Big[\|v_\kappa\|^2_{H^1(\Omega)} + \|\nabla_0^2 v_\kappa\|^2_{L^2(\Omega_1)} + \|\nabla_0^2 h_\kappa\|^2_{H^2(\Gamma)} + \|v_{\kappa t}\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_{\kappa t}\|^2_{L^2(\Gamma)}\Big](t) \\
&+ \int_0^t \Big[\|\nabla v_\kappa\|^2_{L^2(\Omega)} + \|\nabla\nabla_0 v_\kappa\|^2_{L^2(\Omega_1)} + \|\nabla\nabla_0^2 v_\kappa\|^2_{L^2(\Omega_1)} + \|v_{\kappa t}\|^2_{H^1(\Omega)}\Big] ds \\
&\le C' N_3(u_0,F) + C'\int_0^t \Big[\|v_{\kappa t}\|^2_{L^2(\Omega)} + K(s)\Big(\|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)} + \|\nabla_0^2 h_{\kappa t}\|^2_{L^2(\Gamma)}\Big)\Big] ds
\end{aligned}
$$

for some constant C' depending on M and ϵ_1. By the Gronwall inequality and (8.4),

$$
\begin{aligned}
&\sup_{0\le t\le T}\Big[\|v_\kappa\|^2_{H^2(\Omega)} + \|v_{\kappa t}\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_{\kappa t}\|^2_{L^2(\Gamma)} + \|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)} \\
&\qquad + \|q_\kappa\|^2_{H^1(\Omega)}\Big](t) + \|v_\kappa\|^2_{\mathcal{V}^3(T)} + \|q_\kappa\|^2_{L^2(0,T;H^2(\Omega))} \le C(\epsilon_1)N_3(u_0,F).
\end{aligned}
$$

9.3. Weak limits as $\boldsymbol{\kappa \to 0}$. Just as in section 8.8, the weak limit $(v_{\epsilon_1}, h_{\epsilon_1}, q_{\epsilon_1})$ of $(v_\kappa, h_\kappa, q_\kappa)$ as $\kappa \to 0$ exists in $V(T) \times L^2(0,T;H^4(\Gamma)) \times L^2(0,T;H^2(\Omega))$ with estimate

$$
\begin{aligned}
&\sup_{0\le t\le T}\Big[\|v_{\epsilon_1}(t)\|^2_{H^2(\Omega)} + \|v_{\epsilon_1 t}(t)\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_{\epsilon_1 t}(t)\|^2_{L^2(\Gamma)} + \|\nabla_0^4 h_{\epsilon_1}(t)\|^2_{L^2(\Gamma)} \\
&\qquad + \|q_{\epsilon_1}(t)\|^2_{H^1(\Omega)}\Big] + \|v_\kappa\|^2_{\mathcal{V}^3(T)} + \|q_{\epsilon_1}\|^2_{L^2(0,T;H^2(\Omega))} \le C(\epsilon_1)N_3(u_0,F).
\end{aligned}
\tag{9.9}
$$

Equation (9.9) implies that for a.a. $t \in [0,T]$,

$$\|v_\kappa(t)\|_{H^{2.5}(\Gamma)} \le \bar{C}(t)$$

for some $\bar{C}(t)$ independent of κ, and therefore for a.a. $t \in [0,T]$,

$$\kappa\int_\Gamma \Delta_0 v_\kappa \cdot \Delta_0\varphi dS \to 0$$

as $\kappa \to 0$. This observation with (8.20) shows that $(v_{\epsilon_1}, h_{\epsilon_1}, q_{\epsilon_1})$ satisfies, for a.a. $t \in [0,T]$,

$$
\begin{aligned}
&(v_{\kappa t},\varphi)_{L^2(\Omega)} + \frac{\nu}{2}\int_\Omega D_{\bar\eta} v_\kappa : D_{\bar\eta}(\varphi) dx + \sigma\int_\Gamma \bar\Theta \mathcal{L}_{\bar h}(h_\kappa)\Big[-\bar h_{,\sigma}\circ\bar\eta^\tau \varphi^\sigma + \varphi^z\Big] dS \\
&- (\bar a_i^j q_\kappa, \varphi^i_{,j})_{L^2(\Omega)} = \langle \tilde F, \varphi\rangle + \sigma\langle \bar\Theta \mathcal{M}_{\bar h}(-\nabla_0 \bar h\circ\bar\eta^\tau, 1), \varphi\rangle_\Gamma
\end{aligned}
\tag{9.10}
$$

for all $\varphi \in H^{1;2}(\Omega;\Gamma)$. Since (9.10) also defines a linear functional on $H^1(\Omega)$, by the density argument, we have that (9.10) holds for all $\varphi \in H^1(\Omega)$. As $(v_{\epsilon_1}, h_{\epsilon_1}, q_{\epsilon_1})$ are smooth enough, we can integrate by parts and find that $(v_{\epsilon_1}, h_{\epsilon_1}, q_{\epsilon_1})$ satisfies (7.2) with (7.2c) replaced by

$$
[\nu D_{\bar\eta}(v_{\epsilon_1})^j_i - q_{\epsilon_1}\delta^j_i]\bar a^\ell_j N_\ell = \sigma\Big[\bar\Theta[(\mathcal{L}_{\bar h}(h_{\epsilon_1}) + \mathcal{M}(\bar h))(\nabla_0 \bar h, -1)]\circ\bar\eta^\tau\Big] \quad \text{on } (0,T)\times\Gamma.
\tag{9.11}
$$

9.4. $H^{5.5}$-regularity of h_κ. By (9.11), we have the following lemma.

LEMMA 9.1. *For a.a. $t \in [0,T]$, $h_{\epsilon_1}(t) \in H^{5.5}(\Gamma)$ with*

$$\|h_{\epsilon_1}\|^2_{H^{5.5}(\Gamma)} \le C(M)\Big[\|v_{\epsilon_1 t}\|^2_{H^1(\Omega)} + \|\nabla v_{\epsilon_1}\|^2_{L^2(\Omega)} + \|\nabla_0^2 v_{\epsilon_1}\|^2_{H^1(\Omega_1)} + \|\nabla_0^4 h_{\epsilon_1}\|^2_{L^2(\Gamma)} + \|F\|^2_{H^1(\Omega)} + 1\Big], \tag{9.12}$$

and hence

$$\|h_{\epsilon_1}\|^2_{L^2(0,T;H^{5.5}(\Gamma))} \le C(M)e^{C(M)+T} N_3(u_0,F). \tag{9.13}$$

Proof. We write the boundary condition (9.11) as

$$\mathcal{L}_{\bar h}(h_{\epsilon_1}) = \frac{1}{\sigma} J_{\bar h}^{-2}(-\nabla_0 \bar h, 1)\cdot \Big\{\bar\Theta^{-1}\Big[[\nu D_{\bar\eta}(v_{\epsilon_1})^j_i - q_{\epsilon_1}\delta^j_i]\bar a^\ell_j N_\ell\Big]\Big\}\circ \bar\eta^{-\tau} - \mathcal{M}(\bar h). \tag{9.14}$$

By Corollary 7.1, $\mathcal{L}_{\bar h}$ is uniformly elliptic with the elliptic constant ν_1 which is independent of M which defines our convex subset $C_T(M)$. Since $\bar h \in \mathcal{H}(T)$, $\mathcal{M}(\bar h) \in L^2(0,T;H^{2.5}(\Gamma)) \cap L^\infty(0,T;H^1(\Gamma))$, and hence by (8.19), the right-hand side of (9.14) is bounded in $H^{1.5}(\Gamma)$. The important point is that these bounds are independent of ϵ_1. Thus, elliptic regularity of $\mathcal{L}_{\bar h}$ proves the estimate

$$\|h_{\epsilon_1}\|^2_{H^{5.5}(\Gamma)} \le C(M)\Big[\|D_{\bar\eta}(v_{\epsilon_1})\|^2_{H^{1.5}(\Gamma)} + \|q_{\epsilon_1}\|^2_{H^{1.5}(\Gamma)} + 1\Big]$$

so that with (8.4), (9.12) is proved. ◻

9.5. Energy estimates which are independent of ϵ_1. Having estimate (9.12), one can follow exactly the same procedure as in section 9.2 to show that the constant C' appearing in (9.9) is independent of ϵ_1, provided that we have an ϵ_1-independent version of (9.4). By section B.2, we indeed have such an estimate:

$$\begin{aligned}\frac{\nu_1}{2}\|\nabla_0^4 h_{\epsilon_1}(t)\|^2_{L^2(\Gamma)} &\le \int_0^t\int_\Gamma \bar\Theta\Big[[L_{\bar h}(h_{\epsilon_1})(-\nabla_0\bar h,1)]\circ\bar\eta^\tau\Big]\cdot\nabla_0^2(\zeta_1^2\nabla_0^2 v_{\epsilon_1})dSds\\ &+ CN_2(u_0,F) + C\int_0^t K(s)\|\nabla_0^4 h_{\epsilon_1}\|^2_{L^2(\Gamma)}ds + (\delta + Ct^{1/2})\int_0^t \|v_{\epsilon_1}\|^2_{H^3(\Omega)}ds\\ &+ (\delta_1 + Ct^{1/2})\int_0^t \|v_{\epsilon_1 t}\|^2_{H^1(\Omega)}ds\end{aligned}$$

for some constant C depending on M, δ, and δ_1. Therefore, we can conclude that

$$\begin{aligned}&\sup_{0\le t\le T}\Big[\|v_{\epsilon_1}\|^2_{H^2(\Omega)} + \|v_{\epsilon_1 t}\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_{\epsilon_1 t}\|^2_{L^2(\Gamma)} + \|\nabla_0^4 h_{\epsilon_1}\|^2_{L^2(\Gamma)}\\ &\quad + \|q_{\epsilon_1}\|^2_{H^1(\Omega)}\Big](t) + \|v_{\epsilon_1}\|^2_{\mathcal{V}^3(T)} + \|q_{\epsilon_1}\|^2_{L^2(0,T;H^2(\Omega))} \le C(M)e^{C(M)+T}N_3(u_0,F).\end{aligned} \tag{9.15}$$

REMARK 15. *Literally speaking, we cannot use $\nabla_0^2(\zeta_1^2\nabla_0^2 v_{\epsilon_1})$ as a test function in* (9.10)*, since it is not a function in $H^1(\Omega)$. However, since $h_{\epsilon_1} \in H^{5.5}(\Gamma)$ for a.a. $t \in [0,T]$,* (9.10) *also holds for all $\varphi \in H^1(\Omega)' \cap H^{-1.5}(\Gamma)$ and $\nabla_0^2(\zeta_1^2\nabla_0^2 v_{\epsilon_1})$ is a function of this kind.*

9.6. Weak limits as $\epsilon_1 \to 0$. The same argument leads to the fact that weak limits of $(v_{\epsilon_1}, h_{\epsilon_1}, q_{\epsilon_1})$ (denoted by (v,h,q)) as $\epsilon_1 \to 0$ exist and (v,h,q) satisfies (7.1).

9.7. Uniqueness. In this section, we show that for a given $(\tilde{v},\tilde{h}) \in Y_T$, the solution to (7.1) is unique in Y_T. Suppose (v_1, h_1) and (v_2, h_2) are two solutions (in Y_T) to (7.3). Let $w = v_1 - v_2$ and $g = h_1 - h_2$; then w and g satisfy

$$\begin{aligned}\langle w_t, \varphi\rangle + \frac{\nu}{2}\int_\Omega D_{\tilde{\eta}} w : D_{\tilde{\eta}}\varphi dx + \sigma\int_\Gamma \tilde{\Theta}\Big[\tilde{L}_{\tilde{h}}\Big(\int_0^t (\tilde{h}_{,\alpha} w_\alpha - w_z)ds\Big)\Big]\circ\tilde{\eta}^\tau \\ \times(-\tilde{h}_{,\alpha}\circ\tilde{\eta}^\tau\varphi^\alpha + \varphi^z)dS = 0\end{aligned} \tag{9.16}$$

for all $\varphi \in \mathcal{V}_v(T)$ with $w(0) = 0$, where $\tilde{L}$ equals L, except $L_1 = L_2 = 0$. Since w is in $\mathcal{V}_v(T)$, letting $w = \varphi$ in (9.16) leads to

$$\begin{aligned}&\Big[\|v\|^2_{H^1(\Omega)} + \|\nabla_0^2 v\|^2_{L^2(\Omega_1)} + \|\nabla_0^4 h\|^2_{L^2(\Gamma)} + \|v_t\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_t\|^2_{L^2(\Gamma)}\Big](t) \\ &\quad + \int_0^t \Big[\|\nabla v\|^2_{L^2(\Omega)} + \|\nabla\nabla_0 v\|^2_{L^2(\Omega_1)} + \|\nabla\nabla_0^2 v\|^2_{L^2(\Omega_1)} + \|v_t\|^2_{H^1(\Omega)}\Big]ds \\ &\leq C(M)\int_0^t K(s)\Big[\|\nabla_0^4 h\|^2_{L^2(\Gamma)} + \|\nabla_0^2 h_t\|^2_{L^2(\Gamma)}\Big]ds.\end{aligned}$$

Therefore, by the Gronwall inequality and the zero initial condition ($w(0) = 0$), we have that w (and hence g) is identical to zero.

10. Fixed-point argument. From previous sections, we establish a map Θ_T from Y_T into Y_T; i.e., given $(\tilde{v},\tilde{h}) \in C_T(M)$, there exists a unique $\Theta_T(\tilde{v},\tilde{h}) = (v, h)$ satisfying (7.1). Theorem 4.1 is then proved if this mapping Θ_T has a fixed point. We shall make use of the Tychonoff fixed-point theorem which states as follows.

THEOREM 10.1. *For a reflexive Banach space X, and $C \subset X$ a closed, convex, bounded subset, if $F : C \to C$ is weakly sequentially continuous into X, then F has at least one fixed point.*

In order to apply the Tychonoff fixed-point theorem, we need to show that $\Theta(\tilde{v},\tilde{h}) \in C_T(M)$, and this is the case if T is small enough. In the following discussion, we will always assume T is smaller than a fixed constant (for example, say $T \leq 1$) so that the right-hand side of (9.15) can be written as $C(M)N_3(u_0, F)$.

REMARK 16. *The space Y_T is not reflexive. We will treat $C_T(M)$ as a convex subset of X_T and apply the Tychonoff fixed-point theorem on the space X_T.*

Before proceeding with the fixed-point proof, we note that Lemma 6.3 implies that for a short time, the constant $C(M)$ in (8.1) and (8.4) can be chosen to be independent of M. To be more precise, for a.a. $0 < t \leq T_1$,

$$\|q\|^2_{L^2(\Omega)} \leq C\Big[\|v_t\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)} + \|\nabla_0^4 h\|^2_{L^2(\Gamma)} + \|F\|^2_{L^2(\Omega)} + 1\Big], \tag{10.1}$$

$$\begin{aligned}\|v\|^2_{H^3(\Omega)} + \|q\|^2_{H^2(\Omega)} \leq C\Big[&\|v_t\|^2_{H^1(\Omega)} + \|\nabla v\|^2_{H^1(\Omega)} + \|\nabla_0 v\|^2_{H^1(\Omega_1)} \\ &+ \|\nabla_0^2 v\|^2_{H^1(\Omega)} + \|F\|^2_{H^1(\Omega)} + 1\Big],\end{aligned} \tag{10.2}$$

and

$$\begin{aligned}\|h\|^2_{H^{5.5}(\Gamma)} \leq C\Big[&\|v_t\|^2_{H^1(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)} + \|\nabla_0^2 v\|^2_{H^1(\Omega_1)} + \|\nabla_0^4 h\|^2_{L^2(\Gamma)} \\ &+ \|F\|^2_{H^1(\Omega)} + 1\Big]\end{aligned} \tag{10.3}$$

for some constant C independent of M.

10.1. Continuity in time of h. By the evolution equation (7.1d) and the fact that $v \in \mathcal{V}^3(T_1)$, $h_t \in L^2(0,T_1;H^{2.5}(\Gamma))$. Since $h \in L^2(0,T_1;H^{5.5}(\Gamma))$, we have that $h \in \mathcal{C}^0([0,T_1];H^4(\Gamma))$ by the standard interpolation theorem. Although there is no uniform rate that h converges to zero in $H^4(\Gamma)$, we have the following lemma.

LEMMA 10.2. *Let $(v,h) = \Theta_{T_1}(\tilde{v},\tilde{h})$. Then $\|h(t)\|_{H^{2.5}(\Gamma)}$ converges to zero as $t \to 0$, uniformly for all $(\tilde{v},\tilde{h}) \in C_{T_1}(M)$.*

Proof. By the evolution equation (7.1d),

$$\|h(t)\|_{H^{2.5}(\Gamma)} \leq \int_0^t \|\tilde{h}_{,\alpha}v_\alpha - v_z\|_{H^{2.5}(\Gamma)}dS \leq C(M)N_3(u_0,F)^{1/2}t^{1/2}.$$

The lemma follows directly from the inequality. ∎

By Lemma 10.2 and the interpolation inequality, we also have the following lemma.

LEMMA 10.3. *$\|\nabla_0^2 h(t)\|_{H^{1.5}(\Gamma)}$ converges to zero as $t \to 0$, uniformly for all $\tilde{h} \in C_{T_1}(M)$ with estimate*

$$\|\nabla_0^2 h(t)\|_{H^{1.5}(\Gamma)} \leq C(M)N_3(u_0,F)t^{1/4} \tag{10.4}$$

for all $0 < t \leq T_1$.

10.2. Improved energy estimates. In order to apply the fixed-point theorem, we have to use the fact that the forcing F is in $\mathcal{V}^2(T)$. We also define a new constant

$$N(u_0,F) := \|u_0\|^2_{H^{2.5}(\Omega)} + \|F\|^2_{\mathcal{V}^2(T_1)} + \|F\|^2_{L^\infty(0,T_1;L^2(\Omega))} + \|F(0)\|^2_{H^1(\Omega)} + 1.$$

Note that $N_3(u_0,F) \leq N(u_0,F)$.

REMARK 17. *For the linearized problem* (7.1), *we need only $F \in \mathcal{V}^1(T)$ to obtain a unique solution $(v,h) \in Y_T$.*

10.2.1. Estimates for $\nabla_0^2 v$ near the boundary. Note that

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\left[\|\zeta_1\nabla_0^2 v\|^2_{L^2(\Omega)} + \sigma\int_\Gamma \tilde{\Theta}B\tilde{A}^{\alpha\beta\gamma\delta}\nabla_0^2 h_{,\alpha\beta}\nabla_0^2 h_{,\gamma\delta}dS\right] + \frac{\nu}{2}\|\zeta_1 D_{\tilde{\eta}}(\nabla_0^2 v)\|^2_{L^2(\Omega)}\\
&= \langle F, \nabla_0^2(\zeta_1^2\nabla_0^2 v)\rangle - \frac{\nu}{4}\int_\Omega\left[\nabla_0^2(\tilde{a}_i^k\tilde{a}_i^\ell)v^j_{,\ell} + \nabla_0^2(\tilde{a}_i^k\tilde{a}_j^\ell)v^i_{,\ell}\right](\zeta_1^2\nabla_0^2 v^j)_{,k}dx\\
&\quad - \frac{\nu}{2}\int_\Omega\left[\nabla_0(\tilde{a}_i^k\tilde{a}_i^\ell)\nabla_0 v^j_{,\ell} + \nabla_0(\tilde{a}_i^k\tilde{a}_j^\ell)\nabla_0 v^i_{,\ell}\right](\zeta_1^2\nabla_0^2 v^j)_{,k}dx\\
&\quad - \frac{\nu}{2}\int_\Omega D_{\tilde{\eta}}(\nabla_0^2 v)^j_i\tilde{a}_i^k\zeta_1\zeta_{1,k}\nabla_0^2 v^j dx + \int_\Omega q\tilde{a}_k^\ell[\nabla_0^2(\zeta_1^2\nabla_0^2 v^k)]_{,\ell}dx - \sigma\left(\sum_{k=1}^3 I_k + \sum_{k=1}^8 J_k\right),
\end{aligned}$$

where I_k's and J_k's are defined in section B.1 (with $\bar{}$ replaced by $\tilde{}$, and no ϵ_1 and ϵ_2).

As in [7] and [8], we study the time integral of the right-hand side of the identity above in order to prove the validity of the requirement of applying the Tychonoff fixed-point theorem. By interpolation and (9.9),

$$\begin{aligned}
&\int_0^t\int_\Omega\left[\nabla_0^2(\tilde{a}_i^k\tilde{a}_i^\ell)v^j_{,\ell} + \nabla_0^2(\tilde{a}_i^k\tilde{a}_j^\ell)v^i_{,\ell}\right](\zeta_1^2\nabla_0^2 v^j)_{,k}dxds\\
&\leq C\int_0^t \|\tilde{a}\tilde{a}\|_{H^2(\Omega)}\|\nabla v\|_{L^\infty(\Omega)}\|v\|_{H^3(\Omega)}ds
\end{aligned}$$

$$
\begin{aligned}
&\leq C(M)C(\delta)\int_0^t \|v\|_{H^3(\Omega)}^{1/2}\|v\|_{H^1(\Omega)}^{1/2}ds+\delta\|v\|^2_{L^2(0,T;H^3(\Omega))}\\
&\leq C(M)C(\delta)N(u_0,F)^{1/2}\int_0^t \|v\|_{H^3(\Omega)}^{1/2}ds+\delta C(M)N(u_0,F)\\
&\leq C(M)N(u_0,F)\Big[C(\delta)t^{3/4}+\delta\Big].
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\int_0^t\int_\Omega \Big[\nabla_0(\tilde a_i^k\tilde a_i^\ell)\nabla_0 v^j_{,\ell}+\nabla_0(\tilde a_i^k\tilde a_j^\ell)\nabla_0 v^i_{,\ell}\Big](\zeta_1^2\nabla_0^2 v^j)_{,k}dxds\\
&+\int_0^t\int_\Omega D_{\tilde\eta}(\nabla_0^2 v)_i^j\tilde a_i^k\zeta_1\zeta_{1,k}\nabla_0^2 v^j dxds\leq C(M)N(u_0,F)\Big[t^{1/2}+C(\delta)t+\delta\Big].
\end{aligned}
$$

For the pressure term, by interpolation and (8.10),

$$
\begin{aligned}
&\int_0^t\int_\Omega q\tilde a_k^\ell[\nabla_0^2(\zeta_1^2\nabla_0^2 v^k)]_{,\ell}dxds\\
&\leq C(M)\int_0^t\Big[\|q\|_{L^\infty(\Omega)}+\|q\|_{W^{1,4}(\Omega)}+\|q\|_{H^1(\Omega)}\Big]\|v\|_{H^3(\Omega)}ds\\
&\leq C(M)C(\delta)\int_0^t\|q\|^2_{H^1(\Omega)}ds+\delta\Big[\|v\|^2_{L^2(0,T;H^3(\Omega)}+\|q\|^2_{L^2(0,T;H^2(\Omega))}\Big]\\
&\leq C(M)N(u_0,F)\Big[C(\delta)t^{1/2}+\delta\Big].
\end{aligned}
$$

By the estimates already established in Appendix B, with the help of (6.6), it is also easy to see that

$$
\int_0^t\left(\sum_{k=1}^3 I_k+\sum_{k=1}^8 J_k\right)ds\leq C(M)N(u_0,F)\Big[t^{1/4}+t^{1/2}+C(\delta)t^{2/3}+\delta\Big].
$$

Finally, for the forcing term, by the extra regularity we assume on F,

$$
\begin{aligned}
\int_0^t\langle F,\nabla_0^2(\zeta_1^2\nabla_0^2 v)\rangle ds&\leq\int_0^t\|F\|_{H^2(\Omega)}\|v\|_{H^2(\Omega)}ds\leq N(u_0,F)+\int_0^t\|v\|^2_{H^2(\Omega)}ds\\
&\leq N(u_0,F)+C(M)N(u_0,F)t.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\Big[\|\nabla_0^2 v(t)\|^2_{L^2(\Omega_1)}+\sigma E_{\tilde h}(\nabla_0^2 h)\Big]+\nu\int_0^t\|D_{\tilde\eta}(\nabla_0^2 v)\|^2_{L^2(\Omega_1)}ds\\
&\leq\|u_0\|^2_{H^2(\Omega)}+CN(u_0,F)+C(M)N(u_0,F)\Big[C(\delta)(t^{3/4}+t^{2/3}+t^{1/2}+t)+\delta\Big].
\end{aligned}
$$

By Corollary 7.1,

$$
\begin{aligned}
&\Big[\|\nabla_0^2 v(t)\|^2_{L^2(\Omega_1)}+\|\nabla_0^4 h(t)\|^2_{L^2(\Gamma)}\Big]+\int_0^t\|\nabla_0^2 v\|^2_{H^1(\Omega_1)}ds\\
&\leq CN(u_0,F)+C(M)N(u_0,F)\Big[C(\delta)\mathcal{O}(t)+\delta\Big]\quad\text{as}\quad t\to 0,
\end{aligned}
\tag{10.5}
$$

where C depends on ν, σ, ν_1, and the geometry of Γ.

By similar computations, we can also conclude (the (7.27), (8.9), and (9.5) variants) that

$$\begin{aligned}
&\left[\|v(t)\|^2_{L^2(\Omega)} + \|\nabla_0^2 h(t)\|^2_{L^2(\Gamma)}\right] + \int_0^t \|v\|^2_{H^1(\Omega)} ds \\
&\leq CN(u_0,F) + C(M)N(u_0,F)\mathcal{O}(t) \qquad \text{as} \quad t\to 0; \qquad (10.6)\\
&\left[\|\nabla_0 v(t)\|^2_{L^2(\Omega_1)} + \|\nabla_0^3 h(t)|^2_{L^2(\Gamma)}\right] + \int_0^t \|\nabla_0 v\|^2_{H^1(\Omega_1)} ds \\
&\leq CN(u_0,F) + C(M)N(u_0,F)\mathcal{O}(t) \qquad \text{as} \quad t\to 0; \qquad (10.7)\\
&\|\nabla v(t)\|^2_{L^2(\Omega)} + \int_0^t \|v_t\|^2_{L^2(\Omega)} ds \\
&\leq CN(u_0,F) + C(M)N(u_0,F)\mathcal{O}(t) \qquad \text{as} \quad t\to 0, \qquad (10.8)
\end{aligned}$$

where C depends on ν, σ, ν_1, and the geometry of Γ.

10.2.2. $L^2_t H^1_x$-estimate for v_t. For the time-differentiated problem, we are not able to use estimates such as those in sections 8.6 and 10.2.1, since no ϵ_1-regularization is present; nevertheless, we can obtain estimates at the ϵ_1-regularization level and then pass ϵ_1 to the limit once the estimate is found to be ϵ_1-independent. We have that

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|v_t\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|D_{\bar\eta} v_t\|^2_{L^2(\Omega)} + \frac{\sigma}{2}\frac{d}{dt}\int_\Gamma \bar\Theta \bar A^{\alpha\beta\gamma\delta} h_{t,\alpha\beta} h_{t,\gamma\delta} dS \\
&= \langle F_t, v_t\rangle - \nu \int_\Omega \left[(\bar a^k_i \bar a^\ell_j)_t v^j_{,\ell} + (\bar a^k_i \bar a^\ell_j)_t v^i_{,\ell}\right] v^j_{t,k} dx + \int_\Omega q_t \bar a^\ell_{kt} v^k_{,\ell} dx \\
&\quad + \frac{1}{2}\int_\Gamma (\bar\Theta \bar A^{\alpha\beta\gamma\delta})_t h_{t,\alpha\beta} h_{t,\gamma\delta} dS - \int_\Gamma \frac{\bar\Theta}{\sqrt{\det(g_0)}} \left[\sqrt{\det(g_0)} (\bar A^{\alpha\beta\gamma\delta})_t h_{,\alpha\beta}\right]_{,\gamma\delta} h_{tt} dS \\
&\quad - 2\int_\Gamma \bar\Theta_{,\gamma} \bar A^{\alpha\beta\gamma\delta} h_{t,\alpha\beta} h_{tt,\delta} dS - \int_\Gamma \bar\Theta_{,\gamma\delta} \bar A^{\alpha\beta\gamma\delta} h_{t,\alpha\beta} h_{tt} dS \\
&\quad - \int_\Gamma \bar\Theta \left[L_1^{\alpha\beta\gamma} \bar h_{,\alpha\beta\gamma}\right]_t h_{tt} dS - \int_\Gamma \bar\Theta (L_2)_t h_{tt} dS + K_1 + K_3 + K_4 + K_5 + K_6,
\end{aligned}$$

where K_i's are defined in Appendix C (without ϵ_2).

As in the previous section, the time integral of the right-hand side of the identity above is studied. It is easy to see that

$$\begin{aligned}
&\int_0^t \left[\langle F_t, v_t\rangle - \nu\Big((\bar a^k_i \bar a^\ell_j)_t v^j_{,\ell} + (\bar a^k_i \bar a^\ell_j)_t v^i_{,\ell}\Big) v^j_{t,k} + K_1 + K_5 + K_6\right] ds \\
&\leq C(M)N(u_0,F)\left[t^{1/4} + t^{1/2} + C(\delta)(t^{1/2} + t) + \delta\right],
\end{aligned}$$

and by Appendix C, particularly Remark 22,

$$\begin{aligned}
&\int_0^t \int_\Gamma \left[\frac{1}{2}(\bar\Theta \bar A^{\alpha\beta\gamma\delta})_t h_{t,\alpha\beta} h_{t,\gamma\delta} - \frac{\bar\Theta}{\sqrt{\det(g_0)}}\left[\sqrt{\det(g_0)} (\bar A^{\alpha\beta\gamma\delta})_t h_{,\alpha\beta}\right]_{,\gamma\delta} h_{tt}\right. \\
&\qquad \left. - 2\bar\Theta_{,\gamma} \bar A^{\alpha\beta\gamma\delta} h_{t,\alpha\beta} h_{tt,\delta} - \bar\Theta_{,\gamma\delta} \bar A^{\alpha\beta\gamma\delta} h_{t,\alpha\beta} h_{tt}\right] dS ds \\
&\leq C(M)N(u_0,F) t^{1/2}.
\end{aligned}$$

Special treatment is needed for the rest of the terms, and we break this procedure into several steps.

Step 1. Let $B_1 = \int_0^t \int_\Omega (q\bar{a}_k^\ell)_t v^k_{t,\ell} dxds$. By the "divergence-free" condition (7.2b),

$$B_1 = \int_0^t \int_\Omega \bar{a}^\ell_{kt} q v^k_{t,\ell} dxds - \int_0^t \int_\Omega \bar{a}^\ell_{kt} q_t v^k_{,\ell} dxds.$$

By interpolation and (8.1),

$$\begin{aligned}
&\left| \int_0^t \int_\Omega \bar{a}^\ell_{kt} q v^k_{t,\ell} dxds \right| \\
&\le C(M)C(\delta) \int_0^t \|q\|^2_{L^2(\Omega)} ds + \delta \Big[\|q\|^2_{L^2(0,T;H^1(\Omega))} + \|v_t\|^2_{L^2(0,T;H^1(\Omega))} \Big] \\
&\le C(M)N(u_0,F)\Big[C(\delta)t + \delta\Big].
\end{aligned}$$

For the second integral, we have the following identity:

$$\begin{aligned}
\int_0^t \int_\Omega \bar{a}^\ell_{kt} q_t v^k_{,\ell} dxds = \int_\Omega (\bar{a}^\ell_{kt} q v^k_{,\ell})(t) dx - \int_\Omega \bar{a}^\ell_{kt}(0) q(0) u^k_{0,\ell} dx \\
- \int_0^t \int_\Omega (\bar{a}^\ell_{kt} v^k_{,\ell})_t q dxds.
\end{aligned}$$

By the identity $\bar{a}^\ell_{kt} = -\bar{a}^i_k \bar{v}^j_{,i} \bar{a}^\ell_j$,

$$\begin{aligned}
\left| \int_0^t \int_\Omega (\bar{a}^\ell_{kt} v^k_{,\ell})_t q dxds \right| &\le \int_0^t \int_\Omega \left| \Big[\bar{a}^\ell_{ktt} v^k_{,\ell} + \bar{a}^\ell_{kt} v^k_{t,\ell} \Big] q \right| dxds \\
&\le C(M) \int_0^t (1 + \|\bar{v}_t\|_{H^1(\Omega)}) \|\nabla v\|_{L^4(\Omega)} \|q\|_{L^4(\Omega)} ds.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left| \int_0^t \int_\Omega (\bar{a}^\ell_{kt} v^k_{,\ell})_t q dxds \right| \\
&\le C(M)C(\delta)N(u_0,F) \int_0^t \|q\|^{2\alpha}_{H^1(\Omega)} \|q\|^{2(1-\alpha)}_{L^2(\Omega)} ds + \delta \int_0^t (1 + \|\bar{v}_t\|^2_{H^1(\Omega)}) ds \\
&\le C(M)N(u_0,F)^2 \Big[C(\delta)(t + t^{\frac{1-\alpha}{2}}) + \delta \Big],
\end{aligned}$$

where $\alpha = \frac{3}{4}$ if $n = 3$ and $\alpha = \frac{1}{2}$ if $n = 2$.

The second integral equals $\int_\Omega \nabla u_0 : (\nabla u_0)^T q(0) dx$, which is bounded by $CN(u_0,F)$. It remains to estimate the first integral. By adding and subtracting $\int_\Omega \bar{a}^\ell_{kt}(0) q v^k_{,\ell} dx$, we find, by $\bar{a}_t(0) \in H^2(\Omega)$, that

$$\begin{aligned}
\left| \int_\Omega (\bar{a}^\ell_{kt} q v^k_{,\ell})(t) dx \right| &\le \int_\Omega \left| (\bar{a}^\ell_{kt} - \bar{a}^\ell_{kt}(0))(q v^k_{,\ell})(t) \right| dx + \int_\Omega \left| \bar{a}^\ell_{kt}(0) q v^k_{,\ell} \right| dx \\
&\le C \|\bar{a}_t(t) - \bar{a}_t(0)\|_{L^4(\Omega)} \|q\|_{L^2(\Omega)} \|\nabla v\|_{L^4(\Omega)} \\
&\quad + C(\delta_1) \|\nabla v\|^2_{L^2(\Omega)} + \delta_1 \|q\|^2_{L^2(\Omega)}.
\end{aligned}$$

Noting that

$$\|\nabla v\|^2_{L^2(\Omega)} = \left\|\nabla u_0 + \int_0^t \nabla v_t ds\right\|^2_{L^2(\Omega)} \le \left[\|\nabla u_0\|_{L^2(\Omega)} + \int_0^t \|\nabla v_t\|_{L^2(\Omega)} ds\right]^2$$
$$\le 2\Big[\|u_0\|^2_{H^1(\Omega)} + C(M)N(u_0,F)t\Big],$$

(9.9), (6.5c), and (10.1) imply that

$$\left|\int_\Omega \bar{a}^\ell_{kt} q v^k_{,\ell}(t)dx\right| \le C(M)N(u_0,F)t^{1/2} + C(\delta_1)N(u_0,F)$$
$$+ \delta_1\Big[\|v_t\|^2_{L^2(\Omega)} + \|\nabla_0^4 h\|^2_{L^2(\Gamma)}\Big].$$

Summing all the estimates above, we find that

$$|B_1| \le C(\delta_1)N(u_0,F) + C(M)N(u_0,F)^2\Big[C(\delta)(t + t^{\frac{1-\alpha}{2}}) + \delta\Big]$$
$$+ \delta_1\Big[\|v_t\|^2_{L^2(\Omega)} + \|\nabla_0^4 h\|^2_{L^2(\Gamma)}\Big].$$

REMARK 18. *It may be tempting to use an interpolation inequality to show that $q \in \mathcal{C}([0,T];X)$ for some Banach space X by analyzing q_t via Laplace's equation. The problem, however, is that the boundary condition for q_t has low regularity $L^2(0,T;H^{-1.5}(\Gamma))$ (by the fact that $h_t \in L^2(0,T;H^{2.5}(\Gamma))$), and thus standard elliptic estimates do not provide the desired conclusion that $q_t \in L^2(0,T;H^1(\Omega)')$ (and hence by interpolation, $q \in \mathcal{C}([0,T];H^{0.5}(\Omega))$). However, suppose that $q_t \in L^2(0,T;H^1(\Omega)')$; then we can estimate $\int_0^t\int_\Omega \bar{a}^\ell_{kt} q_t v^k_{,\ell} dxds$ by the following method:*

$$\left|\int_0^t\int_\Omega \bar{a}^\ell_{kt} q_t v^k_{,\ell} dxds\right| \le \int_0^t \|\bar{a}^i_k \bar{v}^j_{,i} \bar{a}^\ell_j v^k_{,\ell}\|_{H^1(\Omega)} \|q_t\|_{H^1(\Omega)'} ds$$
$$\le C(M)N(u_0,F)\Big[t + t^{5/8}\Big].$$

Step 2. Let $B_2 = \int_0^t\int_\Gamma \tilde{\Theta}\big[[L_1^{\alpha\beta\gamma}\tilde{h}_{,\alpha\beta\gamma}]_t h_{tt} + (L_2)_t h_{tt}\big]dSds$. It is easy to see that

$$\left|\int_0^t\int_\Gamma \bar{\Theta}(L_2)_t h_{tt} dSds\right| \le C(M)\int_0^t \Big[\|v\|_{L^\infty(\Gamma)} + \|v_t\|_{L^2(\Gamma)}\Big]ds$$
$$\le C(M)N(u_0,F)^{1/2}(t + t^{3/4}).$$

For parts involving L_1, we have

$$\int_0^t\int_\Gamma \tilde{\Theta}\Big[L_1^{\alpha\beta\gamma}\tilde{h}_{,\alpha\beta\gamma}\Big]_t h_{tt} dSds = \int_0^t\int_\Gamma \tilde{\Theta}\Big[L_1^{\alpha\beta\gamma}\Big]_t \bar{h}_{,\alpha\beta\gamma} h_{tt} dSds \quad (\equiv B_2^1)$$
$$+ \int_0^t\int_\Gamma \tilde{\Theta} L_1^{\alpha\beta\gamma} \bar{h}_{t,\alpha\beta\gamma} h_{tt} dSds \quad (\equiv B_2^2).$$

By interpolation,

$$|B_2^1| \le C(M)\int_0^t \|\bar{\Theta}\|_{L^\infty(\Gamma)} \|\tilde{h}\|_{W^{1,4}(\Gamma)} \|h_{tt}\|_{L^4(\Gamma)} dSds$$
$$\le C(M)\int_0^t \Big[\|v\|_{H^2(\Omega)} + \|v_t\|_{H^1(\Omega)}\Big]ds$$
$$\le C(M)N(u_0,F)^{1/2}t^{1/2},$$

while by (6.6) and Corollary 6.5,

$$
\begin{aligned}
|B_2^2| &\le \int_0^t \|\bar\Theta\|_{H^{1.5}(\Gamma)} \|\tilde h_t\|_{H^{2.5}(\Gamma)} \|L_1^{\alpha\beta\gamma}\|_{H^{1.5}(\Gamma)} \|h_{tt}\|_{H^{0.5}(\Gamma)} ds \\
&\le C(M) \|L_1^{\alpha\beta\gamma}\|_{H^{1.5}(\Gamma)} \int_0^t \|\tilde h\|_{H^{2.5}(\Gamma)} \Big[\|v\|_{H^2(\Omega)} + \|v_t\|_{H^1(\Omega)} \Big] ds \\
&\le C(M) N(u_0, F) t^{1/4}.
\end{aligned}
$$

Therefore,

$$
|B_2| \le C(M) N(u_0, F)(t + t^{3/4} + t^{1/4}).
$$

Step 3. Let $B_3 = \int_0^t K_3 ds = \int_0^t \int_\Gamma \bar\Theta [L_{\bar h}(h)]_t [(\bar v \circ \bar\eta^{-\tau}) \cdot (\nabla_0 h_t)] dS ds$. The L_1 and L_2 part of B_3 is bounded by

$$
C(M) \int_0^t \|\bar\Theta\|_{H^{1.5}(\Gamma)} \|\bar v\|_{H^{1.5}(\Gamma)} \|\bar h\|_{H^{3.5}(\Gamma)} \|\bar h_t\|_{H^2(\Gamma)} \|h_t\|_{H^2(\Omega)} ds,
$$

and hence

$$
\left| \int_0^t \bar\Theta \Big[L_1^{\alpha\beta\gamma} \bar h_{,\alpha\beta\gamma} + L_2 \Big]_t [(\bar v \circ \bar\eta^{-\tau}) \cdot (\nabla_0 h_t)] dS ds \right| \le C(M) N(u_0, F) t^{1/4}.
$$

By the integration by parts formula, the highest order part of B_3 can be expressed as

$$
\begin{aligned}
&\int_0^t \int_\Gamma \frac{\bar\Theta(\bar v \circ \bar\eta^{-\tau})}{\sqrt{\det(g_0)}} \Big[\sqrt{\det(g_0)} (\bar A^{\alpha\beta\gamma\delta})_t h_{,\alpha\beta} \Big]_{,\gamma\delta} \nabla_0 h_t dS ds \quad (\equiv B_3^1) \\
&+ \int_0^t \int_\Gamma \bar\Theta(\bar v \circ \bar\eta^{-\tau}) \bar A^{\alpha\beta\gamma\delta} h_{t,\alpha\beta} \nabla_0 h_{t,\gamma\delta} dS ds \quad (\equiv B_3^2) \\
&+ 2 \int_0^t \int_\Gamma [\bar\Theta(\bar v \circ \bar\eta^{-\tau})]_{,\gamma} \bar A^{\alpha\beta\gamma\delta} h_{t,\alpha\beta} \nabla_0 h_{t,\delta} dS ds \quad (\equiv B_3^3) \\
&+ \int_0^t \int_\Gamma [\bar\Theta(\bar v \circ \bar\eta^{-\tau})]_{,\gamma\delta} \bar A^{\alpha\beta\gamma\delta} h_{t,\alpha\beta} \nabla_0 h_t dS ds \quad (\equiv B_3^4).
\end{aligned}
$$

It is easy to see that

$$
\begin{aligned}
|B_3^1| &\le C(M) \int_0^t \|\bar\Theta \bar v \circ \bar\eta^{-\tau}\|_{H^{1.5}(\Gamma)} \|\bar h_t\|_{H^2(\Gamma)} \|h\|_{H^4(\Gamma)} \|h_t\|_{H^2(\Gamma)} dS \\
&\le C(M) N(u_0, F) t
\end{aligned}
$$

and

$$
\begin{aligned}
|B_3^3| &\le C(M) \int_0^t \|\bar\Theta \bar v \circ \bar\eta^{-\tau}\|_{W^{1,4}(\Gamma)} \|\bar A\|_{L^\infty(\Gamma)} \|h_t\|_{H^2(\Gamma)} \|h_t\|_{W^{2,4}(\Gamma)} dS \\
&\le C(M) N(u_0, F) t^{1/2}.
\end{aligned}
$$

For B_3^2, by the integration by parts formula,

$$
\begin{aligned}
B_3^2 &= \frac{1}{2} \int_0^t \int_\Gamma \bar\Theta(\bar v \circ \bar\eta^{-\tau}) \bar A^{\alpha\beta\gamma\delta} \nabla_0 \Big[h_{t,\alpha\beta} h_{t,\gamma\delta} \Big] dS ds \\
&= -\frac{1}{2} \int_0^t \int_\Gamma \frac{1}{\sqrt{\det(g_0)}} \nabla_0 \Big[\sqrt{\det(g_0)} \bar\Theta(\bar v \circ \bar\eta^{-\tau}) \bar A^{\alpha\beta\gamma\delta} \Big] h_{t,\alpha\beta} h_{t,\gamma\delta} dS ds,
\end{aligned}
$$

and hence

$$
\begin{aligned}
|B_3^2| &\leq \int_0^t \Big[\|\nabla_0 \bar{\Theta}\|_{L^4(\Gamma)} \|\bar{v}\bar{A}\|_{L^\infty(\Gamma)} + \|\bar{\Theta}\|_{L^\infty(\Gamma)} \|\bar{v}\bar{A}\|_{W^{1,4}(\Gamma)}\Big] \\
&\qquad \times \|h_t\|_{W^{2,4}(\Gamma)} \|h_t\|_{H^2(\Gamma)} ds \\
&\leq C(M) N(u_0, F)^{1/2} \int_0^t \|v\|_{H^3(\Omega)} ds \\
&\leq C(M) N(u_0, F) t^{1/2}.
\end{aligned}
$$

For B_3^4, noting that

$$
\begin{aligned}
\bar{\Theta}_{,\gamma\delta} &= \det(\nabla_0 \bar{\eta}^\tau)_{,\gamma\delta} \sqrt{\det(G_{\bar{h}}) \circ \bar{\eta}^\tau} + \det(\nabla_0 \bar{\eta}^\tau)_{,\gamma} \sqrt{\det(G_{\bar{h}}) \circ \bar{\eta}^\tau}_{,\delta} \\
&\quad + \det(\nabla_0 \bar{\eta}^\tau)_{,\delta} \sqrt{\det(G_{\bar{h}}) \circ \bar{\eta}^\tau}_{,\gamma} + \det(\nabla_0 \bar{\eta}^\tau) \sqrt{\det(G_{\bar{h}}) \circ \bar{\eta}^\tau}_{,\gamma\delta}
\end{aligned}
$$

and $\|\nabla_0 \det(\nabla_0 \bar{\eta}^\tau)\|_{H^{0.5}(\Gamma)} \leq C(M) t^{1/2}$, we find that

$$
\begin{aligned}
|B_3^4| &\leq C(M) \int_0^t \|\nabla_0 \det(\nabla_0 \bar{\eta}^\tau)\|_{H^{0.5}(\Gamma)} \|\nabla_0^2 h_t\|_{H^{0.5}(\Gamma)} \|\nabla_0 h_t\|_{H^{1.5}(\Gamma)} ds \\
&\quad + C(M) \int_0^t \|\det(\nabla_0 \bar{\eta}^\tau)\|_{L^\infty(\Gamma)} \|\nabla_0 \bar{\eta}^\tau\|^2_{L^\infty(\Gamma)} \|\nabla_0^2 h_t\|_{L^2(\Gamma)} \|\nabla_0 h_t\|_{L^2(\Gamma)} ds \\
&\leq C(M) N(u_0, F) t^{1/2} + C(M) N(u_0, F)^{3/4} \int_0^t \|v\|^{1/2}_{H^3(\Omega)} ds \\
&\leq C(M) N(u_0, F)(t^{1/2} + t^{3/4}).
\end{aligned}
$$

Combining all the estimates, we find that

$$
|B_3| \leq C(M) N(u_0, F)(t + t^{1/2} + t^{3/4}).
$$

Step 4. Let $B_4 = \int_0^t K_4 ds = \int_0^t \int_\Gamma \bar{\Theta}[L_{\bar{h}}(h)]_t [(\nabla_0 \bar{h}, -1)_t \cdot (v \circ \bar{\eta}^{-\tau})] dS ds$. Integrating by parts,

$$
\begin{aligned}
B_4 &= -\int_0^t \int_\Gamma L_{\bar{h}}(h) \Big[\bar{\Theta}_t (\nabla_0 \bar{h}, -1)_t \cdot (v \circ \bar{\eta}^{-\tau}) + \bar{\Theta}(\nabla_0 \bar{h}, -1)_t \cdot (v \circ \bar{\eta}^{-\tau})_t \\
&\quad + \bar{\Theta}(\nabla_0 \bar{h}, -1)_{tt} \cdot (v \circ \bar{\eta}^{-\tau})\Big] dS ds + \int_\Gamma \bar{\Theta} L_{\tilde{h}}(h) [(\nabla_0 \tilde{h}, -1)_t \cdot (v \circ \bar{\eta}^{-\tau})] dS.
\end{aligned}
$$

For the first integral, (6.8) implies that

$$
\begin{aligned}
&\left| \int_\Gamma \bar{\Theta} L_{\tilde{h}}(h) [(\nabla_0 \tilde{h}, -1)_t \cdot (v \circ \bar{\eta}^{-\tau})] dS \right| \\
&\leq \|\bar{\Theta}\|_{L^\infty(\Gamma)} \|L_{\tilde{h}}(h)\|_{L^2(\Gamma)} \|\nabla_0 \tilde{h}_t\|_{L^4(\Gamma)} \|v \circ \bar{\eta}^{-\tau}\|_{L^4(\Gamma)} \\
&\leq C(M) N(u_0, F) \|\tilde{h}_t\|_{H^{1.5}(\Gamma)} \\
&\leq C(M) N(u_0, F) t^{1/8}.
\end{aligned}
$$

It is also easy to see that

$$\begin{aligned}&\left|\int_0^t\int_\Gamma L_{\bar h}(h)\Big[\bar\Theta_t(\nabla_0\bar h,-1)_t\cdot(v\circ\bar\eta^{-\tau})+\bar\Theta(\nabla_0\tilde h,-1)_t\cdot(v\circ\bar\eta^{-\tau})_t\Big]dSds\right|\\&\le C(M)\int_0^t\Big[\|v\|_{L^\infty(\Gamma)}+\|v_t\|_{L^4(\Gamma)}\Big]\|L_{\tilde h}(h)\|_{L^2(\Gamma)}\|\nabla_0\tilde h_t\|_{L^4(\Gamma)}ds\\&\le C(M)N(u_0,F)^{1/2}\int_0^t\Big[\|v\|_{H^3(\Omega)}+\|v_t\|_{H^1(\Omega)}\Big]ds\\&\le C(M)N(u_0,F)t^{1/2}.\end{aligned}$$

For the remaining terms, the $H^{0.5}(\Gamma)$-$H^{-0.5}(\Gamma)$ duality pairing leads to

$$\begin{aligned}&\left|\int_0^t\int_\Gamma\bar\Theta L_{\tilde h}(h)(\nabla_0\tilde h,-1)_{tt}\cdot vdSds\right|\\&\le\int_0^t\|\bar\Theta\|_{H^{1.5}(\Gamma)}\|L_{\tilde h}(h)\|_{H^{0.5}(\Gamma)}\|v\|_{H^{1.5}(\Gamma)}\|\tilde h_{tt}\|_{H^{0.5}(\Gamma)}ds.\end{aligned}$$

By interpolation,

$$\|L_{\tilde h}(h)\|_{H^{0.5}(\Gamma)}\le C(M)\Big[\|h\|^{1/2}_{H^{5.5}(\Gamma)}\|h\|^{1/2}_{H^{3.5}(\Gamma)}+1\Big],$$

and hence

$$\begin{aligned}&\left|\int_0^t\int_\Gamma\tilde\Theta L_{\tilde h}(h)(\nabla_0\tilde h,-1)_{tt}\cdot(v\circ\bar\eta^{-\tau})dSds\right|\\&\le C(M)N(u_0,F)\int_0^t\|\tilde h_{tt}\|_{H^{0.5}(\Gamma)}\Big[\|\nabla_0^5h\|^{1/2}_{L^2(\Gamma)}+1\Big]ds\\&\le C(M)C(\delta)N(u_0,F)\int_0^t\Big[\|\nabla_0^5h\|_{L^2(\Gamma)}+1\Big]ds+\delta C(M)N(u_0,F)\\&\le C(M)N(u_0,F)\Big[C(\delta)(t^{1/2}+t)+\delta\Big].\end{aligned}$$

All the inequalities above give us

$$|B_4|\le C(M)N(u_0,F)\Big[C(\delta)(t^{1/2}+t)+t^{1/8}+\delta\Big].$$

Summing all the estimates above, we find that

$$\begin{aligned}&\Big[\|v_t\|^2_{L^2(\Omega)}+\sigma\int_\Gamma\bar\Theta\bar A^{\alpha\beta\gamma\delta}h_{t,\alpha\beta}h_{t,\gamma\delta}|^2dS\Big](t)+\nu\int_0^t\|D_{\tilde\eta}v_t\|^2_{L^2(\Omega)}ds\\&\le\|v_t(0)\|^2_{L^2(\Omega)}+\sigma\int_\Gamma|G_0^{\alpha\beta}h_{t,\alpha\beta}(0)|^2dS+(C+C(\delta_1))N(u_0,F)\\&\quad+C(M)N(u_0,F)\Big[C(\delta)(t+t^{3/4}+t^{1/2}+t^{1/4}+t^{1/8}+t^{\frac{1-\alpha}{2}})+\delta\Big]\\&\quad+\delta_1\Big[\|v_t\|^2_{L^2(\Omega)}+\|\nabla_0^4h\|^2_{L^2(\Gamma)}\Big],\end{aligned}$$

and by Corollary 7.1,

$$
\begin{aligned}
&\Big[\|v_t(t)\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_t(t)\|^2_{L^2(\Gamma)}\Big] + \int_0^t \|v_t\|^2_{H^1(\Omega)} ds \\
&\le (C + C(\delta_1))N(u_0, F) + C(M)N(u_0, F)\Big[C(\delta)\mathcal{O}(t) + \delta\Big] \\
&\quad + \delta_1 \Big[\|v_t\|^2_{L^2(\Omega)} + \|\nabla_0^4 h\|^2_{L^2(\Gamma)}\Big],
\end{aligned}
\tag{10.9}
$$

where C depends on ν, σ, ν_1, and the geometry of Γ. Since this estimate is independent of ϵ_1, we pass ϵ_1 to zero and conclude that the solution (v, h) to (7.1) also satisfies (10.9).

10.3. Mapping from $C_T(M)$ into $C_T(M)$. In this section, we are going to choose M so that $\Theta(\tilde{v}, \tilde{h}) \in C_T(M)$ if $(\tilde{v}, \tilde{h}) \in C_T(M)$.

Summing (10.5), (10.6), (10.7), (10.8), and (10.9), by (6.5) we find that

$$
\begin{aligned}
&\Big[\|v(t)\|^2_{L^2(\Omega)} + \|\nabla_0 v(t)\|^2_{L^2(\Omega_1)} + \|\nabla_0^2 v(t)\|^2_{L^2(\Omega_1)} + \|v_t(t)\|^2_{L^2(\Omega)} \\
&+ \|\nabla_0^2 h(t)\|^2_{L^2(\Gamma)} + \|\nabla_0^3 h(t)\|^2_{L^2(\Gamma)} + \|\nabla_0^4 h(t)\|^2_{L^2(\Gamma)} + \|\nabla_0^2 h_t(t)\|^2_{L^2(\Gamma)}\Big] \\
&+ \int_0^t \Big[\|v\|^2_{H^1(\Omega)} + \|\nabla_0 v\|^2_{H^1(\Omega_1)} + \|\nabla_0^2 v\|^2_{H^1(\Omega_1)} + \|v_t\|^2_{H^1(\Omega)}\Big] ds \\
&\le (C + C(\delta_1))N(u_0, F) + C(M)N(u_0, F)\Big[C(\delta)\mathcal{O}(t) + \delta\Big] \\
&\quad + \delta_1 \Big[\|v_t\|^2_{L^2(\Omega)} + \|\nabla_0^4 h\|^2_{L^2(\Gamma)}\Big],
\end{aligned}
$$

where C depends on ν, σ, ν_1, and the geometry of Γ. Choosing $\delta_1 = \frac{1}{2}$,

$$
\begin{aligned}
&\Big[\|v(t)\|^2_{L^2(\Omega)} + \|\nabla_0 v(t)\|^2_{L^2(\Omega_1)} + \|\nabla_0^2 v(t)\|^2_{L^2(\Omega_1)} + \|v_t(t)\|^2_{L^2(\Omega)} \\
&+ \|\nabla_0^2 h(t)\|^2_{L^2(\Gamma)} + \|\nabla_0^3 h(t)\|^2_{L^2(\Gamma)} + \|\nabla_0^4 h(t)\|^2_{L^2(\Gamma)} + \|\nabla_0^2 h_t(t)\|^2_{L^2(\Gamma)}\Big] \\
&+ \int_0^t \Big[\|v\|^2_{H^1(\Omega)} + \|\nabla_0 v\|^2_{H^1(\Omega_1)} + \|\nabla_0^2 v\|^2_{H^1(\Omega_1)} + \|v_t\|^2_{H^1(\Omega)}\Big] ds \\
&\le C_1 N(u_0, F) + C(M)N(u_0, F)^2\Big[C(\delta)\mathcal{O}(t) + \delta\Big],
\end{aligned}
$$

where C_1 depends on ν, σ, μ, and the geometry of Γ. Similar to section 8.7, for a.a. $0 < t \le T$,

$$
\begin{aligned}
&\Big[\|v(t)\|^2_{H^2(\Omega)} + \|v_t(t)\|^2_{L^2(\Omega)} + \|\nabla_0^2 h(t)\|^2_{H^2(\Gamma)} + \|\nabla_0^2 h_t(t)\|^2_{L^2(\Gamma)}\Big] \\
&+ \int_0^t \Big[\|v\|^2_{H^3(\Omega)} + \|v_t\|^2_{H^1(\Omega)} + \|q\|^2_{H^2(\Omega)}\Big] ds \\
&\le C_2 N(u_0, F) + C(M)N(u_0, F)^2\Big[C(\delta)\mathcal{O}(t) + \delta\Big]
\end{aligned}
\tag{10.10}
$$

for some constant C_2 depending on C_1.

By (6.6), (6.8), and (7.1d),

$$
\begin{aligned}
\int_0^t \|h_t\|^2_{H^{2.5}(\Gamma)} ds &\le \int_0^t \Big[1 + \|\tilde{h}\|^2_{H^{3.5}(\Gamma)}\Big] \|v\|^2_{H^{2.5}(\Gamma)} ds \\
&\le C(M)N(u_0, F)t^{1/4}
\end{aligned}
\tag{10.11}
$$

and

$$\begin{aligned}\int_0^t \|h_{tt}\|^2_{H^{0.5}(\Gamma)} ds &\le C(M)\int_0^t \Big[\|\tilde{h}_t\|^2_{H^{1.5}(\Gamma)}\|v\|^2_{H^2(\Omega)} + \|\tilde{h}\|^2_{H^{2.5}(\Gamma)}\|v_t\|^2_{H^1(\Omega)}\Big] ds \\ &\le C(M)N(u_0,F)\Big[t^{1/4}+t^{1/2}\Big]. \end{aligned} \tag{10.12}$$

Also, by (10.3) and (10.10),

$$\begin{aligned}\int_0^t \|h\|^2_{H^{5.5}(\Gamma)} ds &\le C\int_0^t \Big[\|v_t\|^2_{H^1(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)} + \|\nabla_0^2 v\|^2_{H^1(\Omega_1)} + \|\nabla_0^4 h\|^2_{L^2(\Gamma)} \\ &\qquad + \|F\|^2_{H^1(\Omega)} + 1\Big] ds \\ &\le C_3 N(u_0,F) + C(M)N(u_0,F)^2\Big[C(\delta)\mathcal{O}(t)+\delta\Big] \end{aligned} \tag{10.13}$$

for some constant C_3 depending on C_2.

Combining (10.10), (10.11), (10.12), and (10.13), we have the following inequality:

$$\begin{aligned}&\Big[\|v(t)\|^2_{H^2(\Omega)} + \|v_t(t)\|^2_{L^2(\Omega)} + \|h(t)\|^2_{H^4(\Gamma)} + \|h_t(t)\|^2_{H^2(\Gamma)}\Big] \\ &+ \int_0^t \Big[\|v\|^2_{H^3(\Omega)} + \|v_t\|^2_{H^1(\Omega)} + \|h\|^2_{H^{5.5}(\Gamma)} + \|h_t\|^2_{H^{2.5}(\Gamma)} + \|h_{tt}\|^2_{H^{0.5}(\Gamma)}\Big] ds \\ &\le (C_2+C_3)N(u_0,F) + C(M)N(u_0,F)^2\Big[C(\delta)\mathcal{O}(t)+\delta\Big].\end{aligned}$$

Let $M = 2(C_2+C_3)N(u_0,F)+1$ (and hence corresponding T_0 and T in Lemma 6.3 and Corollary 7.1 are fixed). Choose $\delta > 0$ small enough (but fixed) so that

$$C(M)N(u_0,F)^2\delta \le \frac{1}{4}$$

and then choose $T > 0$ small enough so that

$$C(M)N(u_0,F)^2 C(\delta) T \le \frac{1}{4}.$$

Then for a.a. $0 < t \le T$,

$$\begin{aligned}&\Big[\|v(t)\|^2_{H^2(\Omega)} + \|v_t(t)\|^2_{L^2(\Omega)} + \|h(t)\|^2_{H^4(\Gamma)} + \|h_t(t)\|^2_{H^2(\Gamma)}\Big] \\ &+ \int_0^t \Big[\|v\|^2_{H^3(\Omega)} + \|v_t\|^2_{H^1(\Omega)} + \|h_t\|^2_{H^{2.5}(\Gamma)} + \|h_{tt}\|^2_{H^{0.5}(\Gamma)}\Big] ds \\ &\le C_2 N(u_0,F) + \frac{1}{2},\end{aligned}$$

and therefore

$$\begin{aligned}&\sup_{0\le t\le T}\Big[\|v(t)\|^2_{H^2(\Omega)} + \|v_t(t)\|^2_{L^2(\Omega)} + \|h(t)\|^2_{H^4(\Gamma)} + \|h_t(t)\|^2_{H^2(\Gamma)}\Big] \\ &+ \|v\|^2_{\mathcal{V}^3(T)} + \|h\|^2_{\mathcal{H}(T)} \le 2C_2 N(u_0,F) + 1,\end{aligned} \tag{10.14}$$

or in other words,

$$\|(v,h)\|^2_{Y(T)} \le 2C_2 N(u_0,F) + 1.$$

REMARK 19. *Equation* (10.14) *implies that for* $(\tilde{v},\tilde{h}) \in C_T(M)$ *(with* M *and* T *chosen as above), the corresponding solution to the linear problem* (7.1) $(v,h) = \Theta_T(\tilde{v},\tilde{h})$ *is also in* $C_T(M)$.

10.4. Weak continuity of the mapping Θ_T.

LEMMA 10.4. *The mapping Θ_T is weakly sequentially continuous from $C_T(M)$ into $C_T(M)$ (endowed with the norm of X_T).*

Proof. Let $(v_p, h_p)_{p\in\mathbb{N}}$ be a given sequence of elements of $C_T(M)$ weakly convergent (in Y_T) toward a given element $(v,h) \in C_T(M)$ (where $C_T(M)$ is sequentially weakly closed as a closed convex set) and let $(v_{\sigma(p)}, h_{\sigma(p)})_{p\in\mathbb{N}}$ be any subsequence of this sequence.

Since $\mathcal{V}^3(T)$ is compactly embedded into $L^2(0,T;H^2(\Omega))$, we deduce the following strong convergence results in $L^2(0,T;L^2(\Omega))$ as $p \to \infty$:

$$(a_\ell^j)_p(a_\ell^k)_p \to a_\ell^j a_\ell^k \quad \text{and} \quad (a_\ell^j)_p(a_k^\ell)_p \to a_\ell^j a_k^\ell, \tag{10.15a}$$

$$[(a_\ell^j)_p(a_\ell^k)_p]_{,j} \to (a_\ell^j a_\ell^k)_{,j} \quad \text{and} \quad [(a_\ell^j)_p(a_k^\ell)_p]_{,j} \to (a_\ell^j a_k^\ell)_{,j}, \tag{10.15b}$$

$$(a_i^k)_p \to a_i^k. \tag{10.15c}$$

Now let $(w_p, g_p) = \Theta_T(v_p, h_p)$ and let q_p be the associated pressure so that $(q_p)_{p\in\mathbb{N}}$ is in a bounded set of $\mathcal{V}^2(T)$. Since X_T is a reflexive Hilbert space, let $(w_{\sigma(p)}, g_{\sigma(p)}, q_{\sigma(p)})_{p\in\mathbb{N}}$ be a subsequence weakly converging in $X_T \times \mathcal{V}^2(T)$ toward an element $(w,g,q) \in X_T \times \mathcal{V}^2(T)$. Since $C_T(M)$ is weakly closed in X_T, we also have $(w,g) \in C_T(M)$.

For each $\phi \in L^2(0,T;H^1(\Omega))$, we deduce from (7.3) (and Remark 6) that

$$\int_0^T \left[(w_t, \phi)_{L^2(\Omega)} + \frac{\mu}{2}\int_\Omega D_\eta w : D_\eta \phi dx + \sigma \int_\Gamma L_h(g)(g_{,\alpha}\phi_\alpha - \phi_z) dS \right.$$
$$\left. + \int_\Omega q a_i^j \phi^i_{,j} dx \right] dt = \int_0^T \langle F, \phi \rangle dt,$$

which with the fact that, from (10.15), for all $t \in [0,T]$, $w \in \mathcal{V}_v$, provides that (w,g) is a solution of (2.16) in $C_T(M)$, i.e., $(w,g) = \Theta_T(v,h)$.

Therefore, we deduce that the whole sequence $(\Theta_T(v_n,h_n))_{n\in\mathbb{N}}$ weakly converges in $C_T(M)$ toward $\Theta_T(v,h)$, which concludes the lemma. ∎

10.5. Uniqueness. For the uniqueness result, we assume that u_0, F, and Γ are smooth enough (e.g., $u_0 \in H^{5.5}(\Omega)$, $F \in \mathcal{V}^4(T)$, Γ is a $H^{8.5}$ surface) so that u_0 and the associated u_1, q_0 satisfy compatibility condition (4.4). Therefore, the solution (v,h,q) is such that $v \in \mathcal{V}^6(T)$, $q \in L^2(0,T;H^5(\Omega))$ and $h \in L^\infty(0,T;H^7(\Gamma)) \cap L^2(0,T;H^{8.5}(\Gamma))$, $h_t \in L^\infty(0,T;H^5(\Gamma)) \cap L^2(0,T;H^{5.5}(\Gamma))$, $h_{tt} \in L^\infty(0,T;H^2(\Gamma)) \cap L^2(0,T;H^{3.5}(\Gamma))$. This implies $a \in L^\infty(0,T;H^5(\Omega))$, and hence by studying the elliptic equation

$$(a_i^\ell a_i^k q_{t,k})_{,\ell} = \left[\nu a_i^\ell (a_p^k a_p^j v^i_{,j})_{,k\ell} + a^\ell_{it} v^i_{,\ell} + a_i^\ell F_{,\ell}\right]_t - [(a_i^\ell a_i^k)_t q_{,k}]_{,\ell} \qquad \text{in} \quad \Omega,$$
$$q_t = J_h^{-2}\left[\left(\sigma L_h(h) N_i - \nu D_\eta(v)_i^\ell a_i^j N_j\right)_t - (a_i^j N_j)_t q\right] a_i^\ell N_\ell \quad \text{on} \quad \Gamma,$$

we find that $q_t \in L^2(0,T;H^2(\Omega))$, and this implies $v_{tt} \in L^2(0,T;H^1(\Omega))$. By the interpolation theorem, we also conclude that $v_t \in \mathcal{C}^0([0,T];H^{2.5}(\Omega))$.

Suppose (v,h,q) and $(\tilde{v},\tilde{h},\tilde{q})$ are two sets of solutions of (1.1). Then

$$(v-\tilde{v})_t - \nu[a_\ell^k D_\eta(v-\tilde{v})^i_\ell]_{,k} = -a_i^k (q-\tilde{q})_{,k} + \delta F, \tag{10.16a}$$

$$a_i^j (v-\tilde{v})^i_{,j} = \delta a, \tag{10.16b}$$

$$\left[\nu[D_\eta(v-\tilde{v})]_i^\ell - (q-\tilde{q})\delta_i^\ell\right] a_\ell^j N_j = \sigma\Theta\left[L_h(h-\tilde{h})(-\nabla_0 h, 1)\right] \circ \eta^\tau + \delta L_1 + \delta L_2 + \delta L_3, \tag{10.16c}$$

$$(h-\tilde h)_t\circ\eta^\tau = [h_{,\alpha}\circ\eta^\tau](v_\alpha-\tilde v_\alpha)-(v_z-\tilde v_z) + \delta h_1+\delta h_2+\delta h_3, \tag{10.16d}$$

$$(v-\tilde v)(0)=0, \tag{10.16e}$$

$$(h-\tilde h)(0)=0, \tag{10.16f}$$

where

$$\delta F = f\circ\eta - f\circ\tilde\eta + \nu[(a^k_\ell a^j_\ell - \tilde a^k_\ell \tilde a^j_\ell)\tilde v^i_{,j}]_{,k} + \nu[(a^k_\ell a^j_i - \tilde a^k_\ell\tilde a^j_i)\tilde v^\ell_{,j}]_{,k} - (a^k_i-\tilde a^k_i)\tilde q_{,k}, \tag{10.17a}$$

$$\delta a = (a^j_i-\tilde a^j_i)\tilde v^i_{,j}, \tag{10.17b}$$

$$\delta L_1 = \sigma\Theta\Big[L_h(\tilde h)(\nabla_0 h-\nabla_0\tilde h,0)\Big]\circ\eta^\tau - \nu(a^k_i a^j_\ell - \tilde a^k_i\tilde a^j_\ell)\tilde v^\ell_{,k}N_j - \nu(a^k_\ell a^j_\ell - \tilde a^k_\ell\tilde a^j_\ell)\tilde v^i_{,k}N_j + (a^j_i-\tilde a^j_i)\tilde q N_j, \tag{10.17c}$$

$$\delta L_2 = \tilde\Theta[L_{\tilde h}(\tilde h)\circ\eta^\tau](\nabla_0\tilde h\circ\eta^\tau-\nabla_0\tilde h\circ\tilde\eta^\tau,0) + \Big[\Theta L_h(\tilde h)\circ\eta^\tau - \tilde\Theta L_h(\tilde h)\circ\tilde\eta^\tau\Big](\nabla_0\tilde h\circ\tilde\eta^\tau,-1), \tag{10.17d}$$

$$\delta L_3 = \Big[[L_h(\tilde h)-L_{\tilde h}(\tilde h)](\nabla_0\tilde h,-1)\Big]\circ\tilde\eta^\tau, \tag{10.17e}$$

$$\delta h_1 = (h_{,\alpha}\circ\eta^\tau - h_{,\alpha}\circ\tilde\eta^\tau)\tilde v_\alpha, \tag{10.17f}$$

$$\delta h_2 = \Big[(h_{,\alpha}-\tilde h_{,\alpha})\circ\tilde\eta^\tau\Big]\tilde v_\alpha, \tag{10.17g}$$

$$\delta h_3 = -(\tilde h_t\circ\eta^\tau - \tilde h_t\circ\tilde\eta^\tau). \tag{10.17h}$$

We will also use δL and δh to denote $\sum_{k=1}^3 L_k$ and $\sum_{k=1}^3\delta h_k$, respectively.

Similar to (11.3) in [8], we also have the following estimates.

LEMMA 10.5. *For $f\in H^2(\Omega)$ and $g\in H^{1.5}(\Gamma)$,*

$$\|f\circ\eta - f\circ\tilde\eta\|_{L^2(\Omega)} \le C\sqrt{t}\|f\|_{H^2(\Omega)}\left[\int_0^t \|v-\tilde v\|^2_{H^1(\Omega)}ds\right]^{1/2}, \tag{10.18}$$

$$\|g\circ\eta^\tau - g\circ\tilde\eta^\tau\|_{L^2(\Gamma)} \le C\sqrt{t}\|g\|_{H^{1.5}(\Gamma)}\left[\int_0^t \|v-\tilde v\|^2_{H^1(\Omega)}ds\right]^{1/2} \tag{10.19}$$

for some constant C.

REMARK 20. *Assuming the regularity of h, h_t, and h_{tt} given in the beginning of this section, we have*

$$\|\delta L_2\|_{H^2(\Gamma)} + \|\delta h_1+\delta h_3\|_{H^{2.5}(\Gamma)} \le C\sqrt{t}\left[\int_0^t \|v-\tilde v\|^2_{H^3(\Omega)}ds\right]^{1/2} \tag{10.20}$$

and

$$\|(\delta L_2)_t\|_{L^2(\Gamma)} + \|(\delta h_1+\delta h_3)_t\|_{H^1(\Gamma)} \le C\left[\|v-\tilde v\|_{H^1(\Omega)} + \sqrt{t}\left(\int_0^t \|v-\tilde v\|^2_{H^2(\Omega)}ds\right)^{1/2}\right] \tag{10.21}$$

and

$$\|\nabla_0^2(\delta h_3)_t\|_{L^2(\Gamma)} \le C\Bigg[\|v-\tilde v\|_{H^1(\Omega)} + \|v-\tilde v\|_{H^3(\Omega)}$$

$$+ \sqrt{t}\|\tilde h_{tt}\|_{H^{3.5}(\Gamma)}\left(\int_0^t \|v-\tilde v\|^2_{H^3(\Omega)}ds\right)^{1/2}\Bigg]. \tag{10.22}$$

By using (10.18) to estimate $\|\delta F\|_{L^2(\Omega)}$, we find that

$$\|\nabla(v-\tilde v)(t)\|^2_{L^2(\Omega)} + \int_0^t \|(v-\tilde v)_t\|^2_{L^2(\Omega)}ds$$
$$\le C(\delta)\int_0^t \left[\|v-\tilde v\|^2_{H^1(\Omega)} + \|h-\tilde h\|^2_{H^4(\Gamma)}\right]ds + (C(\delta)t^2+\delta)\int_0^t \|v-\tilde v\|^2_{H^2(\Omega)}ds$$
$$+\delta\int_0^t \left[\|(v-\tilde v)_t\|^2_{H^1(\Omega)} + \|q-\tilde q\|^2_{H^1(\Omega)}\right]ds. \tag{10.23}$$

For the $L^2_tH^3_x$-estimate for $v-\tilde v$ and the $L^2_tH^1_x$-estimate for $(v-\tilde v)_t$, we have

$$\frac12\frac{d}{dt}\left[\|\zeta_1\nabla_0^2(v-\tilde v)\|^2_{L^2(\Omega)} + 2\sigma E_h(\nabla_0^2(h-\tilde h))\right] + \frac{\nu}{4}\|\zeta_1 D_{\bar\eta}\nabla_0^2(v-\tilde v)\|^2_{L^2(\Omega)}$$
$$\le C\Big[\|\delta F\|^2_{H^1(\Omega)} + \|(v-\tilde v)_t\|^2_{L^2(\Omega)} + \|\nabla(v-\tilde v)\|^2_{L^2(\Omega)} + \|\nabla\nabla_0(v-\tilde v)\|^2_{L^2(\Omega_1')}$$
$$+\|\nabla_0^4(h-\tilde h)\|^2_{L^2(\Gamma)}\Big] + \delta\|v-\tilde v\|^2_{H^3(\Omega)} + D_1 + D_2 + D_3$$

and

$$\frac12\frac{d}{dt}\left[\|(v-\tilde v)_t\|^2_{L^2(\Omega)} + 2\sigma E_h((h-\tilde h)_t)\right] + \frac{\nu}{4}\|\nabla(v-\tilde v)_t\|^2_{L^2(\Omega)}$$
$$\le C\Big[(\|\nabla_0^4(h-\tilde h)\|^2_{L^2(\Gamma)} + \|\nabla_0^2(h-\tilde h)_t\|^2_{L^2(\Gamma)}) + \|\delta F_t\|^2_{H^1(\Omega)'}\Big] + \delta\|v-\tilde v\|^2_{H^3(\Omega)}$$
$$+E_1+E_2+E_3,$$

where

$$D_1 := \int_\Omega \zeta_1^2\nabla_0^2(q-\tilde q)\nabla_0^2\delta a dx, \qquad D_2 := \int_\Gamma \Theta\Big[[L_h(h-\tilde h)]\circ\eta^\tau\Big](\nabla_0^4\delta h)dS,$$
$$D_3 := \int_\Gamma \delta L\cdot\nabla_0^4(v-\tilde v)dS$$

and

$$E_1 := \int_\Omega (q-\tilde q)_t(\delta a)_t dx, \qquad E_2 := \int_\Gamma \Big[\Theta[L_h(h-\tilde h)]\circ\eta^\tau\Big]_t(\delta h)_t dS,$$
$$E_3 := \int_\Gamma (\delta L)_t\cdot(v-\tilde v)_t dS.$$

By using (10.20) to estimate D_i and (10.21), (10.22) to estimate E_i, we obtain

$$\left[\|\nabla_0^2(v-\tilde v)(t)\|^2_{L^2(\Omega_1)} + \|\nabla_0^4(h-\tilde h)(t)\|^2_{L^2(\Gamma)}\right] + \int_0^t \|\nabla\nabla_0^2(v-\tilde v)\|^2_{L^2(\Omega_1)}ds$$
$$\le C(\delta)\int_0^t \left[\|(v-\tilde v)_t\|^2_{L^2(\Omega)} + \|\nabla_0(v-\tilde v)\|^2_{L^2(\Omega)} + \|\nabla_0^4(h-\tilde h)\|^2_{L^4(\Gamma)}\right]ds$$
$$+(C(\delta)t^2+\delta)\int_0^t \|v-\tilde v\|^2_{H^3(\Omega)}ds + \delta\int_0^t \|q-\tilde q\|^2_{H^2(\Omega)}ds \tag{10.24}$$

and

$$
\begin{aligned}
&\Big[\|(v-\tilde v)_t(t)\|^2_{L^2(\Omega)} + \|\nabla_0^2(h-\tilde h)_t\|^2_{L^2(\Gamma)}\Big] + \int_0^t \|\nabla(v-\tilde v)_t\|^2_{L^2(\Omega)}ds \\
&\le C(\delta)\int_0^t \Big[\|v-\tilde v\|^2_{H^1(\Omega)} + \|\nabla_0^4(h-\tilde h)\|^2_{L^2(\Gamma)} + (1+\|\tilde h_{tt}\|^2_{H^{4.5}(\Gamma)}) \\
&\qquad \times \|\nabla_0^2(h-\tilde h)_t\|^2_{L^2(\Gamma)}\Big]ds \\
&\quad + (C(\delta)(t+t^2)+\delta)\int_0^t \|v-\tilde v\|^2_{H^3(\Omega)}ds + \delta\|q-\tilde q\|^2_{L^2(\Omega)} \\
&\quad + \delta\int_0^t \Big[\|(v-\tilde v)_t\|^2_{H^1(\Omega)} + \|q-\tilde q\|^2_{H^2(\Omega)}\Big]ds.
\end{aligned}
\tag{10.25}
$$

Summing (10.23), (10.24), and (10.25), we find that

$$
Y(t) + \int_0^t Z(s)ds \le C(\delta)\int_0^t k(s)Y(s)ds + (C(\delta)(t^2+t)+\delta)\int_0^t Z(s)ds, \tag{10.26}
$$

where

$$
\begin{aligned}
k(t) &= 1+\|\tilde h_{tt}(t)\|^2_{H^{3.5}(\Gamma)}, \\
Y(t) &= \Big[\|v-\tilde v(t)\|^2_{H^1(\Omega)} + \|\nabla_0^2(v-\tilde v)(t)\|^2_{L^2(\Omega_1)} + \|(v-\tilde v)_t(t)\|^2_{L^2(\Omega)} \\
&\quad + \|h-\tilde h\|^2_{H^4(\Gamma)} + \|(h-\tilde h)_t\|^2_{H^2(\Gamma)}\Big], \\
Z(t) &= \|(v-\tilde v)_t(t)\|^2_{H^1(\Omega)} + \|\nabla\nabla_0^2(v-\tilde v)(t)\|^2_{L^2(\Omega_1)}.
\end{aligned}
$$

By letting $\delta = 1/4$ and choosing $T_u \le T$ so that $C(\delta)(T_u^2+T_u) \le 1/4$,

$$
Y(t) + \int_0^t Z(s)ds \le C\int_0^t k(s)Y(s)ds \tag{10.27}
$$

for all $0 < t \le T_u$. Since $Y(0) = 0$, the uniqueness of the solution follows from that $Y(t) = 0$ for all $0 < t \le T_u$.

11. The analysis of the membrane traction.

The analysis of the membrane traction consists of four parts: (1) the modified linearized (and regularized) problem; (2) the κ-independent estimates; (3) the fixed-point argument; and (4) the uniqueness of the solution.

11.1. The modified linearized and regularized problem.

Recall that the membrane traction is

$$
\mathfrak{t}_{\rm mem} = \Big[\mathcal{J}\mathcal{P}''(\mathcal{J}) + 2\mathcal{P}'(\mathcal{J})\Big]\mathcal{J}_{,\beta}g^{\alpha\beta}\eta_{,\alpha} + \Big[\mathcal{J}\mathcal{P}'(\mathcal{J}) + \mathcal{P}(\mathcal{J})\Big]Hn.
$$

For given $\bar v = \rho_{\epsilon_1} * \tilde v$ (and hence $\bar\eta$, $\bar g$, etc.), we define (for fixed but small $\epsilon > 0$)

$$
L^{\epsilon}_{\bar m} = \frac{1}{2}\bar{\mathcal{J}}^{-1}\Big[(\partial_\beta\rho_\epsilon) * \Big(\frac{\bar g}{g_0}\Big)\Big]\Big[\bar{\mathcal{J}}\mathcal{P}''(\bar{\mathcal{J}}) + 2\mathcal{P}'(\bar{\mathcal{J}})\Big]\bar g^{\alpha\beta}\bar\eta_{,\alpha} + \Big[\bar{\mathcal{J}}\mathcal{P}'(\bar{\mathcal{J}}) + \mathcal{P}(\bar{\mathcal{J}})\Big]\bar H\bar n.
$$

For the linearized problem, we change the boundary condition (7.1c) to

$$
\begin{aligned}
[\nu D_{\tilde\eta}(v)^j_i - q\delta^j_i]\tilde a^\ell_j N_\ell = (L^\epsilon_{\bar m})^i + \sigma\tilde\Theta\Big[\mathcal{L}_{\tilde h}(h)(-\nabla_0\tilde h, 1)\Big]\circ\tilde\eta^\tau \quad \text{on} \quad (0,T)\times\Gamma \\
+ \sigma\tilde\Theta\Big[[\mathcal{M}(\tilde h)(-\nabla_0\tilde h, 1)]\circ\tilde\eta^\tau\Big],
\end{aligned}
\tag{11.1}
$$

where we recall that $\bar{\Theta} = \det(\nabla_0 \bar{\eta}^\tau)\sqrt{\det(G_{\bar{h}}) \circ \bar{\eta}^\tau}$. Note that here we treat the membrane traction as a given forcing on the boundary. The regularized problem consists of adding the artificial viscosity, as introduced in (7.2c), in (11.1). Note that here we also mollify $\bar{\mathcal{J}}_{,\beta}$ and use the equality $(\rho_\epsilon * f)_{,\beta} = \rho_{\epsilon,\beta} * f$.

Since $L_{\bar{m}}$ is given as a forcing, all the estimates are essentially the same as those in the previous sections. Therefore, we have a unique solution (v_κ, h_κ) to the regularized problem (with ϵ_1-, ϵ-, and κ-dependent estimates).

11.2. The κ-independent estimates. The introduction of the artificial viscosity is to provide enough regularity for the solution to the linearized problem. As in Appendix A, the κ-independent estimates are obtained by studying the normal component of (A.1). Note that with the help of the mollification operation in (11.1), the corresponding f in (A.1) is also a function in $L^2(0,T;H^{1.5}(\Gamma))$. Therefore, (A.7) is still valid. This κ-independent estimate will enable us to take the limit as $\kappa \to 0$ and obtain the solution $(v_{\epsilon_1}, h_{\epsilon_1})$. Essentially the same proof as in section 9.4 shows that (9.12) still holds, and hence taking the limit as $\epsilon_1 \to 0$, the weak limit (v_ϵ, h_ϵ) solves the linearized problem (7.1), and all the estimates in the previous sections hold with $C(M)$ replaced by $C(M,\epsilon)$.

REMARK 21. *The estimate for (v_ϵ, h_ϵ) still depends on ϵ, where the extra ϵ-regularization is used in the $L^2_t H^3_x$-estimates, which requires estimating the following boundary integral:*

$$\int_\Gamma \frac{1}{2}\bar{\mathcal{J}}^{-1}\left[(\partial_\beta \rho_\epsilon) * \left(\frac{\bar{g}}{g_0}\right)\right]\left[\bar{\mathcal{J}}\mathcal{P}''(\bar{\mathcal{J}}) + 2\mathcal{P}'(\bar{\mathcal{J}})\right]\bar{g}^{\alpha\beta}\bar{\eta}_{,\alpha}\nabla_0^4 v dS.$$

Moreover, even though the estimate for h_{ϵ_1} depends only on the normal component of $L_{\bar{m}}$, in the linearized problem, there are still contributions to the normal direction made by $\bar{g}^{\alpha\beta}\bar{\eta}_{,\alpha}$.

11.3. The fixed-point argument. Similar fixed-point arguments as in section 10 guarantee the existence of a fixed point (which is still denoted by (v_ϵ, h_ϵ)) in the space X_{T_ϵ}; that is, there is a fixed point $(v_\epsilon, h_\epsilon) \in \mathcal{V}^3(T_\epsilon) \times \mathcal{H}(T_\epsilon)$. This fixed point satisfies the boundary condition

$$\begin{aligned}[\nu D_{\eta_\epsilon}(v_\epsilon)^j_i - q_\epsilon \delta^j_i](a_\epsilon)^\ell_j N_\ell = (L^\epsilon_m)^i + \sigma\Theta_\epsilon\Big[\mathcal{L}_{h_\epsilon}(h_\epsilon)(-\nabla_0 h_\epsilon, 1)\Big] \circ {\eta_\epsilon}^\tau \\ + \sigma\Theta_\epsilon\Big[[\mathcal{M}(h_\epsilon)(-\nabla_0 h_\epsilon, 1)] \circ {\eta_\epsilon}^\tau\Big]\end{aligned} \tag{11.2}$$

on $(0,T) \times \Gamma$, where

$$L^\epsilon_m = \frac{1}{2}\mathcal{J}_\epsilon^{-1}\left[\rho_\epsilon * \left(\frac{g_\epsilon}{g_0}\right)\right]_{,\beta}\Big[\mathcal{J}_\epsilon\mathcal{P}''(\mathcal{J}_\epsilon) + 2\mathcal{P}'(\mathcal{J}_\epsilon)\Big]g^{\alpha\beta}_\epsilon \eta_{\epsilon,\alpha} + \Big[\mathcal{J}_\epsilon\mathcal{P}'(\mathcal{J}_\epsilon) + \mathcal{P}(\mathcal{J}_\epsilon)\Big]H_\epsilon n_\epsilon.$$

By studying the tangential component of (11.2), we find that for $\gamma = 1, 2$,

$$\mathcal{J}_\epsilon^{-1}\left[\rho_\epsilon * \left(\frac{g_\epsilon}{g_0}\right)\right]_{,\gamma}\Big[\mathcal{J}_\epsilon\mathcal{P}''(\mathcal{J}_\epsilon) + 2\mathcal{P}'(\mathcal{J}_\epsilon)\Big] = 2[\nu D_{\eta_\epsilon}(v_\epsilon)^j_i - q_\epsilon\delta^j_i](a_\epsilon)^\ell_j N_\ell {\eta_\epsilon}^i_{,\gamma}. \tag{11.3}$$

Take T_ϵ even smaller so that

$$\begin{aligned} &\frac{1}{2} \le \|\Theta_\epsilon\|_{H^{1.5}(\Gamma)} \le \frac{3}{2}, && \frac{1}{2} \le \|a_\epsilon\|_{H^2(\Omega)} \le \frac{3}{2}, \\ &\|v_\epsilon\|_{L^2(0,T_\epsilon;H^3(\Omega))} \le \|u_0\|^2_{H^3(\Omega)} + 1, && \|\eta_\epsilon\|_{H^3(\Omega)} \le |\Omega| + 1. \end{aligned}$$

With these bounds, (11.3) together with the assumptions that $\mathcal{P}$ is strictly convex and $\mathcal{P}$ attains its minimum at $\mathcal{J} = 1$ (that assure that the second bracket of the left-hand side of (11.3) is bounded away from zero) implies that

$$\left\| \nabla_0 \left[\rho_\epsilon * \left(\frac{g_\epsilon}{g_0} \right) \right] \right\|_{H^{1.5}(\Gamma)} \leq C(u_0, \Omega). \tag{11.4}$$

Since (11.4) is independent of the ϵ, we find that

$$\|g_\epsilon\|_{H^{2.5}(\Gamma)} \leq C(u_0, g_0, \Omega). \tag{11.5}$$

Having (11.5), we no longer need ϵ-regularization to estimate the boundary integral in Remark 21 and the study of (A.1), and hence all the estimates in the previous sections are still valid with $C(M)$ replaced by $C(u_0, g_0, \Omega)$. These ϵ-independent estimates allow us to construct a solution (v_ϵ, h_ϵ) in $X(T)$ (where T is independent of ϵ) with the same estimates. The solution of the original problem (1.1) is then the limit of (v_ϵ, h_ϵ) as $\epsilon \to 0$.

11.4. The uniqueness of the solution. The uniqueness of the solution follows from the elliptic estimate

$$\|g - \tilde{g}\|^2_{H^{2.5}(\Gamma)} \leq C\Big[\|v - \tilde{v}\|^2_{H^3(\Omega)} + \|v_t - \tilde{v}_t\|^2_{H^1(\Omega)} \Big],$$

which follows from the equation

$$\left(\frac{g - \tilde{g}}{g_0} \right)_{,\gamma} \mathcal{Q}(\eta) + \left(\frac{\tilde{g}}{g_0} \right)_{,\gamma} \Big[\mathcal{Q}(\eta) - \mathcal{Q}(\tilde{\eta}) \Big] = F(v, q)^\gamma - F(\tilde{v}, \tilde{q})^\gamma,$$

where

$$\mathcal{Q}(\eta) = \mathcal{J}^{-1} \Big[\mathcal{J} \mathcal{P}''(\mathcal{J}) + 2\mathcal{P}'(\mathcal{J}) \Big] \quad \text{and} \quad F(v, q)^\gamma = 2[\nu D_\eta(v)^j_i - q\delta^j_i] a^\ell_j N_\ell \eta^i_{,\gamma}.$$

Appendix A. Elliptic regularity. We establish a κ-independent elliptic estimate for solutions of

$$\frac{\bar{\Theta}}{\sqrt{\det(g_0)}} \Big[\Big(\sqrt{\det(g_0)} \bar{A}^{\alpha\beta\gamma\delta} h_{\kappa,\alpha\beta} \Big)_{,\gamma\delta} (-\nabla_0 \bar{h}, 1) \Big] \circ \bar{\eta}^\tau + \kappa \Delta_0^2 v_\kappa = f, \tag{A.1}$$

where h_κ and v_κ satisfy (7.4) with $h_\kappa \in H^4(\Gamma)$, $v_\kappa \in H^4(\Gamma)$, and $f \in H^{1.5}(\Gamma)$. Letting $w = v_\kappa \circ \bar{\eta}^{-\tau}$, (A.1) is equivalent to

$$\frac{\bar{\Theta}}{\sqrt{\det(g_0)}} \Big[\sqrt{\det(g_0)} \bar{A}^{\alpha\beta\gamma\delta} h_{\kappa,\alpha\beta} \Big]_{,\gamma\delta} (-\nabla_0 \bar{h}, 1) + \kappa \Delta_0^2 w = f \circ \bar{\eta}^\tau, \tag{A.2}$$

which implies that

$$\begin{aligned} &\frac{\bar{\Theta}}{\sqrt{\det(g_0)}} \Big[\sqrt{\det(g_0)} \bar{A}^{\alpha\beta\gamma\delta} h_{\kappa,\alpha\beta} \Big]_{,\gamma\delta} + \kappa J_{\bar{h}}^{-2} \Delta_0^2 w \cdot (-\nabla_0 \bar{h}, 1) \\ &= J_{\bar{h}}^{-2} f \circ \bar{\eta}^\tau \cdot (-\nabla_0 \bar{h}, 1). \end{aligned} \tag{A.3}$$

Recall that $w \cdot (-\nabla_0 \bar{h}, 1) = h_{\kappa t}$.

Let D_h denote the difference quotients (with respect to the surface coordinate system). Taking the inner product of (A.3) with $D_{-h}D_h\nabla_0^4 h_\kappa$, by Corollary 7.1 we find that

$$\nu_1 \int_0^t \|D_h\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)} ds \le C(\epsilon_1)\int_0^t \Big[\|h_\kappa\|^2_{H^2(\Gamma)} + \|f\|^2_{H^1(\Gamma)} + \kappa\|w\|^2_{H^4(\Gamma)}\Big] ds.$$

Since the right-hand side is independent of difference parameter h, it follows that $h_\kappa \in H^5(\Gamma)$ (as it is already a H^4-function) with the estimate

$$\int_0^t \|\nabla_0^5 h_\kappa\|^2_{L^2(\Gamma)} ds \le C(\epsilon_1)\int_0^t \Big[\|h_\kappa\|^2_{H^2(\Gamma)} + \|f\|^2_{H^1(\Gamma)} + \kappa\|w\|^2_{H^4(\Gamma)}\Big] ds. \tag{A.4}$$

Next, we obtain a κ-independent estimate of $\kappa\|w\|^2_{H^4(\Gamma)}$. By taking the inner product of (A.2) with $\nabla_0^2 w$ and $\nabla_0^4 w$, we find that

$$\begin{aligned}&\|\nabla_0^3 h_\kappa(t)\|^2_{L^2(\Gamma)} + \kappa\int_0^t \|w\|^2_{H^3(\Gamma)} ds\\ &\quad\le C(\epsilon_1)\int_0^t \Big[\|\nabla_0^3 h_\kappa\|^2_{L^2(\Gamma)} + \|f\|^2_{L^2(\Gamma)} + \|w\|^2_{H^{2.5}(\Omega)}\Big] ds\end{aligned} \tag{A.5}$$

and

$$\begin{aligned}&\|\nabla_0^4 h_\kappa(t)\|^2_{L^2(\Gamma)} + \kappa\int_0^t \|w\|^2_{H^4(\Gamma)} ds\\ &\le C(\epsilon_1,\delta_1)\int_0^t \Big[\|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)} + \|f\|^2_{H^{1.5}(\Gamma)} + \|w\|^2_{H^3(\Omega)}\Big] ds + \delta_1\int_0^t \|\nabla_0^5 h_\kappa\|^2_{L^2(\Gamma)} dS,\end{aligned} \tag{A.6}$$

where we use (A.5) to estimate $\kappa\int_0^t \|w\|_{H^3(\Gamma)} ds$. Equation (A.6) provides a κ-independent estimate for $\kappa\|w\|^2_{H^4(\Gamma)}$; hence by choosing $\delta_1 > 0$ small enough, (A.4) implies that for all $t \in [0,T]$,

$$\int_0^t \|\nabla_0^2 h_\kappa\|^2_{H^3(\Gamma)} ds \le C'\int_0^t \Big[\|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)} + \|f\|^2_{H^{1.5}(\Gamma)} + \|w\|^2_{H^3(\Omega)}\Big] ds \tag{A.7}$$

for some constant C' depending on ϵ_1.

Appendix B. Inequalities in the estimates for $\nabla_0^2 v$ near the boundary.

B.1. κ-independent estimates. Since $\zeta_1 \equiv 1$ on Γ and

$$\begin{aligned}(-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot\nabla_0^4 v_\kappa &= \nabla_0^4((-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot v_\kappa) - \nabla_0^4(-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot v_\kappa\\ &\quad - 4\nabla_0^3(-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot\nabla_0 v_\kappa - 6\nabla_0^2(-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot\nabla_0^2 v_\kappa\\ &\quad - 4\nabla_0(-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot\nabla_0^3 v_\kappa,\end{aligned}$$

we find that

$$
\begin{aligned}
&\int_\Gamma \bar\Theta\Big[L_{\bar h}(h_\kappa)\circ\bar\eta^\tau\Big]((-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot\nabla_0^2(\zeta_1^2\nabla_0^2 v_\kappa))dS\\
&= -\int_\Gamma \bar\Theta\Big[L_{\bar h}(h_\kappa)\circ\bar\eta^\tau\Big]\Big[\nabla_0^4(-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot v_\kappa + 4\nabla_0^3(-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot\nabla_0 v_\kappa\\
&\qquad\qquad + 6\nabla_0^2(-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot\nabla_0^2 v_\kappa\Big]dS \qquad (\equiv I_1)\\
&\quad - 4\int_\Gamma \bar\Theta\Big[L_{\bar h}(h_\kappa)\circ\bar\eta^\tau\Big](\nabla_0(-\nabla_0\bar h\circ\bar\eta^\tau,1)\cdot\nabla_0^3 v_\kappa)dS \qquad (\equiv I_2)\\
&\quad + \int_\Gamma \frac{\bar\Theta}{\sqrt{\det(g_0)}}\nabla_0^2\Big[\sqrt{\det(g_0)}\Big(L_1^{\alpha\beta\gamma}\tilde h_{,\alpha\beta\gamma}+L_2\Big)\circ\bar\eta^\tau\Big]\nabla_0^2(h_{\kappa t}\circ\bar\eta^\tau)dS \qquad (\equiv I_3)\\
&\quad + \int_\Gamma \frac{2\nabla_0\bar\Theta}{\sqrt{\det(g_0)}}\nabla_0\Big[\sqrt{\det(g_0)}\Big(L_1^{\alpha\beta\gamma}\tilde h_{,\alpha\beta\gamma}+L_2\Big)\circ\bar\eta^\tau\Big]\nabla_0^2(h_{\kappa t}\circ\bar\eta^\tau)dS \qquad (\equiv I_4)\\
&\quad + \int_\Gamma (\nabla_0^2\bar\Theta)\Big[\Big(L_1^{\alpha\beta\gamma}\tilde h_{,\alpha\beta\gamma}+L_2\Big)\circ\bar\eta^\tau\Big]\nabla_0^2(h_{\kappa t}\circ\bar\eta^\tau)dS \qquad (\equiv I_5)\\
&\quad + \int_\Gamma \frac{\bar\Theta}{\sqrt{\det(g_0)}}\Big[(\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta}h_{\kappa,\alpha\beta})_{,\gamma\delta}\circ\bar\eta^\tau\Big]\nabla_0^4(h_{\kappa t}\circ\bar\eta^\tau)dS.
\end{aligned}
$$

The last term of the identity above, by a change of coordinates, can be written as

$$
\begin{aligned}
&\int_\Gamma \frac{\bar\Theta}{\sqrt{\det(g_0)}}\Big[(\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta}h_{\kappa,\alpha\beta})_{,\gamma\delta}\circ\bar\eta^\tau\Big]\nabla_0^4(h_{\kappa t}\circ\bar\eta^\tau)dS\\
&= \int_\Gamma \frac{B}{\sqrt{\det(g_0)}}\nabla_0^2(\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta}h_{\kappa,\alpha\beta})_{,\gamma\delta}\nabla_0^2 h_{\kappa t}dS + R_1\\
&\quad + 2\int_\Gamma \frac{\nabla_0\bar\Theta}{\sqrt{\det(g_0)}}\nabla_0\Big[(\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta}h_{\kappa,\alpha\beta})_{,\gamma\delta}\circ\bar\eta^\tau\Big]\nabla_0^2(h_{\kappa t}\circ\bar\eta^\tau)dS \qquad (\equiv J_1)\\
&\quad + \int_\Gamma \frac{\nabla_0^2\bar\Theta}{\sqrt{\det(g_0)}}\Big[(\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta}h_{\kappa,\alpha\beta})_{,\gamma\delta}\circ\bar\eta^\tau\Big]\nabla_0^2(h_{\kappa t}\circ\bar\eta^\tau)dS \qquad (\equiv J_2)\\
&= \frac12\frac{d}{dt}\int_\Gamma B\bar A^{\alpha\beta\gamma\delta}\nabla_0^2 h_{\kappa,\alpha\beta}\nabla_0^2 h_{\kappa,\gamma\delta}dS + R_1',
\end{aligned}
$$

where $B = b^t\otimes b^t\otimes b^t\otimes b^t$ with $b=\nabla_0\bar\eta^\tau$, and

$$
\begin{aligned}
R_1(t) &= \int_\Gamma b^t\otimes b^t\otimes(\nabla_0 b^t)\otimes(\nabla_0 b^t)\nabla_0(\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta}h_{\kappa,\alpha\beta})_{,\gamma\delta}\nabla_0 h_{\kappa t}dS \quad (\equiv J_3)\\
&\quad + \int_\Gamma b^t\otimes b^t\otimes b^t\otimes(\nabla_0 b^t)\nabla_0(\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta}h_{\kappa,\alpha\beta})_{,\gamma\delta}\nabla_0^2 h_{\kappa t}dS \quad (\equiv J_4)\\
&\quad + \int_\Gamma b^t\otimes b^t\otimes b^t\otimes(\nabla_0 b^t)\nabla_0^2(\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta}h_{\kappa,\alpha\beta})_{,\gamma\delta}\nabla_0 h_{\kappa t}dS \quad (\equiv J_5)
\end{aligned}
$$

and

$$
\begin{aligned}
R_1'(t) &= R_1(t)+J_1(t)+J_2(t)-\frac12\int_\Gamma (B\bar A^{\alpha\beta\gamma\delta})_t\nabla_0^2 h_{\kappa,\alpha\beta}\nabla_0^2 h_{\kappa,\gamma\delta}dS \quad (\equiv J_6)\\
&\quad + 2\int_\Gamma \frac{B}{\sqrt{\det(g_0)}}\nabla_0(\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta})\nabla_0 h_{\kappa,\alpha\beta}\nabla_0^2 h_{\kappa t,\gamma\delta}dS \quad (\equiv J_7)\\
&\quad + \int_\Gamma \frac{B}{\sqrt{\det(g_0)}}\nabla_0^2(\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta})h_{\kappa,\alpha\beta}\nabla_0^2 h_{\kappa t,\gamma\delta}dS \quad (\equiv J_8)
\end{aligned}
$$

$$+ 2\int_\Gamma \frac{B_{,\gamma}}{\sqrt{\det(g_0)}} \nabla_0^2(\sqrt{\det(g_0)}\bar{A}^{\alpha\beta\gamma\delta} h_{\kappa,\alpha\beta}) \nabla_0^2 h_{\kappa t,\delta} dS \quad (\equiv J_9)$$
$$+ \int_\Gamma \frac{B_{,\gamma\delta}}{\sqrt{\det(g_0)}} \nabla_0^2(\sqrt{\det(g_0)}\bar{A}^{\alpha\beta\gamma\delta} h_{\kappa,\alpha\beta}) \nabla_0^2 h_{\kappa t} dS \quad (\equiv J_{10}).$$

It follows that

$$|I_1| \le C(\epsilon_1)(1 + \|\nabla_0^4 h_\kappa\|_{L^2(\Gamma)})\|\nabla_0^2 v_\kappa\|_{H^1(\Omega_1')},$$
$$|I_3| + |I_4| + |I_5| \le C(M)(1 + \|\tilde{h}\|_{H^5(\Gamma)})\|\nabla_0^2 v_\kappa\|_{H^1(\Omega_1)}$$

and hence that

$$|I_1| + |I_3| + |I_4| + |I_5| \le C(\epsilon_1)\Big[\|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)} + \|\tilde{h}\|^2_{H^5(\Gamma)} + 1\Big] + \delta\|v_\kappa\|^2_{H^3(\Omega)}.$$

It follows that

$$|J_2| + |J_3| + |J_5| + |J_{10}| \le C(\epsilon_1)\|\nabla_0^4 h_\kappa\|_{L^2(\Gamma)}\|\nabla_0^2 h_{\kappa t}\|_{L^2(\Gamma)},$$
$$|J_6| \le C(M)(\|\tilde{v}\|_{H^3(\Omega)} + \|\tilde{h}_t\|_{H^{2.5}(\Gamma)})\|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)}.$$

We need only obtain κ-independent estimates for the terms I_2, J_1, J_4, J_7, J_8, and J_9. By the $H^{-0.5}(\Gamma)$-$H^{0.5}(\Gamma)$ duality pairing,

$$|I_2| \le C(M)\Big[\|\nabla_0^2 h_\kappa\|_{H^{2.5}(\Gamma)} + 1\Big]\|v_\kappa\|_{H^{2.5}(\Gamma)}.$$

Therefore, by interpolation and Young's inequality,

$$|I_2| \le C\Big[\|h_\kappa\|^2_{H^4(\Gamma)} + 1\Big] + \delta_1\|\nabla_0^2 h_\kappa\|^2_{H^3(\Gamma)} + \delta\|v_\kappa\|^2_{H^3(\Omega)} \tag{B.1}$$

for some C depending on M, δ, and δ_1.

For J_1, J_4, and J_9, we find that

$$|J_1| + |J_4| + |J_9| \le C(\epsilon_1)\|h_\kappa\|_{H^{4.5}(\Gamma)}\|v_\kappa\|_{H^{2.5}(\Gamma)}$$
$$\le C'\Big[\|\nabla_0^2 h_\kappa\|^2_{H^2(\Gamma)} + 1\Big] + \delta_1\|\nabla_0^2 h_\kappa\|^2_{H^3(\Gamma)} + \delta\|v_\kappa\|^2_{H^3(\Omega)}$$

for some constant C' depending on M, ϵ_1, δ, and δ_1.

For J_7 and J_8, by the $H^{-1.5}(\Gamma)$-$H^{1.5}(\Gamma)$ duality pairing,

$$|J_7| + |J_8| \le C(M)\|B\|_{H^{1.5}(\Gamma)}\|\bar{h}\|_{H^{3.5}(\Gamma)}\|h_\kappa\|_{H^{4.5}(\Gamma)}\|v_\kappa\|_{H^{2.5}(\Gamma)}.$$

Similarly to the estimate in (B.1), we find that

$$|J_7| + |J_8| \le C(M)\Big[\|h_\kappa\|^2_{H^4(\Gamma)} + 1\Big] + \delta_1\|\nabla_0^2 h_\kappa\|^2_{H^3(\Gamma)} + \delta\|v_\kappa\|^2_{H^3(\Omega)}.$$

Summing all the estimates and then integrating in time from 0 to t, by Corollary 7.1 and the fact that B is close to 1 in the uniform norm for T small,

$$\frac{\nu_1}{2}\|\nabla_0^4 h_\kappa(t)\|^2_{L^2(\Gamma)} \le \int_0^t\int_\Gamma \bar{\Theta}\Big[[L_{\bar{h}}(h_\kappa)(-\nabla_0\bar{h},1)]\circ\bar{\eta}^\tau\Big]\cdot\nabla_0^2(\zeta_1^2\nabla_0^2 v_\kappa)dSds$$
$$+ C'\int_0^t K(s)\|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)}ds + C'\int_0^t\Big[\|\tilde{h}\|^2_{H^5(\Gamma)} + 1\Big]ds$$
$$+ \delta\int_0^t\|v_\kappa\|^2_{H^3(\Omega)}ds + \delta_1\int_0^t\|\nabla_0^2 h_\kappa\|^2_{H^3(\Gamma)}ds$$

for some constant C' depending on M, ϵ_1, δ, and δ_1, where

$$K(s) := 1 + \|\tilde{v}\|^2_{H^3(\Omega)} + \|\tilde{h}\|^2_{H^5(\Gamma)} + \|\tilde{h}_t\|^2_{H^{2.5}(\Gamma)}.$$

B.2. ϵ_1-independent estimates. We next obtain ϵ_1-independent estimates for the first two terms of I_1, as well as those for I_2, J_1, J_2, J_3, J_4, J_5, J_9, and J_{10} with h_κ replaced by h_{ϵ_1}. Let

$$I_1^1 = -\int_\Gamma \bar\Theta \Big[L_{\tilde h}(h_{\epsilon_1}) \circ \bar\eta^\tau\Big]\Big[\nabla_0^4(-\nabla_0\bar h \circ \bar\eta^\tau, 1) \cdot v_{\epsilon_1}\Big] dS,$$
$$I_1^2 = -4\int_\Gamma \bar\Theta \Big[L_{\tilde h}(h_{\epsilon_1}) \circ \bar\eta^\tau\Big]\Big[\nabla_0^3(-\nabla_0\bar h \circ \bar\eta^\tau, 1) \cdot \nabla_0 v_{\epsilon_1}\Big] dS.$$

By the $H^{-1.5}(\Gamma)$-$H^{1.5}(\Gamma)$ duality pairing,

$$|I_1^1| + |I_1^2| \le C(M)\|L_{\tilde h}(h_{\epsilon_1})\|_{H^{1.5}(\Gamma)}\|v_{\epsilon_1}\|_{H^{2.5}(\Gamma)}\|(\nabla_0\tilde h)\circ\bar\eta^\tau\|_{H^{2.5}(\Gamma)}.$$

Therefore, by (6.6) and (9.12),

$$\text{(B.2)}\quad |I_1^1| + |I_1^2| \le C(M)t^{1/4}\Big[\|h_{\epsilon_1}\|^2_{H^{5.5}(\Gamma)} + 1\Big]\|v_{\epsilon_1}\|_{H^3(\Omega)}$$
$$\le Ct^{1/2}\Big[\|v_{\epsilon_1 t}\|^2_{H^1(\Omega)} + \|\nabla_0^4 h_{\epsilon_1}\|^2_{L^2(\Gamma)} + \|F\|^2_{H^1(\Omega)} + 1\Big] + (\delta + Ct^{1/2})\|v_{\epsilon_1}\|^2_{H^3(\Omega)}$$

for some constant C depending on M and δ.

For J_1, we use an L^4-L^4-L^2-type of Hölder inequality and conclude that

$$|J_1| \le C(M)t^{1/2}\|h_{\epsilon_1}\|_{H^{5.5}(\Gamma)}\|v_{\epsilon_1}\|_{H^{2.5}(\Gamma)},$$

while for the other J terms, we use the $H^{0.5}(\Gamma)$-$H^{-0.5}(\Gamma)$ duality pairing to obtain

$$|J_2| + |J_3| + |J_4| + |J_5| + |J_9| + |J_{10}| \le C(M)t^{1/2}\|h_{\epsilon_1}\|_{H^{5.5}(\Gamma)}\|v_{\epsilon_1}\|_{H^{2.5}(\Gamma)},$$

and hence all the J terms are bounded by the same right-hand side of the inequality in (B.2). Therefore,

$$\frac{\nu_1}{2}\|\nabla_0^4 h_{\epsilon_1}(t)\|^2_{L^2(\Gamma)} \le \int_0^t\int_\Gamma \bar\Theta\Big[[L_{\bar h}(h_{\epsilon_1})(-\nabla_0\bar h, 1)]\circ\bar\eta^\tau\Big]\cdot\nabla_0^2(\zeta_1^2\nabla_0^2 v_{\epsilon_1})dSds$$
$$+ CN_2(u_0, F) + C\int_0^t K(s)\|\nabla_0^4 h_{\epsilon_1}\|^2_{L^2(\Gamma)}ds + (\delta + Ct^{1/2})\int_0^t \|v_{\epsilon_1}\|^2_{H^3(\Omega)}ds$$
$$+ (\delta_1 + Ct^{1/2})\int_0^t \|v_{\epsilon_1 t}\|^2_{H^1(\Omega)}ds$$

for some constant C depending on M, δ, and δ_1.

Appendix C. $L_t^2H_x^1$-estimates for v_t. By the chain rule and integrating by parts,

$$\int_\Gamma \Big[\bar\Theta[L_{\bar h}(h_\kappa)(-\nabla_0\bar h, 1)]\circ\bar\eta^\tau\Big]_t \cdot v_{\kappa t}dS = \int_\Gamma \bar\Theta_t\Big[L_{\bar h}(h_\kappa)\Big]\circ\bar\eta^\tau(-\nabla_0\bar h\circ\bar\eta^\tau, 1)\cdot v_{\kappa t}dS$$
$$+ \int_\Gamma \bar\Theta\bar\eta_t^\tau\cdot\Big[\nabla_0[L_{\bar h}(h_\kappa)](-\nabla_0\bar h, 1)\Big]\circ\bar\eta^\tau\cdot v_{\kappa t}dS \quad (\equiv K_1)$$
$$+ \int_\Gamma \bar\Theta\Big[[L_{\bar h}(h_\kappa)](\nabla_0\bar h, -1)\Big]_t\circ\bar\eta^\tau\cdot v_{\kappa t}dS \quad (\equiv K_2).$$

The first term is bounded by

$$C(M)\|\bar v\|_{H^3(\Omega)}\Big[\|\nabla_0^4 h_\kappa\|_{L^2(\Gamma)} + 1\Big]\|v_{\kappa t}\|_{L^2(\Gamma)}.$$

After integrating by parts, the most difficult term to estimate in K_1 consists of the integral

$$\int_\Gamma \frac{\bar v}{\sqrt{\det(g_0)}}\Big[[\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta}h_{\kappa,\alpha\beta}]_{,\gamma\delta}(\nabla_0\bar h,-1)\Big]\circ\bar\eta^\tau\nabla_0 v_{\kappa t}dS.$$

Integrating from 0 to t and integrating by parts in time, we find that

$$\begin{aligned}&\int_0^t\int_\Gamma \frac{\bar v}{\sqrt{\det(g_0)}}\Big[[\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta}h_{\kappa,\alpha\beta}]_{,\gamma\delta}(\nabla_0\bar h,-1)\Big]\circ\bar\eta^\tau\nabla_0 v_{\kappa t}dSds\\ &=-\int_0^t\int_\Gamma \frac{\bar v}{\sqrt{\det(g_0)}}\Big[[\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta}h_{\kappa,\alpha\beta}]_{t,\gamma\delta}(\nabla_0\bar h,-1)\Big]\circ\bar\eta^\tau\nabla_0 v_{\kappa}dSds+R_3,\end{aligned}$$

where R_3 is bounded by

$$\begin{aligned}C\int_0^t\Big[1+\|\tilde v_t\|^2_{H^1(\Omega)}\Big]\|\nabla_0^4h_\kappa\|^2_{L^2(\Gamma)}ds+\delta_2\|\nabla_0^4h_\kappa\|^2_{L^2(\Gamma)}\\ +\delta\int_0^t\|v_\kappa\|^2_{H^3(\Omega)}ds+(\delta+Ct^{1/2})\int_0^t\|v_{\kappa t}\|^2_{H^1(\Omega)}ds\end{aligned}$$

for some constant C depending on M, δ, and δ_2. Next, using that

$$[(-\nabla_0\bar h,1)\circ\bar\eta^\tau]\cdot\nabla_0 v_\kappa=b^t(\nabla_0h_{\kappa t})\circ\bar\eta^\tau+b^t(\nabla_0^2\bar h\circ\bar\eta^\tau,0)\cdot v_\kappa$$

and integrating by parts, we find that the integral on the right-hand side is identical to

$$\frac12\int_0^t\int_\Gamma\frac{1}{\sqrt{\det(g_0)}}\nabla_0\Big[\sqrt{\det(g_0)}\bar\Theta\bar v b^t\bar A^{\alpha\beta\gamma\delta}\Big]h_{\kappa t,\alpha\beta}h_{\kappa t,\gamma\delta}dSds+R_4,$$

where

$$|R_4|\le C(M)C(\delta)\int_0^t\|\nabla_0^4h_\kappa\|^2_{L^2(\Gamma)}ds+\delta\int_0^t\|v_\kappa\|^2_{H^3(\Omega)}ds.$$

By interpolation, the integral part is bounded by

$$C\Big[N(u_0,F)+\int_0^t\|\nabla_0^4h_\kappa\|^2_{L^2(\Gamma)}ds\Big]+\delta\int_0^t\|v_\kappa\|^2_{H^3(\Omega)}ds+Ct\int_0^t\|v_{\kappa t}\|^2_{H^1(\Omega)}ds$$

for some constant C depending on M and δ. Therefore, K_1 satisfies

(C.1)
$$\begin{aligned}\Big|\int_0^tK_1ds\Big|\le C\int_0^t\Big[K(s)\Big(\|\nabla_0^4h_\kappa\|^2_{L^2(\Gamma)}+\|\nabla_0^2h_{\kappa t}\|^2_{L^2(\Gamma)}\Big)+1\Big]ds+\delta_2\|\nabla_0^4h_\kappa\|^2_{L^2(\Gamma)}\\ +(\delta+Ct^{1/2})\int_0^t\|v_\kappa\|^2_{H^3(\Omega)}ds+(\delta+Ct^{1/2})\int_0^t\|v_{\kappa t}\|^2_{H^1(\Omega)}ds\end{aligned}$$

for some constant C depending on M, δ, and δ_2.

For K_2, by time differentiating the evolution equation, we find that

$$\begin{aligned}(-\nabla_0\bar h\circ\bar\eta^\tau,1)v_{\kappa t}=h_{\kappa tt}\circ\bar\eta^\tau+\bar v^\tau\cdot(\nabla_0h_{\kappa t})\circ\bar\eta^\tau-\bar v^\tau\cdot(\nabla_0^2\bar h\circ\bar\eta^\tau,0)\cdot v_\kappa\\ -(\nabla_0\bar h_t\circ\bar\eta^\tau,0)\cdot v_\kappa,\end{aligned}$$

and hence (after a change of coordinates)

$$\begin{aligned}K_2 = &\int_\Gamma [L_{\bar h}(h_\kappa)]_t h_{\kappa tt} dS + \int_\Gamma [L_{\bar h}(h_\kappa)]_t [(\bar v^\tau \circ \bar\eta^{-\tau}) \cdot (\nabla_0 h_{\kappa t})] dS \quad (\equiv K_3)\\ &- \int_\Gamma [L_{\bar h}(h_\kappa)]_t [(\nabla_0 \bar h_t, 0) \cdot (v_\kappa \circ \bar\eta^{-\tau})] dS \quad (\equiv K_4)\\ &- \int_\Gamma [L_{\bar h}(h_\kappa)]_t [(\bar v^\tau \circ \bar\eta^{-\tau}) \cdot (\nabla_0^2 \bar h, 0)(v_\kappa \circ \bar\eta^{-\tau})] dS \quad (\equiv K_5)\\ &+ \int_\Gamma [L_{\bar h}(h_\kappa)] [(\nabla_0 \bar h_t, 0) \cdot (v_{\kappa t} \circ \bar\eta^{-\tau})] dS \quad (\equiv K_6).\end{aligned}$$

For the first term, we have

$$\begin{aligned}\int_\Gamma [L_{\bar h}(h_\kappa)]_t h_{\kappa tt} dS = &\frac{1}{2}\frac{d}{dt}\int_\Gamma \bar A^{\alpha\beta\gamma\delta} h_{\kappa t,\alpha\beta} h_{\kappa t,\gamma\delta} dS\\ &+ \int_\Gamma \frac{1}{\sqrt{\det(g_0)}}\Big[\sqrt{\det(g_0)}(\bar A^{\alpha\beta\gamma\delta})_t\Big]_{,\gamma\delta} h_{\kappa,\alpha\beta} h_{\kappa tt} dS \quad (\equiv K_7) + R_5,\end{aligned} \tag{C.2}$$

where R_5 is bounded by

$$C\Big[1 + \|\tilde h_t\|^2_{H^{2.5}(\Gamma)}\Big]\Big[1 + \|\nabla_0^2 h_{\kappa t}\|^2_{L^2(\Gamma)}\Big] + \delta\Big[\|v_\kappa\|^2_{H^2(\Omega)} + \|\nabla_0^2 v_\kappa\|^2_{H^1(\Omega_1')}\Big] + \delta_1 \|v_{\kappa t}\|^2_{H^1(\Omega)}$$

for some constant C depending on M, δ, and δ_1. Also, by the inequality $\|h_{\kappa tt}\|_{L^4(\Gamma)} \le C(M)\big[\|v_\kappa\|_{H^2(\Omega)} + \|v_{\kappa t}\|_{H^1(\Omega)}\big]$,

$$\begin{aligned}|K_7| &\le C \|[\sqrt{\det(g_0)}(\bar A^{\alpha\beta\gamma\delta})_t]_{,\gamma\delta}\|_{H^{-0.5}(\Gamma)} \left\|\frac{1}{\sqrt{\det(g_0)}} h_{\kappa,\alpha\beta} h_{\kappa tt}\right\|_{H^{0.5}(\Gamma)}\\ &\le C(M)C(\delta,\delta_1)\|\tilde h_t\|^2_{H^{2.5}(\Gamma)}\|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)} + \delta\|v_\kappa\|^2_{H^2(\Omega)} + \delta_1\|v_{\kappa t}\|^2_{H^1(\Omega)}.\end{aligned}$$

REMARK 22. *The bound for K_7 can be refined even further as*

$$|K_7| \le C(M)C(\delta)\|\tilde h_t\|^2_{H^{1.5}(\Gamma)}\|\nabla_0^2 h_\kappa\|^2_{H^{1.5}(\Gamma)} + \delta\|v_\kappa\|^2_{H^3(\Omega)} + \delta\|v_{\kappa t}\|^2_{H^1(\Omega)};$$

it is this inequality that will be used in the proof of the fixed-point argument.

It remains to estimate K_3 to K_6. By proper use of Hölder's inequality,

$$\begin{aligned}|K_3| + |K_5| + |K_6| \le &C\Big[1 + \|\tilde h_t\|^2_{H^{2.5}(\Gamma)}\Big]\Big[1 + \|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)}\Big]\\ &+ (\delta + Ct^{1/2})\|v_\kappa\|^2_{H^3(\Omega)} + \delta\|v_{\kappa t}\|^2_{H^1(\Omega)}\end{aligned}$$

for some constant C depending on M and δ. For K_4, most of the terms can be estimated in the same fashion, except the term

$$\int_\Gamma \frac{1}{\sqrt{\det(g_0)}}\Big[\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta} h_{\kappa t,\alpha\beta}\Big][(\nabla_0 \bar h_{t,\gamma\delta}, 0) \cdot (v_\kappa \circ \bar\eta^{-\tau})] dS,$$

which is identical to

$$\int_\Gamma \left\{\frac{1}{\sqrt{\det(g_0)}}\Big[\sqrt{\det(g_0)}\bar A^{\alpha\beta\gamma\delta} h_{\kappa t,\alpha\beta}\Big][(\nabla_0 \bar h_{,\gamma\delta}, 0) \cdot (v_\kappa \circ \bar\eta^{-\tau})]\right\}_t dS \ (\equiv K_8) + R_6,$$

where

$$|R_6| \le C\|\tilde{h}\|^2_{H^{5.5}(\Gamma)}\Big[\|v_\kappa\|^2_{L^2(\Omega)} + \|\nabla_0^2 h_{\kappa t}\|^2_{L^2(\Gamma)}\Big] + \delta\|v_\kappa\|^2_{H^3(\Omega)} + \delta_1\|v_{\kappa t}\|^2_{H^1(\Omega)}$$

for some constant C depending on M, δ, and δ_1. Time integrating K_8 and using the interpolation inequality together with Young's inequality, we find that

$$\begin{aligned}\left|\int_0^t K_8(s)ds\right| &\le C(M)\Big[\|u_0\|^2_{H^{2.5}(\Omega)} + \|\nabla_0^2 h_{\kappa t}\|_{L^2(\Omega)}\|v_\kappa\|_{L^4(\Omega)}\Big]\\ &\le C(M)C(\delta_1,\delta_2)N_3(u_0,F) + \delta_2\|\nabla_0^2 h_{\kappa t}\|^2_{L^2(\Gamma)} + \delta_1\int_0^t \|v_{\kappa t}\|^2_{H^1(\Omega)}ds, \end{aligned}\tag{C.3}$$

where

$$\begin{aligned}N_3(u_0,F) := \|u_0\|^2_{H^{2.5}(\Omega)} &+ \|u_0\|^2_{H^{4.5}(\Gamma)} + \|F\|^2_{L^2(0,T;H^1(\Omega))}\\ &+ \|F_t\|^2_{L^2(0,T;H^1(\Omega)')} + \|F(0)\|^2_{H^1(\Omega)} + 1,\end{aligned}$$

and we use $\|v_\kappa\|^2_{H^1(\Omega)} \le C\big[\int_0^t \|v_{\kappa t}\|^2_{H^1(\Omega)}ds + \|u_0\|^2_{H^1(\Omega)}\big]$ to obtain (C.3), and hence

$$\begin{aligned}\sum_{i=3}^6 |K_i| \le C\Big[1 + \|\tilde{h}\|^2_{H^{5.5}(\Gamma)} + \|\tilde{h}_t\|^2_{H^{2.5}(\Gamma)}\Big]\Big[1 + \|v_\kappa\|^2_{L^2(\Omega)} + \|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)}\Big]\\ + (\delta + Ct^{1/2})\|v_\kappa\|^2_{H^3(\Omega)} + \delta_1\|v_{\kappa t}\|^2_{H^1(\Omega)} + K_8\end{aligned}\tag{C.4}$$

with K_8 satisfying inequality (C.3). Finally, combining all the estimates,

(C.5)

$$\begin{aligned}&\int_0^t \|\nabla_0^2 h_{\kappa t}\|^2_{L^2(\Gamma)}ds \le \int_0^t\int_\Gamma \Big[[L_{\bar{h}}(h_\kappa)(\nabla_0\bar{h},-1)]\circ\bar{\eta}^\tau\Big]_t \cdot v_{\kappa t}dS + CN_3(u_0,F)\\ &+ C\int_0^t K(s)\Big[\|v_\kappa\|^2_{L^2(\Omega)} + \|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)} + \|\nabla_0^2 h_{\kappa t}\|^2_{L^2(\Gamma)}\Big]ds\\ &+ (\delta + Ct^{1/2})\int_0^t \|v_\kappa\|^2_{H^3(\Omega)}ds + (\delta_1 + Ct^{1/2})\int_0^t \|v_{\kappa t}\|^2_{H^1(\Omega)}ds + \delta_2\|\nabla_0^4 h_\kappa\|^2_{L^2(\Gamma)}\end{aligned}$$

for some constant C depending on M, δ, δ_1, and δ_2.

Acknowledgment. We thank the referees for the valuable comments that helped improve the manuscript.

REFERENCES

[1] H. Beirão da Veiga, *On the existence of strong solutions to a coupled fluid-structure evolution problem*, J. Math. Fluid Mech., 6 (2004), pp. 21–52.
[2] M. Boulakia, *Existence of weak solutions for an interaction problem between an elastic structure and a compressible viscous fluid*, J. Math. Pures Appl. (9), 84 (2005), pp. 1515–1554.
[3] R. S. Chadwick, *Axisymmetric indentation of a thin incompressible elastic layer*, SIAM J. Appl. Math., 62 (2002), pp. 1520–1530.
[4] A. Chambolle, B. Desjardins, M. J. Esteban, and C. Grandmont, *Existence of weak solutions for an unsteady fluid-plate interaction problem*, J. Math. Fluid Mech., 7 (2005), pp. 368–404.
[5] C. H. Cheng, D. Coutand, and S. Shkoller, *Navier-Stokes Equations Interacting with a Nonlinear Elastic Solid Shell*, preprint, 2006.

[6] P. G. Ciarlet, *Mathematical Elasticity*, Vol. III, Stud. Math. Appl. 29, North–Holland, Amsterdam, 2000.

[7] D. Coutand and S. Shkoller, *On the motion of an elastic solid inside of an incompressible viscous fluid*, Arch. Ration. Mech. Anal., 176 (2005), pp. 25–102.

[8] D. Coutand and S. Shkoller, *On the interaction between quasilinear elastodynamics and the Navier-Stokes equations*, Arch. Ration. Mech. Anal., 179 (2006), pp. 303–352.

[9] B. Desjardins and M. J. Esteban, *Existence of weak solutions for the motion of rigid bodies in a viscous fluid*, Arch. Ration. Mech. Anal., 146 (1999), pp. 59–71.

[10] B. Desjardins, M. J. Esteban, C. Grandmont, and P. Le Tallec, *Weak solutions for a fluid-structure interaction problem*, Rev. Mat. Complut., 14 (2001), pp. 523–538.

[11] Y. C. Fung, *Biomechanics: Mechanical Properties of Living Tissues*, Springer-Verlag, New York, 1981.

[12] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations*, Vol. I, Springer Tracts Nat. Philos. 38, Springer-Verlag, New York, 1994.

[13] W. M. Gelbart, *Micelles, Membranes, Microemulsions, and Monolayers*, Springer-Verlag, New York, 1995.

[14] C. Grandmont and Y. Maday, *Existence for unsteady fluid-structure interaction problem*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 609–636.

[15] C. Liu and N. J. Walkington, *An Eulerian description of fluids containing visco-elastic particles*, Arch. Ration. Mech. Anal., 159 (2001), pp. 229–252.

[16] L. Miao, U. Seifert, M. Wortis, and H. Dobereiner, *Budding transitions of fluid-bilayer vesicle: The effect of area-difference elasticity*, Phys. Rev. A (3), 49 (1994), pp. 5389–5407.

[17] Z. Ou-Yabg and W. Helfrich, *Bending energy of vesicle membranes: General expressions for the first, second and third variation of the shape energy and applications to spheres and cylinders*, Phys. Rev. E (3), 39 (1989), pp. 5280–5288.

[18] U. Seifert, *Adhesion of vesicles in two dimensions*, Phys. Rev. A (3), 43 (1991), pp. 6803–6814.

[19] U. Seifert, *Curvature-induced lateral phase segregation in two-component vesicles*, Phys. Rev. Lett., 70 (1993), pp. 1335–1338.

[20] D. Serre, *Chute libre d'un solide dans un fluide visqueux incompressible: Existence*, Japan J. Appl. Math., 4 (1987), pp. 33–73.

[21] D. J. Steigmann, *Fluid films with curvature elasticity*, Arch. Ration. Mech. Anal., 150 (1999), pp. 127–152.

[22] H. F. Weinberger, *Variational properties of steady fall in Stokes flow*, J. Fluid Mech., 52 (1972), pp. 321–344.

SIAM J. MATH. ANAL.
Vol. 39, No. 3, pp. 801–818

CLASSICAL SOLUTIONS FOR NONELLIPTIC EULER–LAGRANGE EQUATIONS VIA CONTINUATION*

MARKUS LILLI†

Abstract. We consider a higher-gradient model in one-dimensional nonlinear elasticity. Therefore we consider a physically reasonable stored-energy density W such that $W(\nu)$ goes to infinity for $\nu \searrow 0$ and $\nu \nearrow \infty$. We prescribe a parameter-dependent body-forcing. Our main goal is to show under a certain sign condition for the body-force that global solution branches of the singular perturbed problem converge to a global branch of weak solutions of vanishing capillarity. Moreover, the solution satisfies the first and the second Weierstrass–Erdmann corner conditions, and the strain field is uniformly pointwise positive.

Key words. nonlinear elasticity, singular limits, global continuation

AMS subject classification. 34B15, 74N15, 74G55

DOI. 10.1137/060662186

1. Introduction. This paper is motivated to a large extent by results presented in [10]. There a two-phase model for an elastic solid in the presence of live body-forcing and interfacial or higher-gradient effects is considered, the latter characterized by a small capillarity coefficient $\varepsilon > 0$. The existence of families of equilibria is proved by continuation methods, and the main goal in [10] is the analysis of those solutions in the limit $\varepsilon \searrow 0$. Assuming that the loading is bounded and everywhere nonnegative, uniform a priori bounds on solutions via maximum-principle arguments are obtained. An important by-product of that analysis is a certain monotonicity property for all solutions. In the end, global weak solutions that satisfy the Maxwell condition are obtained.

In this work we pursue precisely the same question, but we allow the loading to change sign exactly one time in a prescribed manner. The latter makes the application of maximum principles much more delicate. We impose very general, physically reasonable growth conditions on both the stored-energy function and the body-force function. Also, in contrast to [10], we impose the pointwise unilateral constraint introduced in [4], ensuring that all deformations of the bar are injective (with the concomitant infinite growth of the stored energy).

The outline of the work is as follows: In the second section we introduce the model, the variational problem under consideration, and the corresponding Euler–Lagrange equation. Afterwards we give the assumptions on the potentials. The existence of solutions of the singular perturbed Euler–Lagrange equation via a global implicit function theorem is fairly routine, which we summarize in section 3. Uniform a priori bounds on solutions are given in section 4. There we pay careful attention to the injectivity constraint and show that the strains are uniformly (pointwise) positive. Furthermore, we prove the unboundedness of the branch. To prove this for every λ is another main difference compared to [10], where a similar result is obtained via a maximum principle which does not apply to the more general body-force considered

*Received by the editors June 6, 2006; accepted for publication (in revised form) February 27, 2007; published electronically August 22, 2007. This work was supported by the Graduiertenkolleg "Nichtlineare Probleme in Analysis, Geometrie und Physik" at the University of Augsburg.
http://www.siam.org/journals/sima/39-3/66218.html
†Institut für Mathematik, Universität Augsburg, 86135 Augsburg, Germany (lilli@math.uni-augsburg.de).

in this work. Our approach also applies in [10]. Hence all the results given there remain valid if one replaces the condition that the body-force is bounded by the more general growth conditions given in Theorem 5. The next section is the core of the paper. We introduce a further assumption on the body-force, namely, that it changes sign exactly once. By a careful application of maximum principles and monotonicity criteria, we prove some properties for all solutions lying on the branch. These properties enable us to obtain a singular limit in a classical function space (in contrast to [21], where, under weaker assumptions on the body-force, the existence of Young-measure solutions was proved). Moreover, we show in section 6 that this limit is a solution of the Euler–Lagrange equation of the unperturbed problem and satisfies the Maxwell conditions, forming a global and unbounded continuum. We remark that the results denoted here, without imposing the pointwise unilateral constraint ensuring that all deformations of the bar are injective, can also be found in [20].

One of the most powerful tools in nonlinear analysis is the Leray–Schauder degree. It was especially used for semilinear elliptic PDEs of second order to prove existence and obtain qualitative properties of classical solutions (see [25], or, for a recent treatment, [17]). Even in quasi-linear equations the degree was used to obtain branches of positive solutions (see, e.g., [19]). In [14], [15], [16] the existence of classical solutions of a nonelliptic Euler–Lagrange equation was proved also by using singular perturbation, which in recent years has become a widely used tool to tackle nonconvex variational problems (see, e.g., [18], [22]).

In three-dimensional elasticity problems, global continuation results were obtained in [9], [11], [12] by employing generalized degree theoretic methods. In this case, the resulting solution continua are characterized not only by the usual two alternatives given in [25], but also by a third concerning the termination of the branch. This phenomenon can happen due to loss of local injectivity, ellipticity, and/or the complementing condition.

2. Formulation. We briefly discuss the physical model behind our analysis. We follow the lines of [13]; see also [8], [21].

We consider a one-dimensional elastic solid placed in a soft loading device. Let $[0,1]$ be the reference configuration in the undeformed state and let $v = v(x)$ be the displacement of the material point occupying position x in the undeformed state. Hence, $u(x) = x + v(x)$ is the placement of the bar, and $u'(x) = 1 + v'(x)$ is the stretch ratio.

We assume the bar to be homogenous, which means that the stored-energy density $W \in C^2(\mathbb{R}^+)$ is independent of the spatial variable and depends only on u'. Furthermore, let $B(\lambda, u, x) : \mathbb{R} \times \mathbb{R}_0^+ \times [0,1] \mapsto \mathbb{R}$ be a potential delivering in general a live body-force

$$-\frac{\partial B}{\partial u}(\lambda, u, x) := b(\lambda, u, x),$$

where $b \in C(\mathbb{R} \times \mathbb{R}_0^+ \times [0,1])$. The left end of the bar is fixed, which assigns the boundary condition $u(0) = 0$. We assume

$$b(0, \cdot, \cdot) \equiv 0. \tag{2.1}$$

We consider a potential $W \geq 0$ and we impose the following assumptions on $W \in C^2(\mathbb{R}^+)$: Because we assume the bar to be stable in the undeformed state, we derive

$$W(1) = \sigma(1) = 0 \text{ and } W''(1) > 0, \tag{2.2}$$

where $\sigma(\nu) := W'(\nu)$ denotes the stress. Moreover,

$$\begin{aligned} &W(\nu) \geq W(1) = 0 \text{ for every } \nu \in \mathbb{R}^+, \\ &\lim_{\nu \searrow 0} W(\nu) = \lim_{\nu \nearrow \infty} W(\nu) = \infty. \end{aligned} \tag{2.3}$$

Condition $(2.3)_2$ reflects the fact that one needs an infinite amount of energy to compress the bar to volume zero and also that an infinite amount of energy is necessary to stretch the bar to infinity.

Furthermore, we assume the existence of ν_1, $\nu_2 \in (1, \infty)$ such that

$$W''(\nu) \begin{cases} < 0 \text{ for } \nu \in (\nu_1, \nu_2), \\ > 0 \text{ otherwise.} \end{cases} \tag{2.4}$$

Let W be a so-called one-well potential, which means

$$\sigma(\nu) \begin{cases} < 0 \text{ for } \nu \in (0, 1), \\ > 0 \text{ for } \nu \in (1, \infty). \end{cases} \tag{2.5}$$

We require the following growth conditions for σ and b:

(i) We assume

$$\lim_{\nu \to \infty} \frac{\sigma(\nu)\nu}{|\nu|^{p+1}} \geq K > 0 \tag{2.6}$$

and

(ii)

$$\sup_{\lambda \in [-\hat{\lambda}, \hat{\lambda}]} |b(\lambda, u, x)| \leq c_3 |u|^r + c_4 \tag{2.7}$$

for some $r < p$, $c_3 = c_3(\hat{\lambda})$, $c_4 = c_4(\hat{\lambda}) \geq 0$. Moreover, we have c_3, $c_4 < \infty$ provided that $\hat{\lambda} < \infty$ and both constants can be chosen independently of x.

The total *potential energy* of the bar is given by

$$\begin{aligned} &J(u) := \int_0^1 [W(u') + B(\lambda, u, x)]\, dx, \\ &u(0) = 0. \end{aligned} \tag{2.8}$$

Here "λ" is a loading parameter. The corresponding Euler–Lagrange equation is

$$\begin{aligned} &\frac{d}{dx} [\sigma(u')] + b(\lambda, u, x) = 0, \\ &u(0) = 0 \quad \text{and} \quad \sigma(u'(1)) = 0. \end{aligned} \tag{2.9}$$

Note that (2.8) has no global minimizer, in general, due to the fact that W is nonconvex. In particular, existence of a solution of (2.9) cannot be guaranteed by the direct methods of the calculus of variations (see [7] for details). For the same reason, the boundary value problem (2.9) is singular, obviating any systematic solution strategy.

Instead we introduce a “relaxed” variational problem by adding an additional strain gradient-term, intended to model interfacial energy:

$$
\begin{aligned}
J_\varepsilon(u) &:= \int_0^1 \Big(\frac{\varepsilon}{2}(u'')^2 + W(u') + B(\lambda, u, x)\Big)\, dx, \\
u(0) &= 0,
\end{aligned}
\tag{2.10}
$$

where $\varepsilon > 0$ is a small parameter. The Euler–Lagrange equation of equilibrium is the fourth-order equation

$$
-\varepsilon u^{(4)} + \frac{d}{dx}\left[\sigma(u')\right] + b(\lambda, u, x) = 0 \tag{2.11}
$$

with boundary conditions

$$
u(0) = u''(0) = u''(1) = 0, \qquad \varepsilon u'''(1) = \sigma(u'(1)). \tag{2.12}
$$

Integration of (2.11) yields the system

$$
\begin{aligned}
u' &= z, \\
-\varepsilon z'' + \sigma(z) &= \int_x^1 b(\lambda, u(\eta), \eta) d\eta, \\
u(0) &= z'(0) = z'(1) = 0.
\end{aligned}
\tag{2.13}
$$

In contrast to (2.8) and (2.9), (2.10) and (2.13) are amenable to existence methods for each $\varepsilon > 0$ (see [3], [7], [10] for details).

3. Global analysis. We just sketch the ideas. The details can be found in [10] and [20]. The only difference is that one has to take care of $(2.3)_2$, which is done by methods introduced in [9].

We define $Y := C^2([0,1])$, we let

$$
Y_0 := \{y \in Y \mid y(0) = 0\}, Y_1 := \left\{y \in Y \mid \int_0^1 y dx = 0\right\},
$$

and, moreover, we define

$$
\mu := \int_0^1 u'(s) ds = \int_0^1 z(s) ds, \quad v := z - \mu.
$$

Finally we define the triple

$$
w := (u, v, \mu) \in \mathcal{W} \equiv Y_0 \times Y_1 \times \mathbb{R},
$$

where $\mathcal{W}$ is endowed with the norm $||w||_{\mathcal{W}} = ||u||_{C^2} + ||v||_{C^2} + |\mu|$ and, in particular, $\mathcal{W}$ is a Banach space. Moreover, let

$$
K := \{(y, \mu) \in Y_1 \times \mathbb{R} \mid y(x) + \mu > 0 \text{ for all } x \in [0,1]\}
$$

and for arbitrary $\delta > 0$

$$
K_\delta := \{(y, \mu) \in K \mid y(x) + \mu > \delta \text{ for all } x \in [0,1]\}.
$$

Obviously $(\lambda, w) := (0, w_0) \in \mathbb{R} \times \mathcal{W}$ is a solution of (2.13), where

$$w_0 := (Id, 0, 1).$$

By an easy calculation one obtains that (2.13) can be written as a compact perturbation of the identity (see [10] for details). Hence the Leray–Schauder degree is well defined, and a well-known argument proved in [25] and generalized in [9], employing the homotopy invariance of the degree and taking the restriction $u' > 0$ into account, yields the following proposition.

Proposition 1. *Let $\mathcal{C}_\varepsilon \subset \mathbb{R} \times \mathcal{W}$ be the connected component of solutions of* (2.13) *containing the trivial solution* $(0, w_0)$. *Then $\mathcal{C}_\varepsilon$ is characterized by at least one of the following alternatives:*

(i) *$\mathcal{C}_\varepsilon$ is unbounded in $\mathbb{R} \times \mathcal{W}$.*
(ii) *$\mathcal{C}_\varepsilon \setminus \{(0, 0)\}$ is connected.*
(iii) *$\mathcal{C}_\varepsilon \setminus \mathbb{R} \times Y_0 \times K_\delta \neq \emptyset$ for every $\delta > 0$.*

Remark 2. Alternative (ii) means that the branch forms a loop. The third alternative means that the branch can terminate and is due to the singularity of W. If (iii) holds, then there exists a sequence $(\lambda_n, u_n, z_n)_{n\in\mathbb{N}} \in \mathcal{C}_\varepsilon$ and $(x_n)_{n\in\mathbb{N}} \in [0, 1]$ with

$$(\lambda_n)_{n\in\mathbb{N}} \text{ is bounded and } z_n(x_n) \searrow 0.$$

4. Properties of the continuum. By a standard phase-plane analysis like the one in [5] it is easy to show that alternative (ii) in Proposition 1 is impossible (see [10] for details).

Furthermore, we have the following theorem.

Theorem 3. *Let σ and b satisfy the growth conditions* (2.6) *and* (2.7). *Fix some arbitrary λ_0 and let $(\lambda, u_\varepsilon^\lambda, v_\varepsilon^\lambda, \mu_\varepsilon^\lambda) \in \mathcal{C}_\varepsilon$ for $\lambda \in [-\lambda_0, \lambda_0]$ and $z_\varepsilon^\lambda := v_\varepsilon^\lambda + \mu_\varepsilon^\lambda$. Then there exists some $\delta > 0$, such that for every $\lambda \in [-\lambda_0, \lambda_0]$ and for every $\varepsilon > 0$ we have*

$$z_\varepsilon^\lambda(x) > \delta$$

for all $x \in [0, 1]$.

We give a sketch of the proof. For a detailed proof we refer to [21].

We prove by contradiction and therefore we assume sequences $(\varepsilon_n)_n$, $(x_n)_n \in [0, 1]$ and $(z_{\varepsilon_n})_n := (z_n)_n$ with

$$\varepsilon_n \searrow 0, \quad z_n(x_n) \searrow 0.$$

Step 1. First we prove by virtue of (2.13), $(2.3)_2$, and (2.7) $||z_n||_{L^{p+1}} \to \infty$.

Step 2. We define

$$A_n := \{x \in [0, 1] \mid z_n(x) > 1\} := \bigcup_{j=1}^{k(n)} (x_1^j(n), x_2^j(n)).$$

Observe that

$$z_n'(x_1^j(n)) \geq 0, \quad z_n'(x_2^j(n)) \leq 0 \tag{4.1}$$

for all j. Multiplying (2.13) by z_n and integrating over A_n yields

$$\int_{A_n} \sigma(x, z_n) z_n dx \leq \int_{A_n} \left[\int_x^1 (b(u_n(\eta), \eta) d\eta + \tau) \, z_n \right] dx \tag{4.2}$$

for every $n \in \mathbb{N}$.

Step 3. From Step 1 we deduce

$$||z_n||_{L^{p+1}(A_n)} \to \infty. \tag{4.3}$$

Multiplying (4.2) by $\frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}}$, we obtain for every $n \in \mathbb{N}$

$$\begin{aligned} &\frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \int_{A_n} \sigma(x, z_n) z_n \, dx \\ &\qquad \le \frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \int_{A_n} \left(\int_x^1 b(u_n(\eta), \eta) d\eta + \tau \right) z_n dx. \end{aligned} \tag{4.4}$$

One can prove the following:

(a) The right side of (4.4) converges to 0 for $n \to \infty$.

(b) We have

$$\lim_{n\to\infty} \frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \int_{A_n} \sigma(x, z_n) z_n \, dx \ge K > 0 \tag{4.5}$$

with K given in (2.6).

By virtue of (4.4) we get

$$\begin{aligned} 0 < K &\le \lim_{n\to\infty} \left(\frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \int_{A_n} \sigma(x, z_n) z_n \, dx \right) \\ &\le \lim_{n\to\infty} \left(\frac{1}{||z_n||^{p+1}_{L^{p+1}(A_n)}} \int_{A_n} \left(\int_x^1 b(u_n(\eta), \eta) d\eta + \tau \right) z_n \, dx \right) = 0, \end{aligned}$$

which is obviously a contradiction.

Remark 4. For fixed λ consider $(u_\varepsilon, v_\varepsilon, \mu_\varepsilon) \in \mathcal{C}_\varepsilon$ with associated z_ε. Then, by Theorem 3 there exists $\delta > 0$ such that $z_\varepsilon > \delta$ for every $\varepsilon > 0$. In particular, Theorem 3 excludes alternative (iii) for every finite λ_0.

Hence the global continuum $\mathcal{C}_\varepsilon$ is unbounded in $\mathbb{R} \times \mathcal{W}$, but one can prove more by the assumed growth conditions on σ and b.

THEOREM 5. *Let σ and b satisfy* (2.6), (2.7) *and let $\varepsilon > 0$. Then the projection of $\mathcal{C}_\varepsilon$ onto the λ-axis is $\mathbb{R}$.*

For the proof we refer to [21]; see also [20].

LEMMA 6. *Let σ and b satisfy the growth conditions as in Theorem* 5 *and let $(z_n)_{n\in\mathbb{N}}$ be a sequence with $z_n \in \mathcal{C}_{\varepsilon_n}$ for all n as $\varepsilon_n \searrow 0$. Then we obtain the following: The sequence $(z_n)_{n\in\mathbb{N}}$ is uniformly bounded in $C^0([0,1])$. Furthermore, there exists a subsequence of $(z_n)_{n\in\mathbb{N}}$, which we again denote by $(z_n)_{n\in\mathbb{N}}$, such that the following holds: We have $\varepsilon_n z_n'' \rightharpoonup^* 0$ in $L^\infty(0,1)$.*

One can find the proof of Lemma 6 in [21]. For the reader's convenience we give a sketch of the proof.

Let $(z_n)_{n\in\mathbb{N}}$ be a sequence as in Lemma 6 and assume the existence of $x_m \in [0,1]$ and n_m such that $z_{n_m}(x_m) > m$ for every $m \in \mathbb{N}$. Without loss of generality, we assume that $x = x_m$ is a maximum of z_{n_m}. By properties (2.5) and (2.6) we deduce

$$\lim_{m\to\infty} \sigma(z_{n_m}(x_m)) = \infty. \tag{4.6}$$

Moreover, we have for every $m \in \mathbb{N}$

$$-\varepsilon_{n_m} z''_{n_m}(x_m) + \sigma(z_{n_m}(x_m)) = \int_{x_m}^1 b(\lambda, u_{n_m}(\eta), \eta) d\eta. \tag{4.7}$$

The growth conditions (2.7) imply that for every $m \in \mathbb{N}$

$$\left| \int_{x_m}^1 b(\lambda, u_{n_m}(\eta), \eta) d\eta \right| \leq c_5,$$

where c_5 is independent of m. Therefore we have by (4.6) and (4.7)

$$z''_{n_m}(x_m) > 0$$

for m sufficiently large. On the other hand, we have $z'_{n_m}(x_m) = 0$ for every $m \in \mathbb{N}$. Hence x_m cannot be a maximum, which contradicts our assumption.

5. Geometric structure of $\mathcal{C}_\varepsilon^+$. In order to obtain a priori estimates for $(z_n)_{n\in\mathbb{N}}$ to establish *strong* convergence in some Sobolev space, we consider a body-force potential b with the following properties.

There exists $x_0 = x_0(\lambda) \in (0, 1)$ such that for every $\lambda \in \mathbb{R}^+$ and $u \in \mathbb{R}_0^+$ we have

(H1) $b(\lambda, u, \cdot)_{|[0,x_0]} \geq 0$,
(H2) $b(\lambda, u, \cdot)_{|[x_0,1]} \leq 0$.
(H3) Furthermore, there exist an open subset $\Omega_2 \subseteq (x_0, 1]$ and $b(\lambda, u, \cdot)_{|\Omega_2} < 0$.
(H4) There exists $\vartheta > 0$ such that $b(\lambda, u, \cdot)_{|(0,\vartheta)} > 0$ for every $\lambda \in \mathbb{R}^+$ and $u \in \mathbb{R}_0^+$.

Example 7. The above assumptions are valid for the following body-force: Let W be some "one-well" potential satisfying the growth conditions with some $p \in \mathbb{N}$, and let $b(\lambda, u, x) := c(\lambda, x)(1 + u^r)$ with $r < p$, $c \in C(\mathbb{R} \times [0, 1], \mathbb{R})$ and c changing sign from + to − at $x_0 = x_0(\lambda)$ for $\lambda > 0$.

Let $\mathcal{C}_\varepsilon^+$ be the unbounded branch obtained by the global implicit function theorem for fixed $\varepsilon > 0$ and $\lambda > 0$, and let $(u, v, \mu) \in \mathcal{C}_\varepsilon^+$, $z := v + \mu$. For our convenience we say $z \in \mathcal{C}_\varepsilon$ if z is constructed as above. Our goal is to show that by virtue of (H1)–(H4) z' changes sign in $(0, 1)$ at most once.

One can find the following definition in [2].

DEFINITION 8. *A function v has exactly k changes of sign in* $[0, 1]$ *provided that the following hold:*

(i) *There are points $s_1 < s_2 < \cdots < s_{k+1} \in [0, 1]$ with $v(s_i)v(s_{i+1}) < 0$ for all $i = 1, \ldots, k$.*
(ii) *We will say that v changes sign at the points $x_1 < \cdots < x_k$ provided $v(x_i) = 0$ and the points in* (i) *can be chosen so that $s_i < x_i < s_{i+1}$ and v does not change sign in $[x_i, x_{i+1}]$.*

Let us further remark that for our forthcoming analysis differentiation of (2.13) turns out to be very useful:

$$\begin{aligned} &h = z', \\ &\varepsilon h'' - \sigma'(z)h = b(\lambda, u, \cdot), \\ &h(0) = h(1) = 0. \end{aligned} \tag{5.1}$$

Remark 9. By the continuation method, (2.4), and $z_{|\{\lambda=0\}} \equiv 1$, there exists λ_0 sufficiently small such that $\sigma'(z_\lambda) > 0$ for every $0 \leq \lambda \leq \lambda_0$.

Hence we derive the following lemma.

LEMMA 10. *Let $z_\lambda \in \mathcal{C}_\varepsilon^+$ and consider λ_0 such that $\sigma'(z_\lambda)_{|\lambda\in[0,\lambda_0]} > 0$. Then z'_λ changes sign at most once for $0 < \lambda \le \lambda_0$.*

Proof. We prove by contradiction and accordingly we assume that z'_λ changes sign at least two times in $[0,1]$. By Definition 8 there exist $s_1, s_2, s_3 \in (0,1)$ with $z'_\lambda(s_i)z'_\lambda(s_{i+1}) < 0$ and $x_1, x_2 \in (0,1)$ such that $s_1 < x_1 < s_2 < x_2 < s_3$ and $z'_\lambda(x_1) = z'_\lambda(x_2) = 0$. By the boundary condition $z'_\lambda(0) = z'_\lambda(1) = 0$ we have three nodal areas of z'_λ in $[0,1]$. On the other hand, we have just one sign change of b, and hence, there exists a nodal area $\Omega' \subseteq [0,1]$ of z'_λ such that

$$\operatorname{sign} b_{|\Omega'} = \operatorname{sign} z'_{\lambda|\Omega'}. \tag{5.2}$$

For every $x \in \Omega'$ we have by (5.1)

$$\begin{aligned} \varepsilon(z'_\lambda)'' - \sigma'(z_\lambda)z'_\lambda &= b, \\ z'_{\lambda|\partial\Omega'} &= 0. \end{aligned} \tag{5.3}$$

By the weak maximum principle we conclude $\operatorname{sign} b_{|\Omega'} \ne \operatorname{sign} z'_{\lambda|\Omega'}$, which contradicts (5.2). ∎

First we give a property of an arbitrary solution of (2.13).

LEMMA 11. *Let z be a solution of* (2.13). *Then there is no negative minimum of z' in $(x_0, 1)$.*

Proof. Suppose there exists $x_1 \in (x_0, 1)$ with

$$z'(x_1) < 0, \quad z''(x_1) = 0, \quad z'''(x_1) \ge 0.$$

With respect to the identity

$$-\sigma'(z(x_1))z'(x_1) = -\varepsilon z'''(x_1) + b(\lambda, u(x_1), x_1) \le 0,$$

we obtain by (2.4)

$$z(x_1) \in [\nu_1, \nu_2] \subseteq (1, \infty). \tag{5.4}$$

On the other hand, we have the identity $\sigma(z(x_1)) = \int_{x_1}^1 b(\lambda, u(\tau), \tau)d\tau \le 0$, which yields $z(x_1) \le 1$ by (2.5), and this contradicts (5.4). ∎

Again maximum principles yield for λ sufficiently small the following lemma.

LEMMA 12. *Let λ_0 be defined by Remark* 9 *and let $z_\lambda \in \mathcal{C}_\varepsilon^+$ for $\lambda < \lambda_0$ and an arbitrary $\varepsilon > 0$. Then there exists no interval $J \subseteq [0, x_0]$ such that $z'_{\lambda|J} \equiv 0$.*

Proof. Suppose there exists an interval $J := [x_\alpha, x_\beta] \subseteq [0, x_0]$ with $z'_{\lambda|J} \equiv 0$. For the remainder of the proof we omit λ. Let $\hat{x} \in (x_\alpha, x_\beta)$ and $z'_{|(0,\hat{x})}$ satisfy the equation

$$\begin{aligned} \varepsilon(z')'' - \sigma'(z)z' &= b \ge 0, \\ z'(0) = z'(\hat{x}) &= z''(\hat{x}) = 0. \end{aligned} \tag{5.5}$$

The weak maximum principle tells us $z'_{|(0,\hat{x})} \le 0$, and the strong maximum principle implies $z'_{|(0,\hat{x})} \equiv 0$ and, in particular, $z'_{|(0,x_\beta)} \equiv 0$. By virtue of (5.5) we have $b_{|(0,x_\beta)} \equiv 0$, contradicting (H4). ∎

LEMMA 13. *Let λ_0 be defined by Remark* 9 *and let $\lambda \in (0, \lambda_0)$ and $z_\lambda \in \mathcal{C}_\varepsilon^+$ for an arbitrary but fixed $\varepsilon > 0$. We assume the existence of an interval $J \subseteq (x_0, 1)$ such* *that* $z'_{\lambda|J} \equiv 0$. *Then we obtain* $z'_\lambda(x) \le 0$ *for every* $x \in [0,1]$. *Moreover, there exists* $\bar{x} \in (x_0, 1]$ *with* $z'_{\lambda|(0,\bar{x})} < 0$ *and* $z'_{\lambda|(\bar{x},1)} \equiv 0$.

Proof. Let $J := [x_\alpha, x_\beta] \subseteq (x_0, 1)$ be an interval with $z'_{\lambda|J} \equiv 0$. Similar to the proof of Lemma 12, one can show that this implies

$$z'_{\lambda|[x_\alpha,1]} \equiv 0.$$

Furthermore (H3) excludes $z'_{\lambda|(x_0,1)} \equiv 0$. Thus there exists an $x \in (x_0, 1)$ such that $z'_\lambda(x) \ne 0$, and we let $\bar{x} \in (x_0, 1)$ such that $z'_\lambda(\bar{x}) = z''_\lambda(\bar{x}) = 0$ and $z'_\lambda(x) \ne 0$ for $x \in (\bar{x} - \delta, \bar{x})$ for some $\delta > 0$. The strong maximum principle applied to (5.1) implies

$$z'_{\lambda|(\bar{x}-\delta,\bar{x})} < 0.$$

The weak maximum principle yields

$$z'_{\lambda|(x_0,1)} \le 0, \quad z'_{\lambda|(x_0,\bar{x})} < 0.$$

By a further application of the maximum principle the proof is done. ∎

By virtue of Lemmas 12 and 13 we can refine our possibilities for z'_λ for λ small.

Corollary 14. *Choose an arbitrary but fixed* $\varepsilon > 0$. *Furthermore, consider* $z_\lambda \in \mathcal{C}^+_\varepsilon$ *for* $\lambda \in (0, \lambda_0)$ *with* λ_0 *such that* $\sigma'(z_\lambda) > 0$. *Then we have the following two possibilities:*

(i) z'_λ *has exactly one sign change, and, moreover,* $z''_\lambda(0) < 0$, $z''_\lambda(1) < 0$, *and the only zero at* $x = x_1$ *of* z'_λ *is nondegenerate with* $z''_\lambda(x_1) > 0$. *In particular, the zero of* z'_λ *in* $(0,1)$ *is unique,* $z'_{\lambda|(0,x_1)} < 0$, $z'_{\lambda|(x_1,1)} > 0$.

(ii) z'_λ *does not change sign, and if there exists an interval* $J \subseteq (0,1)$ *such that* $z'_{\lambda|J} \equiv 0$, *then* $J = (\bar{x}, 1]$ *with* $\bar{x} > x_0$ *and* $z'_{\lambda|(0,\bar{x})} < 0$.

Proof. We let λ_0 be defined as above and we omit λ in the remainder of the proof.

(i) We assume that z' has exactly one sign change at x_1. By virtue of Lemmas 12 and 13 there exists no interval J with $z'_{|J} \equiv 0$, and the weak maximum principle applied to (5.1) implies immediately $z'_{|(0,x_1)} \le 0$ and $z'_{|(x_1,1)} \ge 0$. Hopf's maximum principle yields the assertion.

(ii) This case is identical to Lemmas 12 and 13. ∎

In the remainder of this section we propose a continuation method to obtain the following: The function z'_λ has at most one sign change in $(0,1)$ for every $\lambda \in \mathbb{R}^+$.

Remark 15. Because $\mathcal{C}^+_\varepsilon$ is connected in the C^2-topology by construction of the continuum, we deduce that the corresponding z'_λ are connected in the C^1-topology. A well-known result is the following proposition.

Proposition 16. *There are two (respectively, three) possibilities for creating a new zero on a* C^1*-component:*

1. *There exists* $x_2 \in (0,1)$ *such that* $z'(x_2) = z''(x_2) = 0$ *and, moreover, we have* $\operatorname{sign}(z'(x_2 - \gamma)) = \operatorname{sign}(z'(x_2 + \gamma))$ *for every* $\gamma \in (0, \kappa)$, κ *sufficiently small.*
2. *To create a new zero at the boundary we assume either* $z''(0) = 0$ *or* $z''(1) = 0$.
3. *Another possibility in case* (i) *of Corollary* 14 *is the following: Let* z' *change sign at* $x = x_1$ *with* $z'(x_1) = z''(x_1) = 0$.

In the forthcoming analysis we will often apply the well-known generalized maximum principle to (5.1), where the sign of $\sigma'(z)$ does not matter. In what follows we refer to Hopf's generalized maximum principle by (HGM). For a proof of (HGM) we refer to [24].

Lemma 17. *Let $\lambda_1 \in \mathbb{R}^+$ and let $x = x_1 \in (0,1)$ be the nondegenerate zero of z'_{λ_1} such that $z'_{\lambda_1|(0,x_1)} < 0$ and $z'_{\lambda_1|(x_1,1)} > 0$. Moreover, we assume $z''_{\lambda_1}(0) < 0$ and $z''_{\lambda_1}(1) < 0$. Then we have two possibilities:*

1. *z'_λ has the same properties as z'_{λ_1} for every $\lambda > \lambda_1$.*
2. *We obtain existence of $\lambda_2 > \lambda_1$ with $z'_{\lambda_2} \le 0$. Moreover, we have $z''_{\lambda_2}(0) < 0$, and if there exists an interval $J \subseteq [0,1]$ with $z'_{\lambda_2|J} \equiv 0$, then $J = [x_2, 1]$ for some $x_2 > x_0$.*

Proof. Let x_1 be the nondegenerate zero of z'_{λ_1}, and by assumption we have

$$z'_{\lambda_1|(0,x_1)} < 0 \text{ and } z'_{\lambda_1|(x_1,1)} > 0.$$

Step 1. First we exclude the creation of a new zero if $x_1 \in (x_0, 1)$.

Suppose the existence of $\lambda_2 > \lambda_1$ and $x_2 \in (0, x_0)$ with $z'_{\lambda_2}(x_2) = z''_{\lambda_2}(x_2) = 0$. Moreover, we have $\operatorname{sign}(z'_{\lambda_2}(x_2 - \gamma)) = \operatorname{sign}(z'_{\lambda_2}(x_2 + \gamma))$ for every $\gamma \in (0, \kappa)$, κ sufficiently small. We know by Corollary 14 and the continuation method that $z'_{\lambda_2|(0,x_2)} < 0$, and, furthermore, we have

$$\varepsilon(z')'' - \sigma'(z)z' = b,$$

$$z'(0) = z'(x_2) = 0.$$

(HGM) yields $z'_{\lambda_2|[0,x_2]} \equiv 0$, which is impossible by (H4).

If we assume the existence of $x_3 \in (x_0, x_1)$ with $z'_{\lambda_2}(x_3) = z''_{\lambda_2}(x_3) = 0$, then we have two possibilities:

1. There exists $x' \in (x_3, x_1)$ such that z'_{λ_2} possesses a negative minimum at x', which is impossible by Lemma 11.
2. We have $z'_{\lambda_2|(x_3,1)} \equiv 0$ and we obtain $z'_{\lambda_2} \le 0$.

If we consider $x_4 \in (x_1, 1)$ with $z'_{\lambda_2}(x_4) = z''_{\lambda_2}(x_4) = 0$, then we obtain $z'_{\lambda_2|(x_4,1)} \equiv 0$ and $z'_{\lambda_2} \le 0$ by virtue of (HGM).

We consider the creation of a new zero at the boundary: First we observe that $z''_{\lambda_2}(0) = 0$ for some $\lambda_2 > \lambda_1$ is impossible by (HGM). Moreover, $z''_{\lambda_2}(1) = 0$ predicts again $z'_{\lambda_2} \le 0$.

Step 2. We consider $x_1 \in (0, x_0)$ such that $z'_{\lambda_1|(0,x_1)} < 0$ and $z'_{\lambda_1|(x_1,1)} > 0$.

First we exclude possibility one of Proposition 16: If there exists $x_2 \in (0, x_1)$, respectively, $x_3 \in (x_0, 1)$ with $z'_{\lambda_2}(x_j) = z''_{\lambda_2}(x_j) = 0$ for $j = 2, 3$ and for some $\lambda_2 > \lambda_1$, we obtain again by (HGM)

$$z'_{\lambda_2|(0,x_1)} \equiv 0, \text{ respectively, } z'_{\lambda_2|(x_0,1)} \equiv 0.$$

Both are impossible by virtue of (5.1) and (H3), (H4).

The case $x_4 \in (x_1, x_0)$ with $z'_{\lambda_2}(x_4) = z''_{\lambda_2}(x_4) = 0$ for some $\lambda_2 > \lambda_1$ involves more difficulties because a maximum principle does not apply. First we observe by the continuation method the existence of $\bar{x} \in (x_1, x_4)$ with

$$z'_{\lambda_2}(\bar{x}) > 0, \quad z''_{\lambda_2}(\bar{x}) = 0, \quad z'''_{\lambda_2}(\bar{x}) \le 0.$$

By the identity

$$-\sigma'(z_{\lambda_2}(\bar{x}))z'_{\lambda_2}(\bar{x}) = b(\lambda_2, u(\bar{x}), \bar{x}) - \varepsilon z'''_{\lambda_2}(\bar{x}) \ge 0,$$

we deduce $\sigma'(z_{\lambda_2}(\bar{x})) \le 0$ and therefore

$$z_{\lambda_2}(\bar{x}) \in [\nu_1, \nu_2].$$

Because of $z'_{\lambda_2|(x_1,1)} \geq 0$ we know that $z_{\lambda_2}(1) \geq z_{\lambda_2}(\bar{x}) \geq \nu_1 > 1$, and by virtue of $\sigma(z_{\lambda_2}(1)) = \varepsilon z''_{\lambda_2}(1)$ we get $z''_{\lambda_2}(1) > 0$, which yields

$$z'_{\lambda_2|(1-\varrho,1)} < 0 \quad \text{for } \varrho > 0 \text{ sufficiently small.}$$

But z' does not change sign in $(x_0, 1)$, hence a contradiction.

Next we exclude the creation of new zeros at the boundary: We assume existence of $\lambda_3 > \lambda_1$ with $z''_{\lambda_3}(0) = 0$, respectively, $z''_{\lambda_3}(1) = 0$. Both cases imply by (HGM) $z'_{\lambda_3|(0,x_1)} \equiv 0$, respectively, $z'_{\lambda_3|(x_0,1)} \equiv 0$, which is again impossible by (H3) and (H4).

Excluding possibility 3 of Proposition 16 is obvious. ∎

LEMMA 18. *Let $\lambda_1 \in \mathbb{R}^+$ be such that $z'_{\lambda_1} \leq 0$. Then we have two possibilities:*

1. *For every $\lambda > \lambda_1$ any solution $z_\lambda \in \mathcal{C}_\varepsilon^+$ satisfies the property $z'_\lambda \leq 0$.*
2. *There exists $\lambda_2 > \lambda_1$ such that z'_{λ_2} changes sign exactly once at $x = x_1$, and we have $z'_{\lambda_2|(0,x_1)} < 0$ and $z'_{\lambda_2|(x_1,1)} > 0$. Moreover, all zeros of z'_{λ_2} are nondegenerate.*

Proof. First we treat $z'_{\lambda_1}(x) \leq 0$ for a certain $\lambda_1 \in \mathbb{R}^+$ and for every $x \in [0,1]$. We have to show that z'_λ changes sign at most once for arbitrary $\lambda > \lambda_1$.

Again by (HGM) applied to (5.1) we exclude existence of a solution $z_\lambda \in \mathcal{C}_\varepsilon^+$ satisfying $z''_\lambda(0) = 0$. By the same reason existence of $x_1 \in (0, x_0]$ with $z'_\lambda(x_1) = z''_\lambda(x_1) = 0$ is impossible.

Oscillation of $z'_{\lambda|(x_0,1)}$ cannot occur by virtue of Lemma 11. Moreover, if a change in sign takes place at $x = x_1$, then we obtain by (HGM) $z'_{\lambda|(x_1,1)} > 0$, $z'_{\lambda|(0,x_1)} < 0$ and every zero of z'_λ is nondegenerate. ∎

LEMMA 19. *Let $\lambda_1 \in \mathbb{R}^+$ such that $z'_{\lambda_1} \geq 0$. Then we have two possibilities:*

1. *We have for every $\lambda > \lambda_1$ the property $z'_\lambda \geq 0$.*
2. *There exists $\lambda_2 > \lambda_1$ such that z'_{λ_2} changes sign exactly once. The zeros of z'_{λ_2} are nondegenerate.*

Proof. Let $z'_{\lambda_1|(0,1)} \geq 0$ for a certain $\lambda_1 > 0$. Then we have to show that z'_λ changes sign at most once for every $\lambda > \lambda_1$.

We assume for the remainder of the proof the existence of $\lambda_2 > \lambda_1$ such that z'_{λ_2} changes sign at least twice: Then there exists a local positive maximum at $x := x_\alpha$. Note that $x_\alpha \leq x_0$ holds due to Lemma 11. By virtue of $z'_{\lambda_2}(x_\alpha) > 0$ and $z'''_{\lambda_2}(x_\alpha) \leq 0$, we obtain

$$-\sigma'(z_{\lambda_2}(x_\alpha))z'_{\lambda_2}(x_\alpha) \geq \varepsilon z'''(x_\alpha) - \sigma'(z_{\lambda_2}(x_\alpha))z'_{\lambda_2}(x_\alpha) = b(\lambda_2, u(x_\alpha), x_\alpha) \geq 0.$$

In particular, this implies $z_{\lambda_2}(x_\alpha) \in [\nu_1, \nu_2]$. Hence a necessary condition for the existence of a local positive maximum and a local negative minimum of z'_{λ_2} is the existence of $x \in [0,1]$ such that

$$z_{\lambda_2}(x) \geq \nu_1 > 1.$$

On the other hand, we have

$$0 > \varepsilon z''_{\lambda_1}(1) = \sigma(z_{\lambda_1}(1)), \tag{5.6}$$

and this implies, due to the assumed monotonicity,

$$z_{\lambda_1}(x) < 1 \text{ for every } x \in [0,1].$$

Of course, z_λ is connected with respect to the C^1-topology, and we obtain by virtue of Lemmas 17 and 18 the existence of $\hat{\lambda} \in (\lambda_1, \lambda_2)$ such that $z_{\hat{\lambda}}$ fulfills

$$z'_{\hat{\lambda}_{|(0,1)}} \geq 0 \text{ and there exists } \hat{x} \in [0,1] \text{ with } z_{\hat{\lambda}}(\hat{x}) = 1. \tag{5.7}$$

Due to $z'_{\hat{\lambda}} \geq 0$ we obtain $z''_{\hat{\lambda}}(1) < 0$ by (HGM). As in (5.6) we conclude that $z_{\hat{\lambda}}(1) < 1$, a contradiction to (5.7). ☐

Now we are in a position to state the main result of this section.

THEOREM 20. *Let $\varepsilon > 0$ be arbitrary but fixed and consider an arbitrary $z \in \mathcal{C}^+_\varepsilon$. Then we obtain three possibilities for the shape of z':*

(i) *The function z' has exactly one change in sign at $x = x_1$, and, moreover, we have $z'_{|(0,x_1)} < 0$ and $z'_{|(x_1,1)} > 0$. Furthermore, $z''(0) < 0$, $z''(x_1) > 0$, and $z''(1) < 0$ hold.*

(ii) *We have $z'(x) \leq 0$ for every $x \in [0,1]$, and if there exists an interval $J \subseteq [0,1]$ such that $z'_{|J} \equiv 0$, then $J = [x_\alpha, 1]$ for a certain $x_\alpha > x_0$.*

(iii) *For every $x \in (0,1)$ we obtain $z'(x) > 0$.*

Proof. By virtue of Corollary 14, the continuation method, and Lemmas 17, 18, and 19, we obtain that z'_λ changes sign at most once for every $\lambda \in \mathbb{R}^+_0$, and, moreover, (i) and (ii) of Theorem 20 are valid.

It remains to show (iii) of Theorem 20, and therefore we assume

$$z'(x) \geq 0 \quad \text{for every } x \in [0,1]$$

and, moreover, existence of $\hat{x} \in (0,1)$ such that $z'(\hat{x}) = 0$. (HGM) yields that we have no degenerate zero of z' in $[x_0, 1)$. If we assume $\hat{x} \in (0, x_0)$, we obtain by (H4) and (5.1) that

$$z'_{|(0,\gamma)} \text{ is not identically zero}$$

for some $\gamma > 0$. Hence, by our assumption on z' we get existence of a local maximum $x_2 \in (0, x_1)$ with $z'(x_2) > 0$. In a manner analogous to that in the proof of Lemma 19, we obtain a contradiction, which completes the proof of Theorem 20. ☐

6. Singular limit analysis. Let $z_{\varepsilon_n} \in \mathcal{C}^+_{\varepsilon_n}$ for every $n \in \mathbb{N}$. Our goal here is to pass to the limit for the sequence $(z_{\varepsilon_n})_{n\in\mathbb{N}}$ as $\varepsilon_n \searrow 0$. This is essentially done by Theorem 20 and Helly's theorem. Afterwards we show some qualitative properties of this limit.

LEMMA 21. *Consider for fixed $\lambda \in \mathbb{R}^+$ a sequence $(z_n)_{n\in\mathbb{N}}$ with $z_n \in \mathcal{C}^+_{\varepsilon_n}$ and $\varepsilon_n \searrow 0$. We obtain that $(z_n)_{n\in\mathbb{N}}$ is uniformly bounded in $BV(0,1)$, where $BV(0,1)$ denotes the space of functions with bounded variation defined on $(0,1)$.*

Proof. By Theorem 20, z'_n has at most one sign change, say at $x = x(n)$. Due to the fundamental theorem of calculus we obtain by Lemma 6

$$||z'_n||_{L^1} = \int_0^{x(n)} -z'_n(s)ds + \int_{x(n)}^1 z'_n(s)ds \leq 4\,||z_n||_{C^0} \leq K,$$

where K is independent of n. The same bound obviously holds for $BV(0,1)$. ☐

By Helly's theorem (see [23]) we conclude that the following theorem holds.

THEOREM 22. *For fixed $\lambda \in \mathbb{R}^+$ we consider a sequence $(z_n)_{n\in\mathbb{N}}$ as in Lemma 21. Then there exists a subsequence which we again denote by $(z_n)_{n\in\mathbb{N}}$ and a function $z \in L^\infty(0,1) \cap BV(0,1)$, such that*

$$\lim_{n\to\infty} z_n(x) = z(x) \tag{6.1}$$

for every $x \in [0,1]$. In particular, there exists $\delta > 0$ such that

$$z(x) \geq \delta \text{ for all } x \in [0,1]. \tag{6.2}$$

Moreover, we have for every $x \in [0,1]$

(i)

$$\lim_{n\to\infty} \varepsilon_n z_n''(x) = 0, \tag{6.3}$$

(ii)

$$\lim_{n\to\infty} \sigma(z_n(x)) = \sigma(z(x)) = \int_x^1 b(\lambda, u(\tau), \tau) d\tau, \tag{6.4}$$

where $u = \int_0^x z(s)ds$ is the limit of $(u_n)_{n\in\mathbb{N}}$ in $C^0(0,1)$. In particular, the first Weierstrass–Erdmann corner condition is fulfilled.

Proof. Let $\lambda \in \mathbb{R}^+$ be fixed and consider $(z_n)_{n\in\mathbb{N}}$ as in Lemma 21. Then (6.1) is a consequence of Lemma 21, (6.2) is obvious by Theorem 3, (6.3) is proved by Lemma 6 and (6.1), and (6.4) is obvious by (6.1). $\square$

Hence, the continuation method delivers a stress field

$$\sigma(x, z_n) \to \sigma(x, z) \quad \text{pointwise.}$$

Equation (6.4) expresses force-balance at $x \in [0,1]$. Observe that (6.4) may even be differentiated:

$$\frac{d}{dx}\sigma(x, z(x)) = b(\lambda, u(x), x) \quad \text{in } (0,1).$$

Moreover, z also satisfies the second Weierstrass–Erdmann corner condition. For the proof we refer to [10].

Corollary 23. *Let z be the limit obtained by Theorem 22. Then*

$$x \mapsto W(z(x)) - \sigma(z(x))z(x)$$

is continuous for every $x \in [0,1]$.

The next goal is to show that z jumps at most once in the unit interval. The first step in this direction is the observation that the last two cases in Theorem 20 cannot occur as $\varepsilon \searrow 0$.

Lemma 24. *Let $(z_n)_{n\in\mathbb{N}}$ be as in Lemma 21 for fixed $\lambda > 0$. Then we obtain for n sufficiently large*

$$\operatorname{sign}(z_n') \neq const.$$

Proof. We argue by contradiction: By assumption there exists a sequence $(z_n)_{n\in\mathbb{N}}$ with $\operatorname{sign}(z_n') \equiv const.$ for every $n \in \mathbb{N}$, and we assume without loss of generality $\operatorname{sign}(z_n') \geq 0$. We obtain by Helly's theorem

$$\lim_{n\to\infty} z_n(x) = z(x)$$

for every $x \in [0,1]$, and z is monotonically increasing and satisfies the equation

$$\sigma(z(x)) = \int_x^1 b(\lambda, u(\tau), \tau) d\tau$$

with $u(x) = \int_0^x z(s)ds$. By (H1), $\int_x^1 b(\lambda, u(\tau), \tau)d\tau_{|[0,x_0]}$ is monotonically decreasing, and by (2.4) and the monotonicity of z and b we obtain

$$z_{|[0,x_0]} \in [\nu_1, \nu_2] \subseteq (1, \infty).$$

On the other hand, we have $\sigma(z(x_0)) = \int_{x_0}^1 b(\lambda, u(\tau), \tau)d\tau < 0$ which yields

$$z(x_0) < 1,$$

a contradiction. ∎

In the remainder let $\lambda > 0$ be fixed. We consider a sequence $(\varepsilon_n)_{n\in\mathbb{N}}$ with $\varepsilon_n \searrow 0$ and let $(u_n)_{n\in\mathbb{N}} = ((u_\lambda)_n)_{n\in\mathbb{N}}$ be a sequence of functions on the branch $\mathcal{C}^+_{\varepsilon_n}$ with $u_n \to u$ uniformly in $C^0(0,1)$. Because of (H1) and (H2) on b we define for $\eta > 0$ sufficiently small

$$x_0 := \max\left\{\hat{x} \in [0,1] \mid b(\lambda, u_n(\hat{x}), \hat{x}) = 0 \text{ and } b(\lambda, u_n(x), x)_{|(\hat{x}-\eta,\hat{x})} > 0\right\}. \tag{6.5}$$

Note that by our assumptions x_0 is independent from $n \in \mathbb{N}$ for fixed λ and b changes sign at $x = x_0$.

Furthermore we define $x_s = x_s(n)$ by

$$\begin{aligned} x_s(n) := \Big\{\tilde{x} \in [0,1] \mid &\int_{\tilde{x}}^1 b(\lambda, u_n(\tau), \tau)d\tau = 0 \text{ and} \\ &\int_x^1 b(\lambda, u_n(\tau), \tau)d\tau > 0 \text{ for every } x < \tilde{x}\Big\}. \end{aligned} \tag{6.6}$$

Note that existence of $x_s = x_s(n)$ is by no means guaranteed. But if x_s exists, then $x_s < x_0$ holds. Furthermore we define

$$\tilde{x}_s(n) := \begin{cases} x_s(n) & \text{if } x_s \text{ exists,} \\ 0 & \text{elsewhere.} \end{cases}$$

Due to the obvious fact that $\tilde{x}_s(n) < x_0$, there exists $\varrho(n) > 0$ with $\tilde{x}_s(n) + \varrho(n) = x_0$. We define

$$\varrho_0 := \inf_{n\in\mathbb{N}} \{\varrho_n\}, \tag{6.7}$$

and it is easy to see that $\varrho_0 > 0$.

Lemma 25. *Let $z_\varepsilon \in \mathcal{C}^+_\varepsilon$. We claim the following: For every $\varrho \in (0, \varrho_0)$ there exists $\varepsilon_0 = \varepsilon_0(\varrho)$ such that for all $\varepsilon < \varepsilon_0$*

$$z'_{\varepsilon|(0,x_0-\varrho)} < 0.$$

Proof. We argue by contradiction and we assume the existence of $\varrho_1 \in (0, \varrho_0)$ such that for every $\varepsilon_0 > 0$ there exist $\varepsilon < \varepsilon_0$ and $x_\varepsilon \in (0, x_0 - \varrho_1)$ such that $z'_\varepsilon(x_\varepsilon) \geq 0$. By virtue of Theorem 20 and Lemma 24 we obtain $(z_n)_{n\in\mathbb{N}} \in \mathcal{C}^+_{\varepsilon_n}$ and $x_1 = x_1(n)$ with

$$z'_{n|(0,x_1(n))} < 0 \quad \text{and} \quad z'_{n|(x_1(n),1)} > 0.$$

Due to our assumption we deduce $z'_{n|[x_0-\varrho_1,1]} > 0$ for every $n \in \mathbb{N}$. Helly's theorem yields $z_{n|[x_0-\varrho_1,1]} \to z_{|[x_0-\varrho_1,1]}$ pointwise, and, in particular, $z_{|[x_0-\varrho_1,1]}$ is monotonically increasing. By virtue of (6.6) and (6.7) we obtain for every $x \in [x_0-\varrho_1, x_0]$

$$\sigma(z(x)) = \int_x^1 b(\lambda, u(\tau), \tau)d\tau \leq 0, \tag{6.8}$$

where $u(x) = \int_0^x z(s)ds$, and hence $z_{|[x_0-\varrho_1,x_0]} \leq 1$. Furthermore, $x \mapsto \int_x^1 b(\lambda, u(\tau), \tau)d\tau$ is monotonically decreasing for $x \in (x_0-\varrho_1, x_0)$. This yields either (i)

$$z_{|[x_0-\varrho_1,x_0]} \in (\nu_1, \nu_2)$$

or (ii)

$$z_{|[x_0-\varrho_1,x_0]} \equiv const.$$

Then (i) contradicts $z_{|[x_0-\varrho_1,x_0]} \leq 1$, and (ii) implies $\int_x^1 b(\lambda, u(\tau), \tau)d\tau = const.$ for every $x \in [x_0-\varrho_1, x_0]$, contradicting (6.5). $\square$

Let $(u_n)_{n\in\mathbb{N}} \in (\mathcal{C}^+_{\varepsilon_n})_n$ be a sequence for fixed $\lambda > 0$ with $\varepsilon_n \searrow 0$, and by virtue of Theorem 22 we obtain functions u, z such that $\lim_{n\to\infty} u_n =: u$ and $\lim_{n\to\infty} z_n =: z = u'$.

We define

$$\begin{aligned} \nu_a &:= \min\left\{\nu \in \mathbb{R}^+ \mid \sigma(\nu) = \sigma(\nu_2)\right\}, \\ \nu_b &:= \max\left\{\nu \in \mathbb{R}^+ \mid \sigma(\nu) = \sigma(\nu_1)\right\}, \end{aligned} \tag{6.9}$$

and

$$A := \left\{x \in [0,1] \mid \int_x^1 b(\lambda, u(\tau), \tau)d\tau \in \mathbb{R}\backslash[\sigma(\nu_a), \sigma(\nu_b)]\right\},$$

and due to the identity

$$\sigma(z(x)) = \int_x^1 b(\lambda, u(\tau), \tau)d\tau, \tag{6.10}$$

we have $A = \{x \in [0,1] \mid z(x) \in \mathbb{R}\backslash[\nu_a, \nu_b]\}$. Also by (6.10) we deduce that $z_{|A}$ is continuously differentiable.

Furthermore, we define

$$B := [0,1]\backslash A,$$

and $B \subseteq [0, x_0-\varrho_0)$ is obvious by (6.6), (6.7). We assume the existence of $\lambda \in \mathbb{R}^+$ and u obtained by the limit process such that

$$\int_{x_2}^1 b(\lambda, u(\tau), \tau)d\tau \in B \quad \text{for a certain } x_2 \in [0, x_0].$$

In particular, we have $\int_0^1 b(\lambda, u(\tau), \tau)d\tau > 0$. Thus $x_s(n) \in (0,1)$ defined in (6.6) exists for every $n \in \mathbb{N}$, and by virtue of Lemma 25 we have

$$z'_{n|(0,x_0-\frac{\varrho_0}{2})} < 0$$

for all $n \in \mathbb{N}$. By Helly's theorem we obtain that

$$z_{|B} \text{ is monotonically decreasing,}$$

where z denotes the pointwise limit of $(z_n)_{n\in\mathbb{N}}$. In particular, z has at most finitely many jumps in B. Moreover, we have that

$$\int_x^1 b(\lambda, u(\tau), \tau) d\tau_{|B} = \sigma(z)_{|B} \text{ is monotonically decreasing.}$$

Let $\nu_m, \nu_M \in \mathbb{R}^+$, $\nu_m < \nu_M$, be the Maxwell points associated with W. Then we obtain the following theorem.

THEOREM 26. *Let $(z_n)_{n\in\mathbb{N}}$ be a sequence with $z_n \in \mathcal{C}^+_{\varepsilon_n}$ for every $n \in \mathbb{N}$, $\varepsilon_n \searrow 0$ and for fixed $\lambda > 0$. Let z be the pointwise limit of the sequence. Furthermore, let $u(x) := \int_0^x z(s)ds$. Then z has at most one jump in $(0,1)$ in accordance with the first and the second Weierstrass–Erdmann corner conditions.*

If $\int_0^1 b(\lambda, u(\tau), \tau)d\tau < \sigma(\nu_m)$, then z has no jump, and if $\int_0^1 b(\lambda, u(\tau), \tau)d\tau > \sigma(\nu_1)$, then z suffers exactly one jump discontinuity at $x_J \in (0, x_0)$ satisfying

$$\int_{x_J}^1 b(\lambda, u(\tau), \tau) d\tau = \sigma(\nu_m). \tag{6.11}$$

Moreover, we obtain $z(x_J - 0) = \nu_M$ and $z(x_J + 0) = \nu_m$.

If $\int_0^1 b(\lambda, u(\tau), \tau)d\tau \in [\sigma(\nu_m), \sigma(\nu_1)]$, then z has either no jump or exactly one at $x = x_J$ such that (6.11) holds.

Proof. The proof is obvious due to the fact that z satisfies the first and the second Weierstrass–Erdmann corner conditions and, moreover, z is monotonically decreasing in B. Note that z cannot jump in the region where z is monotonically increasing. ∎

We define for some $q < \infty$ the set

$$\Sigma_0^+ = \Big\{(\lambda, u) \in \mathbb{R} \times W^{1,q}(0,1) \mid (\lambda_n, u_n, v_n, \mu_n) \in \bar{\mathcal{C}}^+_{\varepsilon_n}, \lambda_n \to \lambda, u_n \to u \\ \text{in } W^{1,q}(0,1) \text{ as } \varepsilon_n \searrow 0\Big\}. \tag{6.12}$$

Then one can prove the following theorem in the same way as in [10] using [1].

THEOREM 27. *The set $\Sigma_0^+ \subseteq \mathbb{R} \times W^{1,q}(0,1)$ is a continuum having projection $[0, \infty)$ onto the parameter axis.*

The question of whether the obtained solution of the Euler–Lagrange equation (2.13) is also the global minimizer of the nonconvex variational problem (2.8) is answered only in a very special case: We consider a body-force $b = b(\lambda, x)$ independent of the placement u and let z_* be the jump solution obtained by the singular limit process if $\int_0^1 b(\lambda, \tau)d\tau > \sigma(\nu_m)$, and a continuous differentiable solution otherwise. Then we deduce the following theorem.

THEOREM 28. *Let (λ, z_*) be as described above. Then $J(\lambda, u)$ defined by (2.8) attains its minimum at $u_* := \int_0^x z_*(s)ds$; i.e.,*

$$\min_u \big\{J(\lambda, u) \mid u \in W^{1,p+1}(0,1),\, u(0) = 0\big\} = J(\lambda, u_*).$$

We omit the proof because it is completely analogous to the proof of Theorem 6.4 in [10].

7. Concluding remark. By our continuation method we were not able to prove that a solution of (2.13) lying on the branch is the global minimizer of (2.10). Hence, in general, we cannot prove that the limit u obtained in Theorem 22 is a global minimizer of (2.8). However, our results give some hints in this direction, because u satisfies the first and second Weierstrass–Erdmann corner conditions, and, moreover, u' is strictly bounded away from 0. Note that the existence of a global minimizer of (2.8) in the live load case is entirely unclear, because [6] cannot be applied to this problem.

Acknowledgments. I am indebted to Prof. H. Kielhöfer and Prof. T. Healey for helpful hints and discussions and to the referees for their valuable comments.

REFERENCES

[1] J. C. Alexander, *A primer on connectivity*, in Fixed Point Theory, E. Fadell and G. Fournier, eds., Springer-Verlag, New York, 1981, pp. 455–483.

[2] N. D. Alikakos and P. W. Bates, *On the singular limit in a phase field model of phase transitions*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 5 (1988), pp. 141–178.

[3] S. S. Antman and H. Brezis, *The existence of orientation-preserving deformations in nonlinear elasticity*, in Nonlinear Analysis and Mechanics: Heriot–Watt Symposium, Vol. II, R. J. Knops, ed., Pitman, Boston, London, 1978, pp. 1–29.

[4] J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1977), pp. 337–403.

[5] J. Carr, M. E. Gurtin, and M. Slemrod, *Structured phase transitions on a finite interval*, Arch. Rational Mech. Anal., 86 (1984), pp. 317–351.

[6] A. Cellina and G. Colombo, *On a classical problem of the calculus of variations without convexity assumptions*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 97–106.

[7] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, 1989.

[8] J. L. Ericksen, *Equilibrium of bars*, J. Elasticity, 5 (1975), pp. 191–201.

[9] T. J. Healey, *Global continuation in displacement problems of nonlinear elastostatics via the Leray–Schauder degree*, Arch. Ration. Mech. Anal., 152 (2000), pp. 273–282.

[10] T. J. Healey and H. Kielhöfer, *Global continuation via higher-gradient regularization and singular limits in forced one-dimensional phase transitions*, SIAM J. Math. Anal., 31 (2000), pp. 1307–1331.

[11] T. J. Healey and P. Rosakis, *Unbounded branches of classical injective solutions to the forced displacement problem in nonlinear elastostatics*, J. Elasticity, 49 (1997), pp. 65–78.

[12] T. J. Healey and H. C. Simpson, *Global continuation in nonlinear elasticity*, Arch. Rational Mech. Anal., 143 (1998), pp. 1–28.

[13] R. D. James, *Co-existent phases in the one-dimensional static theory of elastic bars*, Arch. Rational Mech. Anal., 72 (1979/80), pp. 99–140.

[14] H. Kielhöfer, *Pattern formation of the stationary Cahn–Hilliard equation model,* Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 1219–1243.

[15] H. Kielhöfer, *Minimizing sequences selected via singular perturbations, and their pattern formation*, Arch. Ration. Mech. Anal., 155 (2000), pp. 261–276.

[16] H. Kielhöfer, *Critical points of nonconvex and noncoercive functionals*, Calc. Var. Partial Differential Equations, 16 (2003), pp. 243–272.

[17] H. Kielhöfer, *Bifurcation Theory. An Introduction with Applications to PDEs*, Appl. Math. Sci. 156, Springer-Verlag, New York, 2004.

[18] R. V. Kohn and S. Müller, *Surface energy and microstructure in coherent phase transitions*, Comm. Pure Appl. Math., 47 (1994), pp. 405–435.

[19] S. Krömer and M. Lilli, *Branches of positive solutions of quasilinear elliptic equations on non-smooth domains*, Nonlinear Anal., 64 (2006), pp. 2183–2202.

[20] M. Lilli, *The Effect of a Singular Perturbation to a* 1*D Nonconvex Variational Problem*, Augsburger Schriften zur Mathematik, Physik und Informatik 10, Logos Verlag, Berlin, 2005.

[21] M. Lilli, T. J. Healey, and H. Kielhöfer, *Singular perturbation as a selection criterion for Young-measure solutions*, SIAM J. Math. Anal., 39 (2007), pp. 195–209.

[22] S. Müller, *Singular perturbations as a selection criterion for periodic minimizing sequences*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 169–204.

[23] I. P. Natanson, *Theory of Functions of a Real Variable*, Frederick Ungar, New York, 1955.
[24] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Springer-Verlag, New York, 1984.
[25] P. H. Rabinowitz, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.

SIAM J. MATH. ANAL.
Vol. 39, No. 3, pp. 819–837

OBSTACLE AND BOUNDARY DETERMINATION FROM SCATTERING DATA*

GEN NAKAMURA† AND MOURAD SINI‡

Abstract. In this paper, we are concerned with the identification of complex obstacles from the scattering data for the acoustic problem. The complex obstacle is characterized by its shape and the boundary values of the impedance coefficient. We establish pointwise formulas which can be used to reconstruct the shape of the obstacle and give explicitly the values of the surface impedance as a function of the far field. In addition, these formulas enable us to distinguish and recognize the coated and the noncoated parts of the obstacle.

Key words. inverse scattering, far-field, impedance boundary, singularity analysis, numerics

AMS subject classifications. 35P25, 35R30, 78A45

DOI. 10.1137/060658667

1. Introduction.

Let D be a bounded domain of $\mathbb{R}^m$, $m \geq 3$, such that $\mathbb{R}^m \setminus \overline{D}$ is connected. In addition, we assume that its boundary ∂D is of class C^2. Precisely, for every point $a \in \partial D$, there exists a rigid transformation of coordinates under which $a = 0$ and a $C^2(B_{m-1}(0,r))$-function f such that

$$f(0) = \frac{\partial f}{\partial x_i}(0) = 0, \quad i = 1, 2, \ldots, m, \tag{1}$$

and

$$D \cap B(0,r) := \{x \in B(0,r); x_m > f(x_1, x_2, \ldots, x_{m-1})\}$$

in terms of the new coordinates where $B_{m-1}(0,r)$ and $B(0,r)$ are the $(m-1)$-dimensional and the m-dimensional balls of centers 0 with some radius $r > 0$.

The propagation of time-harmonic acoustic fields in a homogeneous media is governed by the *Helmholtz equation*

$$\Delta u + \kappa^2 u = 0 \quad \text{in} \quad \mathbb{R}^m \setminus \overline{D}, \tag{2}$$

where κ is the real positive *wave number*. At the boundary of the scatterers we assume that the total field u satisfies the *impedance boundary condition*

$$\frac{\partial u}{\partial \nu} + i\lambda u = 0 \text{ on } \partial D_I \tag{3}$$

with some function λ on ∂D and the *Dirichlet condition*

$$u = 0 \text{ on } \partial D_D, \tag{4}$$

where ∂D_D and ∂D_I are open surfaces in ∂D such that

$$\partial D = \overline{\partial D_I} \cup \overline{\partial D_D} \text{ and } \partial D_I \cap \partial D_D = \emptyset.$$

*Received by the editors May 2, 2006; accepted for publication (in revised form) April 9, 2007; published electronically August 24, 2007.

http://www.siam.org/journals/sima/39-3/65866.html

†Department of Mathematics, Hokkaido University, Sapporo, 060-0810, Japan (gnaka@math.sci.hokudai.ac.jp). This author was partially supported by the Japan Society for Promotion of Science.

‡Corresponding author. RICAM, Altenbergerstrasse 69, Linz, A-4040, Austria (mourad.sini@oeaw.ac.at). This author was supported by Impedance Imaging Research Center of Kyung Hee University, Korea, via MOST/KOSEF (R11-2002-103) and the SFB-project F1308 at RICAM, Austria.

The unit normal ν on ∂D is directed inside D. We assume that λ is a Hölder continuous function of order $\beta \in (0,1]$ and $\lambda_- < \lambda(x)$ on ∂D_I, where λ_- is a positive constant. The part ∂D_I is referred to by the coated part of ∂D, and ∂D_D is the noncoated part as it is commonly used in the radar detection theory; see [4]. The obstacle D is characterized by its shape, ∂D_D, ∂D_I, and the surface impedance distributed on ∂D_I. We call such obstacles *complex obstacles*.

Given an incident field u^i which satisfies $\Delta u^i + \kappa^2 u^i = 0$ we look for a solution $u := u^i + u^s$ of (2) and (3) where the *scattered field* u^s is assumed to satisfy the Sommerfeld radiation condition

$$\lim_{r\to\infty} r^{\frac{m-1}{2}} \left(\frac{\partial u^s}{\partial r} - i\kappa u^s \right) = 0, \tag{5}$$

where $r = |x|$ and the limit is uniform with respect to all of the directions $\hat{x} := \frac{x}{|x|}$.

The mixed problem (2)–(5) is well posed. More generally, for $f \in H^{\frac{1}{2}}(\partial D_D)$ and $h \in H^{-\frac{1}{2}}(\partial D_I)$, there exists a unique solution $u \in H^1_{loc}(R^m \setminus \overline{D})$ of the mixed problem

$$\begin{cases} (\Delta + \kappa^2)u = 0 \quad \text{in } R^m \setminus \overline{D}, \\ u = f \quad \text{on } \partial D_D, \\ \frac{\partial u}{\partial \nu} + i\lambda u = h \quad \text{on } \partial D_I, \\ \lim_{r\to\infty} r^{\frac{m-1}{2}} (\frac{\partial u}{\partial r} - i\kappa u) = 0, \end{cases} \tag{6}$$

and the solution satisfies

$$\|u\|_{H^1(\Omega_R \cap (R^m \setminus \overline{D}))} \leq C_R(\|f\|_{H^{1/2}(\partial D_D)} + \|h\|_{H^{-\frac{1}{2}}(\partial D_I)}), \tag{7}$$

where Ω_R is a disk of radius R; and C_R is positive constant depending on R, see [4] for more details.

It is well known (see [6]) that this reflected field satisfies the following asymptotic property:

$$u^s(x) = \frac{e^{i\kappa r}}{r^{\frac{m-1}{2}}} u^\infty(\hat{x}) + O(r^{-\frac{m+1}{2}}), \quad r \to \infty, \tag{8}$$

where the function $u^\infty(\cdot)$ defined on the unit sphere $\mathbb{S}^m$ of R^m is called the far field associated to the incident field u^i. Taking particular incident fields given by the plane waves, $u^i(x,d) := e^{i\kappa d\cdot x}$, $d \in \mathbb{S}^m$, we define the far-field pattern $u^\infty(\hat{x}, d)$ for $(\hat{x}, d) \in \mathbb{S}^m \times \mathbb{S}^m$. Analogously, for an incident point source $\Phi(\cdot, z)$, where

$$\Phi(x,y) := \frac{1}{(m-2)\sigma_m} \frac{e^{i\kappa|x-y|}}{|x-y|^{m-2}}, \quad x \neq y, x, y \in \mathbb{R}^m,$$

is the fundamental solution of $\Delta + \kappa^2$ in $\mathbb{R}^m$ and σ_m is the surface of the unit sphere in $\mathbb{R}^m$, we denote the scattered field by $\Phi^s(\cdot, z)$ and its far-field pattern by $\Phi^\infty(\cdot, z)$. The problem we are concerned with is the following.

Complex obstacle reconstruction problem. *Given* $u^\infty(\cdot,\cdot)$ *on* $\mathbb{S}^m \times \mathbb{S}^m$ *for the scattering problem* (2)–(5) *reconstruct the shape of the obstacle* D*, distinguish the coated part* ∂D_D *from* ∂D_I*, and reconstruct the surface impedance* $\lambda(x)$*.*

The uniqueness character of this problem is already known; see [4]. The part of the problem consisting of the effective detection of the shape of the obstacle ∂D can be justified, for instance, via the linear sampling method, the factorization method, or the probing methods (the probe method or, equivalently, the singular sources method); see [16, 17] for a review of these methods. Our goal in this paper is to show that not only the shape but the full complex obstacle can be reconstructed. Many efforts have been made regarding the determination of the surface impedance function $\lambda(x)$. We refer to the paper [11] where an optimization method has been proposed by assuming that the shape of the obstacle is known in advance. A different method is given in [7], where the authors first reduce the far-field data to the near-field data, and then from these near-field data they propose a moment method to reconstruct λ. Another work is [5]; see also [4], where the authors computed the L^∞-norm of λ. As a consequence, if λ is known to be constant $\lambda = \lambda_0$, then they compute λ_0. All of these works use a part of or the total far field. We mention the work [2], where the authors use only one incident wave to detect $\lambda(x)$. Assuming that the whole surface ∂D is coated and known, they first compute the total field and then use the impedance boundary condition to give the values of $\lambda(x)$. By the unique continuation, there is no open subset of ∂D in which the normal derivative of the total field may vanish. However, there can be infinitely many points in ∂D at which the total field vanishes. By avoiding these points, it is possible to reconstruct the value of $\lambda(x)$, and then by $\lambda \in C(\partial D)$, it is possible to know λ on the whole ∂D. Hence this method cannot sample each point x of ∂D to obtain the value $\lambda(x)$. To remedy this difficulty, the authors propose a regularization method.

We want to contribute to this problem by giving pointwise formulas to reconstruct fully the complex obstacle. Indeed, these formulas simultaneously reconstruct the shape of the obstacle, distinguish between the coated and the noncoated parts, and detect $\lambda(x)$ directly from the far-field pattern defined on any small open part of the unit sphere $\mathbb{S}^m$.

To justify our formulas, we need to analyze the asymptotic behavior of the Green's function, of the mixed boundary problem, near ∂D. The impedance function $\lambda(x)$ appears in the asymptotic behavior of the imaginary part of this Green's function with respect to the source parameter z; see the proof of Proposition 3.1. In the 2-dimensional case the imaginary part of the corresponding Green's function is bounded with respect to the source parameter z. This is why we consider the m-dimensional case with $m > 2$. For the 2-dimensional case, we need to use more singular sources to capture the values of the surface impedance. This has been analyzed in [13], and the corresponding formulas have been justified theoretically and tested numerically. We refer to that paper for more details on how the formulas are used numerically.

Regarding the stability issue for detecting the surface impedance, in case we know the shape, we cite the results of [12, 18], where the authors use one incident wave and give interesting and optimal results. Another interesting question is to consider the stability of the complex obstacle. We leave this for future investigations.

The rest of the paper is organized as follows. In section 2, we present the results as Theorem 2.1. In section 3, we give the proof of this theorem by splitting it into two propositions, which we prove in sections 3.1 and 3.2.

2. Presentation of the results. It is well known (see [6]) that the scattered field associated with the Herglotz incident field $v_g^i := v_g$ defined by

$$v_g(x) := \int_{\mathbb{S}^m} e^{i\kappa x\cdot d} g(d)\, ds(d), \quad x \in \mathbb{R}^m, \tag{9}$$

with $g \in L^2(\mathbb{S}^m)$, is given by

$$v_g^s(x) := \int_{\mathbb{S}^m} u^s(x,d) g(d)\, ds(d), \quad x \in \mathbb{R}^m \setminus D, \tag{10}$$

and its far field is given by

$$v_g^\infty(\hat{x}) := \int_{\mathbb{S}^m} u^\infty(\hat{x}, d) g(d)\, ds(d), \quad \hat{x} \in \mathbb{S}^m. \tag{11}$$

We will need the following identity:

$$u^\infty(\hat{x}, d) = -\frac{1}{(m-2)\sigma_m} \int_{\partial D} \left\{ \frac{\partial u^s(y,d)}{\partial \nu} e^{-i\kappa \hat{x}\cdot y} - \frac{\partial e^{-i\kappa\hat{x}\cdot y}}{\partial \nu} u^s(y,d) \right\} ds(y) \tag{12}$$

given by using the Green's formula in $\mathbb{R}^m \setminus \overline{D}$ for $u^s(\cdot, d)$ and $\Phi(\cdot, y)$ and their asymptotic behavior at infinity (see [6, Theorem 2.5]). The representation of the scattered field $\Phi^s(x,z)$ for $x, z \in \mathbb{R}^m \setminus \overline{D}$ is given by the Green's formula

$$\Phi^s(x,z) = -\int_{\partial D} \left\{ \frac{\partial \Phi^s(y,z)}{\partial \nu(y)} \Phi(x,y) - \Phi^s(y,z) \frac{\partial \Phi(x,y)}{\partial \nu(y)} \right\} ds(y), \quad x, z \in \mathbb{R}^3 \setminus \overline{D}. \tag{13}$$

Let $a \in \partial D$ and a sequence of points

$$(z_p)_{p\in\mathbb{N}} \subset \mathbb{R}^m \setminus \overline{D} \tag{14}$$

such that z_p tends to a. We consider the sequence of point sources $\Phi(\cdot, z_p)$. We set D_a^p a C^2-regular open set such that $\overline{D} \subset D_a^p$, $z_p \in \mathbb{R}^n \setminus \overline{D_a^p}$ for every $p \in \mathbb{N}$ and that the Dirichlet interior problem on D_a^p is uniquely solvable. In this case, the Herglotz wave operator H defined from $L^2(\mathbb{S}^m)$ to $L^2(\partial D_a^p)$ by

$$H(g)(x) := v_g(x) = \int_{\mathbb{S}^m} e^{i\kappa x\cdot d} g(d)\, ds(d) \tag{15}$$

is injective, compact with a dense range; see [6]. Hence by the Tikhonov regularization method (see [8, 10]) we can construct a sequence g_n^p in $L^2(\mathbb{S}^m)$ such that for every p fixed

$$\|v_{g_n^p} - \Phi(\cdot, z_p)\|_{L^2(\partial D_a^p)} \to 0, \quad n \to \infty. \tag{16}$$

Since both $v_{g_n^p}$ and $\Phi(\cdot, z_p)$ satisfy the same Helmholtz equation in D_a^p, (16) implies that

$$\|v_{g_n^p} - \Phi(\cdot, z_p)\|_{H^{\frac{1}{2}}(\partial D)} \to 0, \quad n \to \infty.$$

Multiplying (12) by $g_n^p(\hat{x}) g_n^p(d)$ and integrating over $\mathbb{S}^m$, we have

$$\begin{aligned}
&\int_{\mathbb{S}^m}\int_{\mathbb{S}^m} u^\infty(-\hat{x}, d) g_n^p(\hat{x}) g_n^p(d)\, ds(\hat{x}) ds(d) \\
&= -\frac{1}{(m-2)\sigma_m} \int_{\partial D} \left\{ \int_{\mathbb{S}^m} \frac{\partial u^s(y,d)}{\partial \nu} g_n^p(d)\, ds(d) \cdot \int_{\mathbb{S}^m} e^{i\kappa\hat{x}\cdot y} g_n^p(\hat{x})\, ds(\hat{x}) \right. \\
&\quad \left. - \int_{\mathbb{S}^m} \frac{\partial e^{i\kappa\hat{x}\cdot y}}{\partial \nu} g_n^p(\hat{x})\, ds(\hat{x}) \cdot \int_{\mathbb{S}^m} u^s(y,d) g_n^p(d)\, ds(d) \right\} ds(y) \\
&= -\frac{1}{(m-2)\sigma_m} \int_{\partial D} \left\{ \frac{\partial v^s_{g_n^p}}{\partial \nu}(y) v^i_{g_n^p}(y) + \frac{\partial v^i_{g_n^p}}{\partial \nu}(y) v^s_{g_n^p}(y) \right\} ds(y). \qquad (17)
\end{aligned}$$

From (17), we have

$$\lim_{n\to\infty}\int_{\mathbb{S}^m}\int_{\mathbb{S}^m} u^\infty(-\hat{x},d)\; g_n^p(\hat{x})\; g_n^p(d)\; ds(\hat{x})ds(d)$$

$$= -\frac{1}{(m-2)\sigma_m}\int_{\partial D}\frac{\partial(\Phi^s)}{\partial\nu(y)}(y,z_p)\Phi(y,z_p) - \Phi^s(y,z_p)\frac{\partial(\Phi)}{\partial\nu(y)}(y,z_p)\; ds(y) \tag{18}$$

$$= \frac{1}{(m-2)\sigma_m}\Phi^s(z_p,z_p). \tag{19}$$

For a point $a\in\partial D$, we choose the sequence $(z_p)_{p\in\mathbb{N}}$ included in $C_{a,\theta}$, where $C_{a,\theta}$ is the cone with center a, angle $\theta\in[0,\frac{\pi}{2})$, and axis $-\nu(a)$, where $\nu(a)$ is the unit normal of ∂D directed inside D.

Theorem 2.1. *The surface of the obstacle can be localized via the formulas:*

$$\left|32\pi^2\lim_{n\to\infty}\mathrm{Re}\int_{\mathbb{S}^m}\int_{\mathbb{S}^m} u^\infty(-\hat{x},d)g_n^p(\hat{x})g_n^p(d)ds(x)ds(d)\right|$$

$$= \frac{1}{|(z_p-a)\cdot\nu(a)|} + O(|\ln(|z_p-a|)|) \tag{20}$$

for $m=3$ and

$$\left|\frac{(m-2)\sigma_m 2^{2m-2}\pi^{m-1}}{\sigma_{m-1}}\lim_{n\to\infty}\mathrm{Re}\int_{\mathbb{S}^m}\int_{\mathbb{S}^m} u^\infty(-\hat{x},d)g_n^p(\hat{x})g_n^p(d)ds(x)ds(d)\right|$$

$$= \frac{1}{|(z_p-a)\cdot\nu(a)|^{m-2}} + O(|z_p-a|^{3-m}) \tag{21}$$

for $m>3$.

In addition, we have the following formulas for distinguishing the coated part from the noncoated part of the obstacle and for detecting the surface impedance.

I. *The case $m=3$.*

I.1. *If $a\in\partial D_I$, then we have*

$$\lim_{p\to\infty} -8\pi^2\frac{\lim_{n\to\infty}\mathrm{Im}\int_{\mathbb{S}^m}\int_{\mathbb{S}^m} u^\infty(-\hat{x},d)g_n^p(\hat{x})g_n^p(d)\; ds(\hat{x})ds(d)}{|\ln|(z_p-a)\cdot\nu(a)||^s}$$

$$= \begin{cases}\infty \;\; \textit{if } s\in[0,1),\\ \lambda(a) \;\; \textit{if } s=1,\\ 0 \;\; \textit{if } s>1.\end{cases} \tag{22}$$

I.2. *If $a\in\partial D_D$, then for all $s>0$ we have*

$$\lim_{p\to\infty}\frac{\lim_{n\to\infty}\mathrm{Im}\int_{\mathbb{S}^m}\int_{\mathbb{S}^m} u^\infty(-\hat{x},d)g_n^p(\hat{x})g_n^p(d)\; ds(\hat{x})ds(d)}{|\ln|(z_p-a)\cdot\nu(a)||^s} = 0. \tag{23}$$

II. *The case $m>3$.*

II.1. *If $a\in\partial D_I$, then we have*

(24)

$$\lim_{p\to\infty}\frac{(m-2)\sigma_m(2\pi)^{m-1}}{\sigma_{m-1}}(2(z_p-a)\cdot\nu(a))^{m-4+s}$$

$$\lim_{n\to\infty}\mathrm{Im}\int_{\mathbb{S}^m}\int_{\mathbb{S}^m} u^\infty(-\hat{x},d)g_n^p(\hat{x})g_n^p(d)\; ds(\hat{x})ds(d) = \begin{cases}\infty \;\; \textit{if } s\in[0,1),\\ \lambda(a) \;\; \textit{if } s=1,\\ 0 \;\; \textit{if } s>1.\end{cases}$$

II.2. *If $a \in \partial D_D$, then for all $s > 0$ we have*

$$\lim_{p\to\infty}((z_p - a)\cdot\nu(a))^{m-3+s}$$
$$\lim_{n\to\infty}\mathrm{Im}\int_{\mathbb{S}^m}\int_{\mathbb{S}^m} u^\infty(-\hat{x},d)g_n^p(\hat{x})g_n^p(d)\,ds(\hat{x})ds(d) = 0. \tag{25}$$

Remark 2.2. 1. In the case $m = 3$, from (22) and (23), we can localize the coated part of the obstacle by taking any $s \in (0,1)$, then taking $s = 1$ in (22), we obtain the pointwise values of the surface impedance. Similarly, we have the same conclusions for $m \geq 4$.

2. We stated the results by using the full far-field pattern, i.e., $(\theta, d) \in \mathbb{S}^m \times \mathbb{S}^m$. We used this information to approximate the point sources $\Phi(\cdot, z_p)$ by Herglotz functions defined on the whole unit sphere $\mathbb{S}^m$. However, this approximation is also justified if we define the Herglotz functions on any subsurface $\gamma \subset \mathbb{S}^m$, and hence the results of Theorem 2.1 are also valid if we replace $\mathbb{S}^m$ by any subsurface $\gamma \subset \mathbb{S}^m$.

3. Proof of Theorem 2.1. Let $\Gamma_{\lambda(a)}$ be a local Green's function satisfying

$$\begin{cases} \Delta\Gamma_{\lambda(a)} = -\delta(x,z) \text{ in } \mathbb{R}^m_+, \\ \left(\frac{\partial\Gamma_{\lambda(a)}}{\partial\nu} + i\lambda(a)\Gamma_{\lambda(a)}\right)(x_1, x_2, \dots, x_{m-1}, 0) = 0 \end{cases} \tag{26}$$

and Γ_D defined by

$$\Gamma_D(x,z) := \Gamma(x,z) - \Gamma(x^*, z),$$

where $x = (x_1, x_2, \dots, x_m)$, $x^* = (x_1, x_2, \dots, -x_m)$, and $\Gamma(x,z) = \frac{1}{(m-2)\sigma_m|x-z|^{m-2}}$.

It is clear that $\Gamma_D(x,z)$ satisfies

$$\begin{cases} \Delta\Gamma_D = -\delta(x,z) \text{ in } \mathbb{R}^m_+, \\ \Gamma_D(x_1, x_2, \dots, x_{m-1}, 0) = 0. \end{cases} \tag{27}$$

We state the following propositions. Their proofs will be given in sections 3.1 and 3.2, respectively.

Proposition 3.1. *The local Green function $\Gamma_{\lambda(a)}$ is given by*

$$\Gamma_{\lambda(a)}(x,z) := \Gamma(x,z) + \frac{1}{2(2\pi)^{m-1}}\int_{\mathbb{R}^{m-1}} e^{i(x'-z')\cdot\xi'}e^{-(x_m+z_m)|\xi'|}\frac{|\xi'| + i\lambda(a)}{|\xi'|(|\xi'| - i\lambda(a))}d\xi', \tag{28}$$

where $x' = (x_1, x_2, \dots, x_{m-1})$ and $\xi' = (\xi_1, \xi_2, \dots, \xi_{m-1})$.

In addition, we have the following asymptotics for the function $(\Gamma_{\lambda(a)} - \Gamma)(x,z)$:

$$\mathrm{Re}(\Gamma_{\lambda(a)} - \Gamma)(z,z) = \frac{\sigma_{m-1}}{2(2\pi)^{m-1}}\frac{1}{(2z_m)^{m-2}} + O\left(\frac{1}{(z_m)^{m-3}}\right),$$

$$\lim_{z_3\to 0^+} -2\pi\frac{\mathrm{Im}(\Gamma_{\lambda(a)} - \Gamma)(z,z)}{\ln(z_3)} = \lambda(a) \text{ for } m = 3$$

and

$$\lim_{z_m\to 0^+}\frac{(2\pi)^{m-1}}{\sigma_{m-1}}(2z_m)^{m-3}\mathrm{Im}(\Gamma_{\lambda(a)} - \Gamma)(z,z) = \lambda(a) \text{ for } m > 3.$$

Proposition 3.2. *If $a \in \partial D_I$, then there exist $\delta(a) > 0$ and $C > 0$ such that*

$$|\Phi^s(x,z)-(\Gamma_{\lambda(a)}-\Gamma)(x,z)| \le C\begin{cases} |\ln|z-a|| & \text{if } m=3 \\ \frac{1}{|z-a|^{m-3}} & \text{if } m>3 \end{cases} \quad \text{for } (x,z)\in B_+(a,\delta(a))\cap C_{a,\theta}, \tag{29}$$

and

$$|\mathrm{Im}(\Phi^s(x,z)-(\Gamma_{\lambda(a)}-\Gamma)(x,z))| \le \frac{C}{|z-a|^{m-3-\beta}}+C \quad \text{for } m\ge 3 \text{ and } (x,z)\in B_+(a,\delta(a))\cap C_{a,\theta}, \tag{30}$$

where $B_+(a,\delta(a)) := B(a,\delta(a))\cap(\mathbb{R}^3\setminus D)$ and $B(a,\delta(a))$ is the ball of center a and radius $\delta(a)$.

Similarly, if $a\in\partial D_D$, we obtain (29) *and* (30) *by replacing $\Gamma_{\lambda(a)}$ by Γ_D.*

End of the proof of Theorem 2.1.

1. Let $a\in\partial D_I$; i.e., we have the impedance boundary condition around a. By a rigid transformation of coordinates, we can assume that $a=(0,0,0,\dots,0)$. Using (19) and Propositions 3.1 and 3.2 we obtain the formulas (20), (21), (22) and (24).
2. Let $a\in\partial D_D$; i.e., we have the Dirichlet boundary condition around a. Similarly, we can assume that $a=(0,0,0)$. Using (19), Proposition 3.2, and the fact that $\mathrm{Im}\Gamma_D = 0 = \mathrm{Im}\Gamma$ we obtain (23) and (25). ∎

The rest of this section is devoted to proving Propositions 3.1 and 3.2.

3.1. Proof of Proposition 3.1. We set $\Gamma_{\lambda(a)}(x,z) := \Gamma(x,z)+w(x,z)$, and then $w(x,z)$ satisfies

$$\begin{cases} \Delta w(x,z)=0 \text{ in } \mathbb{R}^m_+, \\ (\partial_{x_3}+i\lambda(a))w(x,z) = -(\partial_{x_3}+i\lambda(a))\Gamma(x,z) \text{ on } \partial\mathbb{R}^m_+. \end{cases} \tag{31}$$

The first part of this proposition is to show the following explicit form of $w(x,z)$.

Lemma 3.3.

$$w(x,z)=\frac{1}{2(2\pi)^{m-1}}\int_{\mathbb{R}^{m-1}} e^{i(x'-z')\cdot\xi'}e^{-(x_m+z_m)|\xi'|}\frac{|\xi'|+i\lambda(a)}{|\xi'|(|\xi'|-i\lambda(a))}d\xi'.$$

Proof of Lemma 3.3. We represent $w(x,z)$ using up-going and down-going operators $U_\pm$

$$w(x,z) := (U_\pm(x_m)\phi)(x') := \frac{1}{(2\pi)^{m-1}}\int_{\mathbb{R}^{m-1}} e^{ix'\cdot\xi'\mp x_m|\xi'|}\hat{\phi}_\pm(\xi',z)d\xi', \tag{32}$$

where $\hat{\phi}_\pm$ is the $(m-1)$-dimensional Fourier transform of $\phi_\pm$. The goal is to find $\phi_\pm$ or $\hat{\phi}_\pm$. We start by the corresponding representation of $\Gamma(x,z)$. We write

$$\Gamma(x,z)=\begin{cases} \Gamma_+(x,z) \text{ in } x_m>z_m, \\ \Gamma_-(x,z) \text{ in } x_m<z_m, \end{cases} \tag{33}$$

and then $\Delta\Gamma_\pm = 0$ in $\pm(x_m-z_m)>0$ with the transmission conditions

$$\begin{cases} \Gamma_+|_{x_m=z_m+0} = \Gamma_-|_{x_m=z_m-0}, \\ \partial_{x_m}\Gamma_+|_{x_m=z_m+0} - \partial_{x_m}\Gamma_-|_{x_m=z_m-0} = -\delta(x',z'). \end{cases} \tag{34}$$

Now we look for $\Gamma_\pm$ in the form

$$\Gamma_\pm(x,z) = U_\pm(x_m - z_m)\psi_\pm(x', z')$$

and try to determine the potentials $\psi_\pm$.

Clearly, from the definition of $U_\pm$, we have

$$\Delta\Gamma_\pm = 0 \text{ in } \pm(x_m - z_m) > 0,$$

and from the first equation of (34), we get

$$\psi_+ = \psi_-. \tag{35}$$

Let us now consider the second equation of (34). We set

$$(Bf)(x') := \int_{\mathbb{R}^{m-1}} e^{ix'\cdot\xi'}(-|\xi|)\hat{f}(\xi')\bar{d}\xi',$$

and then we deduce that

$$\partial_{x_m} U_\pm(x_m) = \pm B U_\pm(x_m). \tag{36}$$

The second point of (34) implies that

$$B\psi_+ + B\psi_- = -\delta(x' - z').$$

Taking the Fourier transform, we have

$$-|\xi'|\hat{\psi}_+ - |\xi'|\hat{\psi}_- = -e^{iz'\cdot\xi'}, \tag{37}$$

and by combining (35) with (37), we end up with

$$\hat{\psi}_\pm(\xi', z') = \frac{1}{2}|\xi'|^{-1}e^{-iz'\cdot\xi'}. \tag{38}$$

Now we go back to $w(x,z)$. We set $\phi_\pm := \phi$ in (32), i.e.,

$$w(x,z) = (U_+(x_m)\phi)(x', z),$$

and then from (36) we have

$$(\partial_{x_3} + i\lambda(a))w|_{x_m=0} = B\phi + i\lambda(a)\phi, \tag{39}$$

because $U_+(0)\phi = \phi$. By Fourier transform, the right-hand side of (39) becomes

$$-|\xi'|\hat{\phi}(\xi') + i\lambda(a)\hat{\phi}(\xi'). \tag{40}$$

By similar computations for the fundamental solution $\Gamma(x,z)$, we have

$$\begin{aligned} -(\partial_{x_m}\Gamma_- + i\lambda(a)\Gamma_-)\,|_{x_m=0} &= -(-B\Gamma_- + i\lambda(a)\Gamma_-)\,|_{x_m=0} \\ &= -(-U_-(-z_m)B\psi_- + i\lambda(a)U_-(-z_m)\psi_-)(x') \end{aligned} \tag{41}$$

because

$$BU_\pm = U_\pm B.$$

The Fourier transform of (41) is

$$-e^{-z_m|\xi'|}(|\xi'|+i\lambda(a))\hat{\psi}_-(\xi',z'), \tag{42}$$

and hence by combining (40) with (42), we obtain

$$\hat{\phi}(\xi',z) = \frac{|\xi'|+i\lambda(a)}{|\xi'|-i\lambda(a)} e^{-z_m|\xi'|}\hat{\psi}_-(\xi',z'). \tag{43}$$

Using (35), we have

$$\hat{\phi}(\xi',z) = \frac{1}{2}\frac{|\xi'|+i\lambda(a)}{|\xi'|(|\xi'|-i\lambda(a))} e^{-z_m|\xi'|}e^{-iz'\cdot\xi'}.$$

Finally (32) becomes

$$w(x,z) = \frac{1}{2(2\pi)^{m-1}}\int_{\mathbb{R}^{m-1}} e^{i(x'-z')\cdot\xi'}e^{-(x_m+z_m)|\xi'|}\frac{|\xi'|+i\lambda(a)}{|\xi'|(|\xi'|-i\lambda(a))}d\xi'. \tag{44}$$ ∎

Next we deal with the second part of the proposition. From Lemma 3.3, we have

$$w(z,z) = \frac{1}{(2\pi)^{m-1}}\int_{\mathbb{R}^{m-1}} e^{-(2z_m)|\xi'|}\frac{1}{2}\frac{|\xi'|+i\lambda(a)}{|\xi'|(|\xi'|-i\lambda(a))}d\xi'. \tag{45}$$

We start with the case $m=3$. Using polar coordinates, we write

$$\begin{aligned} w(z,z) &= \frac{1}{4\pi^2}\int_0^{2\pi} d\theta \int_0^\infty e^{-2z_3 r}\frac{1}{2r}\left(1+\frac{2i\lambda(a)}{r-i\lambda(a)}\right) r dr \\ &= \frac{1}{2\pi}\int_0^\infty e^{-2z_3 r}\frac{1}{2r}\left(1+\frac{2i\lambda(a)}{r-i\lambda(a)}\right) r dr. \end{aligned}$$

After some computations, we obtain

$$\mathrm{Re}w(z,z) = \frac{1}{8\pi z_3} + O(1). \tag{46}$$

Similarly we obtain

$$2\pi \mathrm{Im} w(z,z) = \lambda(a)\left[-\ln(\lambda(a)) + \int_0^\infty e^{-2t}\ln(t^2+z_3^2\lambda(a)^2)dt - \ln(z_3)\right]. \tag{47}$$

Hence

$$2\pi\frac{\mathrm{Im}w(z,z)}{\ln(z_3)} = -\lambda(a) + \frac{\lambda(a)(-\ln(\lambda(a)) + \int_0^\infty e^{-2t}\ln(t^2+z_3^2\lambda^2)dt)}{\ln(z_3)},$$

which gives the formula:

$$\lim_{z_3\to 0^+} -2\pi\frac{\mathrm{Im}w(z,z)}{\ln(z_3)} = \lambda(a). \tag{48}$$

For $m>3$, we use also the hyperspherical coordinates and get

$$w(z,z) = \frac{\sigma_{m-1}}{2(2\pi)^{m-1}}\int_0^\infty e^{-2z_m r}\frac{1}{2r}\left(1+\frac{2i\lambda(a)}{r-i\lambda(a)}\right) r^{m-1}dr.$$

Hence

$$\mathrm{Re}w(z,z) = \frac{\sigma_{m-1}}{2(2\pi)^{m-1}} \frac{1}{(2z_m)^{m-2}} + O\left(\frac{1}{(z_m)^{m-3}}\right)$$

and

$$\mathrm{Im}w(z,z) = \frac{\sigma_{m-1}}{(2\pi)^{m-1}} \frac{\lambda(a)}{(2z_m)^{m-3}} + O(1).$$

This ends the proof of Proposition 3.1. ∎

3.2. Proof of Proposition 3.2. We assume that the point a is on ∂D_I. The case where a is on ∂D_D is similar and easier.

Let $\tilde{\Phi}^s$ be the corresponding solution as Φ^s replacing ∂D_I by ∂D (i.e., taking $\partial D_D = \emptyset$). We set $G_\lambda(x,z) := \tilde{\Phi}^s(x,z) + \Phi(x,z)$, the Green's function of the problem (2),(3),(5). We set also $G_{\lambda(a)}(x,z)$ to be the Green's function of (2),(3),(5) when the function $\lambda(x)$ is replaced by the constant function $\lambda(a)$. For both of the Green's functions we assumed $\partial D_D = \emptyset$. Finally, we set $G^0_{\lambda(a)}$ to be the Green's function satisfying

$$\left\{\begin{array}{l} \Delta G^0_\lambda = -\delta \text{ in } \Omega \setminus \overline{D}, \\ \frac{\partial G^0_\lambda}{\partial \nu}(x,z) + i\lambda(a) G^0_\lambda(x,z) = 0 \text{ on } \partial D, \\ G^0_{\lambda(a)}(\cdot, z) = 0 \text{ on } \partial\Omega, \end{array}\right. \tag{49}$$

with an arbitrary fixed C^2-regular domain Ω containing $\overline{D}$.

We have the following lemma.

LEMMA 3.4. *For every $R > 0$, there exists a positive constant $C := C(R)$ such that*

1. $|G_\lambda(x,z)| \leq \frac{C}{|x-z|^{m-2}}$,
2. $|\nabla G_\lambda(x,z)| \leq \frac{C}{|x-z|^{m-1}}$ *for* $(x,z) \in (\mathbb{R}^m \setminus D) \cap B(0,R)$.

Proof of Lemma 3.4. These properties are known for general equations and boundary conditions. We refer to [19, 20], where these results are justified for boundary value problems stated on bounded domains. Since the arguments are local, these estimates are also justified for exterior problems. ∎

The function $\tilde{\Phi}^s - \Phi^s$ satisfies

$$\left\{\begin{array}{l} (\Delta + \kappa^2)(\tilde{\Phi}^s - \Phi^s) = 0 \text{ in } \mathbb{R}^m \setminus \overline{D}, \\ \tilde{\Phi}^s - \Phi^s(x,z) = \tilde{\Phi}^s + \Phi \text{ on } \partial D_D, \\ \frac{\partial(\tilde{\Phi}^s - \Phi^s)}{\partial \nu}(x,z) + i\lambda(x)(\tilde{\Phi}^s - \Phi^s)(x,z) = 0 \text{ on } \partial D_I, \\ (\tilde{\Phi}^s - \Phi^s)(\cdot, z) \text{ satisfies Sommerfeld radiation conditions.} \end{array}\right. \tag{50}$$

For z near a, Lemma 3.4 implies that $(\tilde{\Phi}^s + \Phi)(\cdot, z)$ is bounded in $H^{1/2}(\partial D_D)$. The well-posedness of (50) (see [5]) implies that $(\tilde{\Phi}^s - \Phi^s)(\cdot, z)$ is bounded in $H^1_{loc}(\mathbb{R}^m \setminus \overline{D})$. Introducing a cutoff function around the point a and knowing that $\tilde{\Phi}^s(\cdot, z)$ and $\Phi(\cdot, z)$ and their derivatives are bounded for x near ∂D_D and z near a (which is in ∂D_I), we deduce that $(\tilde{\Phi}^s - \Phi^s)$ is bounded for x and z near a. This implies that we can replace Φ^s by $\tilde{\Phi}^s$ in Proposition 3.2. In addition, by setting

$$\tilde{\Phi}^s - (\Gamma_{\lambda(a)} - \Gamma) = G_\lambda - \Gamma_{\lambda(a)} - (\Phi - \Gamma),$$

and knowing that $(\Phi-\Gamma)(x,z)$ is bounded in $\mathbb{R}^m$, then the proof of Proposition 3.2 is reduced to considering the term $G_\lambda - \Gamma_{\lambda(a)}$. We split the rest of the proof into the following three lemmas.

LEMMA 3.5. *There exist $\delta(a)>0$ and $C(R)>0$ such that $|G_\lambda(x,z)-G_{\lambda(a)}(x,z)|\leq C(R)|z-a|^{3-m+\beta}+C(R)$ for $z\in B(a,\delta(a))\cap C_{a,\theta}$ and $x\in\mathbb{R}^m\setminus D$.*

LEMMA 3.6. *There exists $C>0$ such that $|G_{\lambda(a)}(x,z)-G^0_{\lambda(a)}(x,z)|\leq C|z-a|^{4-m}+C$ for z near D and $x\in\Omega\setminus D$.*

LEMMA 3.7. *There exist $C>0$ and $\delta(a)>0$ such that*

1. $|\mathrm{Im}G^0_{\lambda(a)}(x,z)-\mathrm{Im}\Gamma_{\lambda(a)}(x,z)|\leq C|z-a|^{4-m}+C$ *for* $(x,z)\in B(a,\delta(a))\cap C_{a,\theta}$.
2. $|\mathrm{Re}G^0_{\lambda(a)}(x,z)-\mathrm{Re}\Gamma_{\lambda(a)}(x,z)|\leq C|\ln|z-a||$ *for* $(x,z)\in B(a,\delta(a))\cap C_{a,\theta}$, *if* $m=3$.
3. $|\mathrm{Re}G^0_{\lambda(a)}(x,z)-\mathrm{Re}\Gamma_{\lambda(a)}(x,z)|\leq C|z-a|^{3-m}$ *for* $(x,z)\in B(a,\delta(a))\cap C_{a,\theta}$, *if* $m>3$.

In the proofs of these last lemmas we do not, in general, specify the interdependency of the constants appearing in the estimates. However, we distinguish the constant depending on the angle θ and the ones which do not depend.

Proof of Lemma 3.5. We set $R(x,z):=G_\lambda(x,z)-G_{\lambda(a)}(x,z)$. Then it satisfies

$$
(51)\qquad \begin{cases} (\Delta+\kappa^2)R(x,z)=0 \text{ in } \mathbb{R}^m\setminus\overline{D},\\ \frac{\partial R(x,z)}{\partial\nu}+i\lambda(a)R(x,z)=-i(\lambda-\lambda(a))G_\lambda(x,z) \quad \text{on} \quad \partial D,\\ R(\cdot,z) \text{ satisfies the Sommerfeld radiation condition.}\end{cases}
$$

From (51), we have the representation:

$$
R(x,z)=-\int_{\partial D} i(\lambda(y)-\lambda(a))G_{\lambda(a)}(y,x)G_\lambda(y,z)ds(y) \text{ for } (x,z)\in\mathbb{R}^m\setminus\overline{D}.
$$

Hence letting x tend to ∂D we have

$$
(52)\qquad R(x,z)=-\int_{\partial D} i(\lambda(y)-\lambda(a))G_{\lambda(a)}(y,x)G_\lambda(y,z)ds(y) \text{ for } x\in\partial D \text{ and } z\in\mathbb{R}^m\setminus\overline{D}.
$$

From the assumption on the regularity of the surface impedance $\lambda(x)$, we have

$$
|\lambda(y)-\lambda(a)|\leq C|y-a|^\beta.
$$

It is clear that $|y-a|\leq c(\theta)|y-z|$ for $y\in\partial D_I$ and $z\in C_{a,\theta}\cap B(a,\delta(a))$ with a positive constant $c(\theta)$ depending on the angle θ. This is due to the fact that ∂D_I and $C_{a,\theta}\cap B(a,\delta(a))$ are separated, i.e., $\partial D_I\cap C_{a,\theta}\cap B(a,\delta(a))=\{a\}$. From the inequality

$$
\frac{|\lambda(y)-\lambda(a)|}{|y-z|^{m-2}}\leq\frac{c(\theta)^\beta C}{|y-z|^{m-2-\beta}}
$$

and point 1 of Lemma 3.4, we have

$$
|R(x,z)|\leq\int_{\partial D}\frac{c(\theta)^\beta C}{|y-z|^{m-2-\beta}|y-x|^{m-2}}dy\leq\frac{C}{|x-z|^{m-3-\beta}}+C
$$

and then

$$
\max_{x\in\partial D}|R(x,z)|\leq\frac{C}{|z-a|^{m-3-\beta}}+C \text{ for } z\in C_{a,\theta}\cap B(a,\delta(a)).
$$

Now the solvability of the forward problem

$$\left\{\begin{array}{l}(\Delta+\kappa^2)R(x,z)=0 \text{ in } \mathbb{R}^m\setminus\overline{D},\\ |R(\cdot,z)|\leq \frac{C}{|z-a|^{m-3-\beta}}+C \quad \text{on} \quad \partial D,\\ R(\cdot,z) \text{ satisfies the radiation conditions}\end{array}\right. \tag{53}$$

implies the desired estimate for $R(x,z)$ for $x\in\mathbb{R}^m\setminus D$ and $z\in C_{a,\theta}\cap B(0,R)$. ∎

Proof of Lemma 3.6. We recall that $G^0_{\lambda(a)}$ satisfies

$$\left\{\begin{array}{l}\Delta G^0_\lambda(x,z)=-\delta(x,z) \text{ in } \Omega\setminus\overline{D},\\ \frac{\partial G^0_\lambda}{\partial\nu}(x,z)+i\lambda G^0_\lambda(x,z)=0 \text{ on } \partial D,\\ G^0_\lambda(x,z)=0 \text{ on } \partial\Omega.\end{array}\right. \tag{54}$$

Then $G_{\lambda(a)}-G^0_{\lambda(a)}$ is the solution of the problem

$$\left\{\begin{array}{l}\Delta(G_{\lambda(a)}-G^0_{\lambda(a)})(x,z)=\kappa^2G_{\lambda(a)}(x,z) \text{ in } \Omega\setminus\overline{D},\\ \frac{\partial(G_{\lambda(a)}-G^0_{\lambda(a)})}{\partial\nu}(x,z)+i\lambda(a)(G_{\lambda(a)}-G^0_{\lambda(a)})(x,z)=0 \text{ on } \partial D,\\ (G_{\lambda(a)}-G^0_{\lambda(a)})(x,z)=G_{\lambda(a)}(x,z) \text{ on } \partial\Omega.\end{array}\right. \tag{55}$$

Using integral representation for the solution of (55) and Lemma 3.4 applied for $G_{\lambda(a)}$ and $G^0_{\lambda(a)}$, we have the desired estimate for $(G_{\lambda(a)}-G^0_{\lambda(a)})(x,z)$ for x in $\Omega\setminus D$ and z near ∂D. ∎

Proof of Lemma 3.7. We can assume without loss of generality that $a=(0,0,0,\ldots,0)$ by using a rigid transformation of coordinates. Let $\xi=F(x)$ be the following local change of variables:

$$\left\{\begin{array}{l}\xi'=x',\\ \xi_m=x_m-f(x'),\end{array}\right. \tag{56}$$

where f is the function defined in the introduction. We have the following properties:

$$\left\{\begin{array}{l}c_1|x-z|\leq|F(x)-F(z)|\leq c_2|x-z|,\\ |F(x)-x|\leq c_3|x|^2,\\ |DF(x)-I|\leq c_4|x|,\end{array}\right. \tag{57}$$

for x,z near the point a, where $c_i, i=1,\ldots,4$, are positive constants, which is due to the C^2 regularity of ∂D.

Let x,z be near the point a. We set $\tilde{G}^0_{\lambda(a)}(\xi,\eta):=G^0_{\lambda(a)}(x,y)$, where $\xi=F(x)$ and $\eta:=F(z)$. Then $\tilde{G}^0_{\lambda(a)}(\cdot,\eta)$ satisfies

$$\left\{\begin{array}{l}\nabla\cdot B(\xi)\nabla\tilde{G}^0_{\lambda(a)}=-\delta(\xi-\eta) \text{ near } F(a),\\ |J^{-T}\nu|B\nabla\tilde{G}^0_{\lambda(a)}\cdot\nu+i\lambda(a)\tilde{G}^0_{\lambda(a)}=0 \text{ on } \partial\mathbb{R}^m_+ \text{ near } F(a),\end{array}\right. \tag{58}$$

where $B=JJ^T$ and $J=\frac{\partial\xi}{\partial x}(F^{-1}(\xi))$. From the properties (57), we have

$$|J(\xi)-J(0)|\leq c|\xi|, |B(\xi)-B(0)|\leq c|\xi|,$$

and $J(0)=B(0)=I$.

First step. Using similar notations as in the proof of Lemma 3.5, we write $\tilde{R}(\xi,\eta) = \tilde{G}^0_{\lambda(a)}(\xi,\eta) - \Gamma_{\lambda(a)}(\xi,\eta)$, and hence

(59)
$$\begin{cases} \Delta\tilde{R}(\xi,\eta) = \nabla\cdot(I-B)\nabla\tilde{G}^0_{\lambda(a)} \text{ near } F(a) \\ \nabla\tilde{R}\cdot\tilde{\nu} + i\lambda(a)\tilde{R} = (I-B)\nabla\tilde{G}^0_{\lambda(a)}\cdot\tilde{\nu} + i\lambda(a)(1-|J^{-T}|)\tilde{R} \text{ on } \partial\mathbb{R}^m_+ \text{ near } F(a). \end{cases}$$

Let $B_r^+ := B(F(a),r)\cap[F(D)]^c$ for r small enough; then by (59) and using the local Green's function $\Gamma_{\lambda(a)}$, the solution $\tilde{R}$ has the following representation:

$$-\tilde{R}(\xi,\eta) + \int_{\partial B_r^+}\frac{\partial\tilde{R}(z,\xi)}{\partial\nu}\Gamma_{\lambda(a)}(z,\eta)ds(z) - \int_{\partial B_r^+}\frac{\partial\Gamma_{\lambda(a)}(z,\eta)}{\partial\nu}\tilde{R}(z,\xi)ds(z)$$

$$= -\int_{B_r^+}(I-B)\nabla\tilde{G}^0_{\lambda(a)}(z,\xi)\cdot\nabla\Gamma_{\lambda(a)}(z,\eta)dz + \int_{\partial B_r^+}(I-B)\nabla\tilde{G}^0_{\lambda(a)}(z,\xi)\cdot\nu\Gamma_{\lambda(a)}(z,\eta)ds(z)$$

for ξ and η in B_r^+. We write $\partial B_r^+ = S_r\cup S_r^c$, where $S_r = \partial B_r^+\cap\partial(F(D))$. Using the impedance boundary condition on S_r, the last equation becomes

$$-\tilde{R}(\xi,\eta) - \int_{S_r} i\lambda(a)\tilde{R}(z,\xi)\cdot\Gamma_{\lambda(a)}(z,\eta)ds(z) + \int_{S_r} i\lambda(a)\Gamma_{\lambda(a)}(z,\eta)\tilde{R}(z,\xi)ds(z)$$

$$= -\int_{B_r^+}(I-B)\nabla\tilde{G}^0_{\lambda(a)}(z,\eta)\cdot\nabla\Gamma_{\lambda(a)}(z,y)dz + \int_{\partial B_r^+}(I-B)\nabla\tilde{G}^0_{\lambda(a)}(z,\xi)\cdot\nu\Gamma_{\lambda(a)}(z,\eta)ds(z)$$

$$+\int_{S_r^c}\frac{\partial}{\partial\nu}\tilde{R}(z,\xi)\Gamma_{\lambda(a)}(z,\eta)ds(z) + \int_{S_r^c}\frac{\partial}{\partial\nu}\Gamma_{\lambda(a)}(z,\eta)\tilde{R}(z,\xi)ds(z)$$

$$-\int_{S_r}(I-B)\nabla\tilde{G}^0_{\lambda(a)}(z,\xi)\cdot\nu\Gamma_{\lambda(a)}(z,\eta)ds(z) - i\lambda(a)\int_{S_r}(1-|J^{-T}\nu|)\tilde{R}(z,\eta)\Gamma_{\lambda(a)}(z,\xi)ds(z).$$

After simplification we have

$$\begin{aligned} -\tilde{R}(\xi,\eta) = &-\int_{B_r^+}(I-B)\nabla\tilde{G}^0_{\lambda(a)}(z,\xi)\cdot\nabla\Gamma_{\lambda(a)}(z,\eta)dz \\ &+\int_{S_r^c}(I-B)\frac{\partial\tilde{G}^0_{\lambda(a)}(z,\xi)}{\partial\nu}\Gamma_{\lambda(a)}(z,\eta)ds(z) \\ &+\int_{S_r^c}\frac{\partial\tilde{R}}{\partial\nu}(z,\xi)\Gamma_{\lambda(a)}(z,\eta)ds(z) + \int_{S_r^c}\frac{\partial}{\partial\nu}\Gamma_{\lambda(a)}(z,\eta)\tilde{R}(z,\xi)ds(z) \\ &-i\lambda(a)\int_{S_r}(1-|J^{-T}\nu|)\tilde{R}(z,\eta)\Gamma_{\lambda(a)}(z,\xi)ds(z). \end{aligned} \tag{60}$$

Taking the imaginary part in the last equality, we have

$$
\begin{aligned}
-\mathrm{Im}\tilde{R}(\xi,\eta) = &-\int_{B_r^+}(I-B)\nabla(\mathrm{Im}\tilde{G}^0_{\lambda(a)})(z,\xi)\cdot\nabla(\mathrm{Re}\Gamma_{\lambda(a)})(z,\eta)dz\\
&-\int_{B_r^+}(I-B)\nabla(\mathrm{Re}\tilde{G}^0_{\lambda(a)})(z,\xi)\cdot\nabla(\mathrm{Im}\Gamma_{\lambda(a)})(z,\eta)dz\\
&+\int_{S_r^c}(I-B)\frac{\partial\mathrm{Im}\tilde{G}^0_{\lambda(a)}}{\partial\nu}(z,\xi)\mathrm{Re}\Gamma_{\lambda(a)}(z,\eta)ds(z)\\
&+\int_{S_r^c}(I-B)\frac{\partial\mathrm{Re}\tilde{G}^0_{\lambda(a)}}{\partial\nu}(z,\xi)\mathrm{Im}\Gamma_{\lambda(a)}(z,\eta)ds(z)\\
&+\int_{S_r^c}\frac{\partial\mathrm{Im}\tilde{R}}{\partial\nu}(z,\xi)\mathrm{Re}\Gamma_{\lambda(a)}(z,\eta)ds(z)\\
&+\int_{S_r^c}\frac{\partial\mathrm{Re}\tilde{R}}{\partial\nu}(z,\xi)\mathrm{Im}\Gamma_{\lambda(a)}(z,\eta)ds(z)\\
&+\int_{S_r^c}\frac{\partial}{\partial\nu}\mathrm{Im}\Gamma_{\lambda(a)}(z,\eta)\mathrm{Re}\tilde{R}(z,\xi)ds(z)\\
&+\int_{S_r^c}\frac{\partial}{\partial\nu}\mathrm{Re}\Gamma_{\lambda(a)}(z,\eta)\mathrm{Im}\tilde{R}(z,\xi)ds(z)\\
&-\lambda(a)\int_{S_r}(1-|J^{-T}\nu|)\mathrm{Re}[\tilde{R}(z,\eta)\Gamma_{\lambda(a)}(z,\xi)]ds(z).
\end{aligned}\tag{61}
$$

We have for $\Gamma_{\lambda(a)}$ similar estimates as in Lemma 3.4. In particular, we have

$$|\nabla(\mathrm{Re}\Gamma_{\lambda(a)})(x,z)|\leq c|x-z|^{1-m}.$$

It is of importance to remark that the imaginary parts have fewer singularities. Indeed, we will prove the following lemma.

LEMMA 3.8. *For every $R>0$, there exists $c:=c(R)$ such that*

$$|\nabla(\mathrm{Im}\Gamma_{\lambda(a)})(x,z)|\leq c|x-z|^{2-m+s}d^{-s}(x,\partial B_+)$$

and

$$|\nabla(\mathrm{Im}\tilde{G}^0_{\lambda(a)})(x,z)|\leq c|x-z|^{2-m+s}d^{-s}(x,\partial B_+)$$

for x,z in $B_+(0,R):=B(0,R)\cap\mathbb{R}^m_+$.

Proof of Lemma 3.8. From (26), we deduce that $\mathrm{Im}\Gamma_{\lambda(a)}(\cdot,z)$ satisfies

$$
\begin{cases}\Delta(\mathrm{Im}\Gamma_{\lambda(a)})(\cdot,z)=0 \text{ in } \mathbb{R}^m_+,\\ \frac{\partial}{\partial\nu}\mathrm{Im}\Gamma_{\lambda(a)}(\cdot,z)=-\lambda(a)\mathrm{Re}\Gamma_{\lambda(a)}(\cdot,z) \text{ on } \partial\mathbb{R}^m_+.\end{cases}\tag{62}
$$

Let Ω be a regular domain in $\mathbb{R}^m$ symmetric with respect to the plane $\{x_m=0\}$. We state the problem (62) on $\Omega_+:=\Omega\cap\mathbb{R}^m_+$. Let G_+ be the Neumann Green's function of the Laplace on Ω_+. From (62) we can write

$$\mathrm{Im}\Gamma_{\lambda(a)}(x,z)=\int_{\partial\Omega_+}G_+(x,y)\frac{\partial\mathrm{Im}\Gamma_{\lambda(a)}}{\partial\nu}(z,y)ds(y).\tag{63}$$

The boundary condition in (62) on $\partial\Omega_+\cap\partial\mathbb{R}^m_+$ gives

$$
\begin{aligned}
\mathrm{Im}\Gamma_{\lambda(a)}(x,z)=&-\int_{\partial\Omega_+\cap\partial\mathbb{R}^m_+}\lambda(a)G_+(x,y)\mathrm{Re}\Gamma_{\lambda(a)}(z,y)ds(y)\\
&-\int_{\partial\Omega_+\setminus\partial\mathbb{R}^m_+}G_+(x,y)\frac{\partial\mathrm{Im}\Gamma_{\lambda(a)}}{\partial\nu}(z,y)ds(y).
\end{aligned}
$$

Taking the derivatives, we have

$$\nabla_x(\mathrm{Im}\Gamma_{\lambda(a)})(x,z) = -\int_{\partial\Omega_+\cap\partial\mathbb{R}^m_+} \lambda(a)\nabla_x G_+(x,y)\mathrm{Re}\Gamma_{\lambda(a)}(z,y)ds(y) - \int_{\partial\Omega_+\setminus\partial\mathbb{R}^m_+} \nabla_x G_+(x,y)\frac{\partial \mathrm{Im}\Gamma_{\lambda(a)}}{\partial\nu}(z,y)ds(y). \tag{64}$$

Taking Ω large enough to contain $B_+(0,R)$ and using the estimates of G_+ and $\Gamma_{\lambda(a)}$ given in Lemma 3.4, we deduce that the second term in the right-hand side is bounded for x, z in $B_+(0,R)$ because $\partial\Omega_+ \setminus \partial\mathbb{R}^m_+$ is away from $B_+(0,R)$. The first term can be estimated by

$$\begin{aligned}
&\left|\int_{\partial\Omega_+\cap\partial\mathbb{R}^m_+} \lambda(a)\nabla_x G_+(x,y)\mathrm{Re}\Gamma_{\lambda(a)}(z,y)ds(y)\right| \\
&\quad\le c\int_{\partial\Omega_+\cap\partial\mathbb{R}^m_+} |x-y|^{1-m}|y-z|^{2-m}ds(y) \\
&\quad\le c\int_{\partial\Omega_+\cap\partial\mathbb{R}^m_+} |x-y|^{1-m+s}|y-z|^{2-m}ds(y)d^{-s}(x,\partial\Omega_+) \\
&\quad\le c_s|x-z|^{2-m+s}d^{-s}(x,\partial\Omega_+),
\end{aligned}$$

where c_s is a positive constant behaving as $\frac{1}{1-s}$; see [14] (or [21] for Lipschitz surfaces). Similar arguments can be applied for $\tilde{G}^0_{\lambda(a)}$. □

Using this lemma and the estimates of $\tilde{G}^0_{\lambda(a)}$ in (61), we have

$$\begin{aligned}
|\mathrm{Im}\tilde{R}(\xi,\eta)| \le c\Bigg[&\int_{B_r^+} |z|d^{-s}(z,\partial B_r^+)|z-\xi|^{2-m+s}|z-\eta|^{1-m}dz \\
&+\int_{B_r^+} |z||z-\xi|^{1-m}d^{-s}(z,\partial B_r^+)|z-\eta|^{2-m+s}dz \\
&+\int_{S_r^c} d^{-s}(z,\partial B_r^+)|z-\xi|^{2-m+s}|z-\eta|^{1-m}ds(z) \\
&+\int_{S_r^c} |z-\xi|^{1-m}d^{-s}(z,\partial B_r^+)|z-\eta|^{2-m+s}ds(z) \\
&+\int_{S_r} |z||z-\xi|^{2-m}|z-\eta|^{2-m}ds(z)\Bigg],
\end{aligned} \tag{65}$$

where c is a positive constant independent on ξ and η.

Taking $\eta \in C_{F(a),\theta}$ and ξ on S_r away from S_r^c, from (65) we will show the desired estimate of $\mathrm{Im}\tilde{R}(\xi,\eta)$.

The third and the fourth integrals in (65) are bounded because ξ is away from S_r^c. Let us consider the first integral. For $\eta \in C_{F(a),\theta}$ and ξ on S_r, we have the inequality

$$|\xi| \le c(\theta)|\xi-\eta|, \tag{66}$$

with some positive constant $c(\theta)$. We decompose B_r^+ as $B_r^+ = B_{r_1}^+ \cup B_{r_2}^+$, where $B_{r_1}^+ := B_r^+ \cap C_{F(a),\tilde{\theta}}$, where $\theta < \tilde{\theta} < \frac{\pi}{2}$ and $B_{r_2}^+ := B_r^+ \setminus B_{r_1}^+$.

Let us consider $\int_{B_{r_1}^+} |z|d^{-s}(z,\partial B_r^+)|z-\xi|^{2-m+s}|z-\eta|^{1-m}dz$. For $z \in B_r^+$, we have $|z| \le c(\tilde{\theta})d(z,\partial B_r^+)$ for some positive constant $c(\tilde{\theta})$. Hence we need to consider the integral:

$$\int_{B_{r_1}^+} |z|^{1-s}|z-\xi|^{2-m+s}|z-\eta|^{1-m}dz.$$

At this point, we make use of an argument from [1, pp. 209–210]. We decompose the last integral as the sum of

$$I_1(\xi,\eta) := \int_{\{|z|<4c(\theta)h\}\cap B^+_{r_1}} |z|^{1-s}|z-\xi|^{2-m+s}|z-\eta|^{1-m}dz$$

and

$$I_2(\xi,\eta) := \int_{\{|z|>4c(\theta)h\}\cap B^+_{r_1}} |z|^{1-s}|z-\xi|^{2-m+s}|z-\eta|^{1-m}dz,$$

where $h := |\xi-\eta|$. After the change of variables $t := \frac{z}{h}$, we obtain

$$I_1(\xi,\eta) \leq h^{4-m}4c(\theta)\int_{|t|<4c(\theta)} |t|^{1-s}\left|\frac{\xi}{h}-t\right|^{2-m+s}\left|\frac{\eta}{h}-t\right|^{1-m}dt.$$

Since $|\frac{\xi}{h}-\frac{\eta}{h}| = 1$, then from [14, Chapter 2, section 11], the integral in the left-hand side is bounded. Hence

$$I_1(\xi,\eta) \leq c|\xi-\eta|^{4-m}. \tag{67}$$

Let us now consider the term $I_2(\xi,\eta)$. From the inequality (66), we have

$$|z| \ \leq |z-\xi| + |\xi| \leq |z-\xi| + c(\theta)|h| \leq |z-\xi| + \frac{1}{4}|z|,$$

and hence

$$|z| \leq \frac{4}{3}|z-\xi|. \tag{68}$$

Similarly we get $|z| \leq \frac{4c(\theta)}{3c(\theta)-1}|z-\eta|$. In (66), we can always take $c(\theta) > \frac{1}{3}$. Hence

$$I_2(\xi,\eta) \leq c\int_{r>|z|>4c(\theta)h} |z|^{4-2m}dz \leq C|\xi-\eta|^{4-m} + C. \tag{69}$$

Summing up (67) and (69) implies

$$\int_{B^+_{r_1}} |z|d^{-s}(z,\partial B^+_r)|z-\xi|^{2-m+s}|z-\eta|^{1-m}dz \leq C|\xi-\eta|^{4-m} + C. \tag{70}$$

Now we deal with $\int_{B^+_{r_2}}|z|d^{-s}(z,\partial B^+_r)|z-\xi|^{2-m+s}|z-\eta|^{1-m}dz$. In this case $d(z,\partial B^+_{r_2}) = |z_m|$. Since $\eta \in C_{F(a),\theta}$, then $|z-\eta|^{1-m} > c > 0$, since $C_{F(a),\theta}$ and $B^+_{r_2}$ are separated sets.

Hence, using this information with the fact that $|z| \leq \frac{4}{3}|z-\xi|$, we have

$$\int_{B^+_{r_2}} |z|d^{-s}(z,\partial B^+_r)|z-\xi|^{2-m+s}|z-\xi|^{1-m}dz \leq \int_{B^+_{r_2}} |z|^{3-m+s}|z_m|^{-s}dz.$$

By using the Holder inequality and choosing s small enough we have $\int_{B^+_{r_2}} |z|^{3-m+s}|z_m - \xi|^s dz < \infty$. This means that the first integral of (65) is estimated by $C|\xi-\eta|^{4-m} + C$.

Arguing as before for the first terms of (65), we deduce that the second term has a similar estimate. From the inequality (68), the last integral is bounded by $c\int_{S_r} |z-\xi|^{3-m}|z-\eta|^{2-m}ds(z)$, which is itself bounded by $C|\xi-\eta|^{4-m} + C$. Finally we deduce that the integral in (65) is also bounded by $C|\xi-\eta|^{4-m} + C$.

Second step. We go back to estimate $R(x,z) := G^0_{\lambda(a)}(x,z) - \Gamma_{\lambda(a)}(x,z)$. We have

$$R(x,z) = G^0_{\lambda(a)}(x,z) - \Gamma_{\lambda(a)}(F(x),F(z)) + \Gamma_{\lambda(a)}(F(x),F(z)) - \Gamma_{\lambda(a)}(x,z),$$

which we write as

$$\begin{aligned} R(x,z) = \tilde{R}(F(x),F(z)) &+ (\Gamma_{\lambda(a)}(F(x),F(z)) - \Gamma_{\lambda(a)}(F(x),z)) \\ &+(\Gamma_{\lambda(a)}(F(x),z) - \Gamma_{\lambda(a)}(x,z)). \end{aligned} \tag{71}$$

Let us show that $F(C_{a,\theta} \cap B(0,\delta(a))) \subset C_{F(a),\theta'}$ for some $\theta' \in (0,\pi/2)$ and $\theta \leq \theta'$ with $\delta(a)$ small enough. Recall that $a = (0,0,0)$. We set $M := \max_{(z_1,z_2)\in\overline{B_2}(0,\delta(a))} |f(z_1, z_2)|$ and then $M = |f(z_1^0, z_2^0)|$ for some point (z_1^0, z_2^0). We draw the cone with vertex a, axis $(-\nu(a))$, and having the point (z_1^0, z_2^0, M) on its boundary. We denote by ϕ the angle between this cone and the plane $\{z_3 = 0\}$. Since $|\nabla f(0,0)| = 0$ and $f \in C^1(0,\delta(a))$, $\phi = \phi(\delta(a))$ tends to zero if $\delta(a)$ tends to zero. Hence taking $\delta(a)$ small enough, if necessary, we can assume that $0 \leq \phi < \pi/2 - \theta$. Finally we define the cone with center a, axis $(-\nu(a))$, and angle $\theta' := \phi + \theta$. Then $F(C_{a,\theta}) \subset C_{F(a),\theta'}$.

Let x be near a such that $F(x) \in S_r$ and away from S_r^c and $z \in C_{a,\theta} \cap B(a,\delta(a))$. Then $(F(x),F(z)) \in S_r \times C_{F(a),\theta'}$ and hence

$$\mathrm{Im}\tilde{R}(F(x),F(z)) \leq C|F(x) - F(z)|^{4-m} + C \leq C|x-z|^{4-m} + C$$

as shown in the first step.

In the following we need to estimate the second and the third terms of the right-hand side of (71). We write

$$|\Gamma_{\lambda(a)}(F(x),F(z)) - \Gamma_{\lambda(a)}(F(x),z)| \leq |\nabla_z \mathrm{Im}\Gamma_{\lambda(a)}(F(x),\cdot)|_{L^\infty(V_{z,\epsilon})}|F(z) - z| \tag{72}$$

and

$$|\Gamma_{\lambda(a)}(F(x),z) - \Gamma_{\lambda(a)}(x,z)| \leq |\nabla_x \mathrm{Im}\Gamma_{\lambda(a)}(\cdot,z)|_{L^\infty(V_{x,\epsilon})}|F(x) - x|, \tag{73}$$

where $V_{z,\epsilon}$ is an open set containing z and $F(z)$ such that $F(x) \in V^c_{z,\epsilon}$, and $V_{x,\epsilon}$ is an open set containing x and $F(x)$ such that $z \in V^c_{x,\epsilon}$.

From the representation (64) in the proof of Lemma 3.8, we have

$$|\nabla_z(\mathrm{Im}\Gamma_{\lambda(a)}(F(x),\cdot))|_{L^\infty(V_{z,\epsilon})} < c[d(F(x),\partial V_{z,\epsilon})]^{2-m},$$

and similarly, we have

$$|\nabla_x(\mathrm{Im}\Gamma_{\lambda(a)}(\cdot,z))|_{L^\infty(V_{x,\epsilon})} < c[d(z,\partial V_{x,\epsilon})]^{2-m},$$

where c is independent of $V_{z,\epsilon}$ and $V_{x,\epsilon}$.

Hence (72) and (73) become

$$|\Gamma_{\lambda(a)}(F(x),F(z)) - \Gamma_{\lambda(a)}(F(x),z)| \leq c[d(F(x),\partial V_{z,\epsilon})]^{2-m}|F(z) - z|$$

and

$$|\Gamma_{\lambda(a)}(F(x),z) - \Gamma_{\lambda(a)}(x,z)| \leq c[d(z,\partial V_{x,\epsilon})]^{2-m}|F(x) - x|,$$

respectively. Since $\epsilon > 0$ is arbitrary, we deduce, by shrinking $V_{z,\epsilon}$ to tend to a line joining the points z and $F(z)$ and similarly $V_{x,\epsilon}$ to tend to a line joining the points x and $F(x)$, that there exists $c > 0$ such that

$$(74)\quad \begin{cases} |\Gamma_{\lambda(a)}(F(x), F(z)) - \Gamma_{\lambda(a)}(F(x), z)| \leq c|F(x) - F(z)|^{2-m}|F(z) - z| \\ \text{or} \\ |\Gamma_{\lambda(a)}(F(x), F(z)) - \Gamma_{\lambda(a)}(F(x), z)| \leq c|F(x) - z|^{2-m}|F(z) - z| \end{cases}$$

and

$$(75)\quad \begin{cases} |\Gamma_{\lambda(a)}(F(x), z) - \Gamma_{\lambda(a)}(x, z)| \leq c|x - z|^{2-m}|F(x) - x| \\ \text{or} \\ |\Gamma_{\lambda(a)}(F(x), z) - \Gamma_{\lambda(a)}(x, z)| \leq c|F(x) - z|^{2-m}|F(x) - x|. \end{cases}$$

Recalling that

$$|x| \leq c(\theta)|x - z|$$

for x near a and $z \in C_{a,\theta} \cap B(a, \delta(a))$ and using (57), we have

$$|x - z| \leq |x - F(x)| + |F(x) - z| \leq c|x|^2 + |F(x) - z| \leq c(\theta)|x - z|^2 + |F(x) - z|.$$

Hence

$$|x - z|[1 - c(\theta)|x - z|] \leq |F(x) - z|.$$

Taking $\delta(a)$ small enough so that we have $(1 - c(\theta)|x - z|^2) \geq 1/2$,

$$|x - z| \leq 2|F(x) - z|.$$

From (57), we have

$$(76)\quad |F(x) - x| \leq c|x|^2 \leq c(\theta)|x - z|^2$$

and also

$$(77)\quad |F(z) - z| \leq c|z|^2 \leq c(\theta)|x - z|^2.$$

Since $|x - z| \leq c|F(x) - F(z)|$, using (76) and (77) in (74) and (75), we deduce that the second and the third terms of (71) are estimated by $C|x - z|^{4-m}$.

This means that

$$(78)\quad |\mathrm{Im}R(x, z)| \leq C|x - z|^{4-m} + C$$

for $x \in B(a, \delta(a))$ such that $F(x) \in S_r$ and $z \in C_{a,\theta} \cap B(a, \delta(a))$.

For $z \in C_{a,\theta} \cap B(a, \frac{\delta(a)}{2})$ and $x \in [\partial B(a, \delta(a))] \cap \mathbb{R}^m \setminus \overline{D}$, we have

$$(79)\quad |\mathrm{Im}R(x, z)| \leq C$$

with some positive constant c, because $C_{a,\theta} \cap B(a, \frac{\delta(a)}{2})$ and $\partial B(a, \delta(a)] \cap \mathbb{R}^3 \setminus \overline{D}$ are separated sets. Since in $B(a, \delta(a)) \cap (\mathbb{R}^3 \setminus \overline{D})$ we have $\Delta_x \mathrm{Im}R(x, z) = 0$, using (78) and (79), we have, by the maximum principle,

$$|\mathrm{Im}R(x, z)| \leq c(\theta)|z - a|^{4-m} + C$$

for $x \in [\mathbb{R}^3 \setminus D] \cap B(a, \delta(a))$ and $z \in C_{a,\theta} \cap B(a, \frac{\delta(a)}{2})$.

Considering the real part of $R(x, z)$, by similar arguments as for $\mathrm{Im}R(x, z)$, we prove that

$$|\mathrm{Re}R(x, z)| \leq \begin{cases} c(\theta)|\ln |z-a|| \text{ if } m = 3, \\ c(\theta)\frac{1}{|z-a|^{m-3}} \text{ if } m > 3. \end{cases}$$

This ends the proof of Lemma 3.7. $\square$

REFERENCES

[1] G. Alessandrini and M. Di Cristo, *Stable determination of an inclusion by boundary measurements*, SIAM J. Math. Anal., 37 (2005), pp. 200–217.

[2] I. Akduman and R. Kress, *Direct and inverse scattering problems for inhomogeneous impedance cylinders of arbitrary shape*, Radio Sci., 38 (2003), pp. 1055–1064.

[3] F. Cakoni and D. Colton, *Open problems in the qualitative approach to inverse electromagnetic scattering theory*, European J. Appl. Math., 16 (2005), pp. 1–15.

[4] F. Cakoni and D. Colton, *Qualitative Methods in Inverse scattering Theory*, Interaction of Mechanics and Mathematics, Springer, New York, 2006.

[5] D. Colton and F. Cakoni , *The determination of the surface impedance of a partially coated obstacle from far field data*, SIAM. J. Appl. Math., 64 (2004), pp. 709–723.

[6] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer, Berlin, 1998.

[7] J. Cheng, J. Liu, and G. Nakamura, *Recovery of Boundaries and Types of Multiple Obstacles from Far Field Pattern*, preprint, Hokkaido University, Japan, 2005.

[8] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems. Mathematics and its Applications*, Kluwer Academic, Dorderecht, 1996.

[9] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer Ser. Appl. Math. Sci., Springer, Berlin, 1998.

[10] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, Springer Ser. Appl. Math. Sci. 120, Springer, New York, 1996.

[11] R. Kress and W. Rundell, *Inverse scattering for shape and impedance*, Inverse Problems, 17 (2001), pp. 1075–1085.

[12] C. Labreuche, *Stability of the recovery of surface impedances in inverse scattering*, J. Math. Anal. Appl., 231 (1999), pp. 161–176.

[13] J. Liu, G. Nakamura, and M. Sini, *Reconstruction of the shape and surface impedance from acoustic scattering data for arbitrary cylinder*, SIAM J. Appl. Math., 67 (2007), pp. 1124–1146.

[14] C. Miranda, *Partial Differential Equations of Elliptic Type*, Springer, New York, 1970.

[15] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, London, 2000.

[16] R. Potthast, *Point Sources and Multipoles in Inverse Scattering Theory*, Res. Notes Math. 427, Chapman-Hall/CRC, Boca Raton, FL, 2001.

[17] R. Potthast, *Sampling and Probe methods—An algorithmical view*, Computing, 75 (2005), pp. 215–235.

[18] E. Sincich, *Stable determination of the surface impedance of an obstacle by far field measurements*, Siam J. Math. Anal., 38 (2006), pp. 434–451.

[19] V. A. Solonnikov, *On Green's matrices for elliptic boundary value problems*, I, Proc. Steklov Inst. Math, 110 (1970), pp. 123–170.

[20] V. A. Solonnikov, *The Green's matrices for elliptic boundary value problems.* II. *Boundary value problems of mathematical physics*, Trudy. Math. Inst. Steklov, 116 (1971), pp. 181–216.

[21] N. Valdivia, *Uniqueness in inverse obstacle scattering with conductive boundary condictions*, Appl. Anal., 83 (2004), pp. 825–851.

SIAM J. MATH. ANAL.
Vol. 39, No. 3, pp. 838–862

ON A CLASS OF NONCONVEX BOLZA PROBLEMS RELATED TO BLATZ–KO ELASTIC MATERIALS*

P. CELADA† AND S. PERROTTA‡

Abstract. We study the existence of solutions to Bolza problems for a special class of one-dimensional, nonconvex integrals. These integrals describe the possibly singular, radial deformations of certain rubberlike materials called Blatz–Ko materials.

Key words. nonconvex variational problems, Bolza problems, existence of minimizers, cavitation, Blatz–Ko materials

AMS subject classifications. 49J05, 73C50

DOI. 10.1137/060658941

1. Introduction.

In this paper we investigate the existence of solutions to Bolza problems for a special class of one-dimensional, nonconvex integrals of the form

$$\int_0^1 r^{N-1}\Phi\left(v'(r)\,,\frac{v(r)}{r}\right)dr+\theta(v(0)), \tag{1.1}$$

where $\Phi\colon \mathbb{R}^2_+ \to [0\,,+\infty)$ and $\theta\colon [0\,,+\infty) \to [0\,,+\infty)$ are smooth functions and the competing functions $v > 0$ satisfy the unilateral constraint $v' > 0$ a.e. and the one-point boundary condition $v(1) = \lambda$ for a given $\lambda > 1$. We emphasize that Φ need not be convex with respect to the derivative v'.

This special class of variational problems arises in mathematical models describing the deformations of some foam rubberlike materials, including the so-called *Blatz–Ko materials*.

The variational elasticity problem that motivates the model can be described as follows. Let the open unit ball B_1 be the reference configuration of a hyperelastic, isotropic material with stored energy density W so that the total energy corresponding to a smooth deformation u with given displacement $u(x) = \lambda x > 1$ at the boundary $|x| = 1$ is given by

$$E_0(u)=\int_{B_1} W(Du(x))\,dx.$$

Physical arguments require that admissible deformations u be orientation preserving, i.e., $\det Du > 0$, and that $W(Du) \to +\infty$ as $\det Du \to +\infty$ and $\det Du \to 0_+$. The first hypothesis means that no interpenetration of matter occurs, and the others mean that infinite energy is required for unbounded expansion of the body or compression to a single point.

We restrict our analysis to the special case of radial deformations, i.e., $u(x) = v(|x|)x/|x|$ with $v(r) > 0$ for $0 < r < 1$ and $v(1) = \lambda > 1$. In this case, the total

*Received by the editors May 4, 2006; accepted for publication (in revised form) January 24, 2007; published electronically August 29, 2007.
http://www.siam.org/journals/sima/39-3/65894.html

†Dipartimento di Matematica, Università degli Studi di Parma, V.le G. P. Usberti 53/A, I-43100 Parma, Italy (pietro.celada@unipr.it).
‡Dipartimento di Matematica Pura ed Applicata, Università degli Studi di Modena e Reggio Emilia, Via Campi 213/B, I-41100 Modena, Italy (perrotta@unimore.it).

energy, by a change of variables, turns out to be

$$E_0(u) = J_0(v) = \omega_N \int_0^1 r^{N-1} \Phi\left(v'(r)\,, \frac{v(r)}{r}\right) dr, \tag{1.2}$$

where ω_N is the $(N-1)$-dimensional measure of the surface of B_1 and Φ is associated with the stored energy density W; see section 2. We look for those radial deformations that minimize the total energy among all radial deformations and we emphasize here that, on the grounds of experimental evidence, we want to include among the feasible functions those v satisfying $v(0) > 0$, i.e., corresponding to deformations u which are singular at the origin. The variational problem thus obtained was considered by Ball in [1], and it is known as the problem of *cavitation*. Indeed, the existence of optimal radial deformations with $v(0) > 0$ for large enough displacement λ at the boundary can be interpreted as the occurrence of a spherical fracture—a *cavity*—inside the body. Ball in [1] extensively studied this problem with different possible boundary conditions—displacement and dead load traction—and for both compressible and incompressible materials. With regard to cavitation, we mention also [31], [19] as well as [21] and [7], which address the full three-dimensional problem. See also [13] for a description of cavitation in the language of currents. Among the results of [1], we are especially interested in those regarding isotropic, compressible materials whose stored energy density $W(Du)$, as a function of the singular values $\lambda_1, \dots, \lambda_N$ of the deformation gradient Du, takes the form

$$W(Du) = \sum_n w_0(\lambda_n) + w(\lambda_1 \cdots \lambda_N). \tag{1.3}$$

In Ball's paper [1], w_0 and w are suitable strictly convex functions with superlinear growth at infinity, and these hypotheses play a crucial role in the analysis. Ball then shows that minimizers in the restricted class of radial deformations exist and, moreover, that cavitation occurs for large enough displacements λ. Since Ball's seminal paper, the problem of cavitation has been extensively investigated, and several other aspects, including stability, have been considered. See [23], [25], [26], [27], [28] and [29], to mention just a few. We refer to [15] and the references therein for a recent survey of theoretical and experimental results about cavitation. Yet, apart from the contribution by Müller and Spector [21] mentioned above and related papers, investigation on cavitation has been carried out so far mostly in the radial case, and little is known in the unrestricted case, in particular, about the symmetry of possible minimizers. As to this question, we mention the uniqueness result by Knops and Stuart [17] and the existence result by Spector [30], which single out two situations where minimizers in the unrestricted case are radially symmetric. None of these results, however, applies to the case considered here.

The hypotheses of [1] do not include the case that w in (1.3) is nonconvex, i.e., that the model can possibly exhibit phase transition, and in particular do not include the case that w is given by

$$w_{BK}(t) = \frac{2\alpha}{t} + (1-\alpha)\left(2t + \frac{1}{t^2}\right), \qquad t > 0, \tag{1.4}$$

with $\alpha = -0.19$. Indeed, $w_{BK}(t) \to +\infty$ as $t \to 0_+$ and $t \to +\infty$ as expected, but w_{BK} is at the same time asymptotically linear and concave as $t \to +\infty$. This special choice of w was proposed by Blatz and Ko in [3] on the grounds of experimental results; see also [18] and [14]. It is supposed to describe the behavior of some foam rubberlike materials, now called Blatz–Ko materials.

The problem of cavitation was studied also by Marcellini in [19]. Marcellini's approach to the problem is based on the idea that, contrary to Ball's approach, the energy corresponding to a singular, radial deformation v featuring cavitation cannot be taken equal to (1.2) but must be defined elastically by lower semicontinuity or relaxation, i.e., choosing the energy of the radial deformation associated to v to be

$$J_V(v) = \inf \left\{ \liminf_{k\to+\infty} J_0(v_k) : \; v_k \to v \right\},$$

where the greatest lower bound is taken among all regular deformations v_k with $v_k(0) = 0$ (no cavitation occurs), and the convergence is in the strong Sobolev sense. Marcellini's main result is the derivation of a representation formula for the resulting relaxed energy J_V which turns out to be

$$J_V(v) = \omega_N \int_0^1 r^{N-1}\Phi\left(v'(r)\,, \frac{v(r)}{r}\right) dr + c\frac{\omega_N}{N}\left[v(0)\right]^N . \tag{1.5}$$

It is important to emphasize here that relaxation is made with respect to strong Sobolev convergence; otherwise the energy density Φ would be replaced by its convex envelope Φ^{**} with respect to v', in contrast with the experimental results of [3] which suggest for Blatz–Ko materials a nonconvex dependence on the determinant of the deformation gradient.

The additional term appearing in the relaxed energy J_V is thus proportional to the N-dimensional measure of the cavity and can be interpreted as the contribution to the total energy of the singular part of the Jacobian determinant of the radial deformation $u(x) = v(|x|)x/|x|$. The coefficient c appearing in (1.5) depends on Φ and, when Φ comes from a stored energy density W as those considered in [1] (see (1.3)), it is given by the recession w^∞ of the convex function w at $t = 1$ (see (2.4)). Clearly, the additional term in the definition of J_V penalizes the occurrence of cavitation, and, moreover, according to this model and contrary to Ball's, singular radial deformations require infinite energy for superlinear w since $w^\infty = +\infty$.

Following Müller and Spector's full three-dimensional model [21], a further radial model is considered in [4], where the energy associated with a radial deformation v is given by

$$J_S(v) = \omega_N \int_0^1 r^{N-1}\Phi\left(v'(r)\,, \frac{v(r)}{r}\right) dr + \omega_N w^\infty \left[v(0)\right]^{N-1} .$$

In this case, the contribution of the cavity to the total energy is proportional to the surface of the cavity as in [21], and we refer to [4] for a detailed discussion of the features of this model and for a comparative analysis with Ball's and Marcellini's models.

All the models considered so far thus lead, in the restricted case of radial deformations, to the problem of minimizing either an energy of the form J_0 defined by (1.2) when W is superlinear with respect to the Jacobian as in Ball's model [1] or an energy of the form (1.1) which we rewrite as

$$J(v) = \omega_N \int_0^1 r^{N-1}\Phi\left(v'(r)\,, \frac{v(r)}{r}\right) dr + \omega_N w^\infty \theta(v(0))$$

when W is linear in the Jacobian as in Marcellini's model [19] and in [4]. In these latter cases, the function $\theta\colon [0\,,+\infty) \to [0\,,+\infty)$ appearing in the definition of J is continuous, increasing, and null at zero, and the coefficient $0 < w^\infty < +\infty$ is given by (2.4). In both cases, minimization is subject to the one-point boundary condition $v(1) = \lambda > 1$.

When the stored energy density W is polyconvex, the resulting Φ is convex with respect to the derivative v' so that J_0 is lower semicontinuous for the natural convergence and the existence of minimizers for J_0 follows from standard arguments as in Ball [1]. If, in addition, the function

$$t \in [0\,,+\infty) \mapsto \frac{t^N}{N} - \theta(t)$$

is increasing, the same argument applies to J (Theorem 3.4). By contrast, when W fails to be polyconvex, Φ need not be convex with respect to v', and minimizers of J and J_0 need not exist. Our aim in this paper is the proof of fairly general attainment results for the minimum problems for J (Theorem 3.3) and J_0 (Theorem 3.5) when Φ is nonconvex with respect to v'. We show in particular (Theorem 2.1) that the result for J (linear growth) applies to the special choice of $w = w_{BK}$ corresponding to Blatz–Ko materials.

Our analysis in this paper mainly deals with the case of nonpolyconvex energy densities W that grow linearly as the Jacobian $\det Du \to +\infty$, and this leads to the one-dimensional integral J which comes from relaxing J_0 as in Marcellini [19]. Yet, the very same arguments and proofs apply unchanged to the superlinear case, which turns out to be even simpler. We thus emphasize that our existence results apply to elastomers which can possibly exhibit phase transition because of the lack of polyconvexity in both cases of "soft" elastomers, i.e., having linear growth in the Jacobian, as well as in the case of "hard" or nearly incompressible elastomers featuring very fast growth in the Jacobian. However, as a disclaimer, we wish to emphasize also that we do not address in this paper the issue of the onset of cavitation, i.e., whether optimal radial deformations for Blatz–Ko materials are singular at the origin or not, and we refer to [4] for a detailed discussion of this problem in the polyconvex case.

The issue of attainment for one-dimensional, nonconvex variational problems has been extensively studied in recent years, and we refer to the references in [20] for a far from exhaustive list of contributions on this subject. See also the references in [5] for more recent years. In particular, for the radial case we mention [6], [8], and [9]. The existence results we prove here are partially based on the ideas developed in [5].

The rest of the paper is organized as follows. In section 2, we briefly introduce the notation, we describe with more details the physical model and the variational problems we end up with, and we state our main result on the existence of optimal, radially symmetric deformations for Blatz–Ko materials (Theorem 2.1). Then, in section 3, we study the associated one-dimensional, nonconvex Bolza problem for J (linear growth) and we state an attainment result for it (Theorem 3.3). In the same section, we consider also the nonconvex variational problem for J_0 (superlinear growth) and we state an attainment result in this case as well (Theorem 3.5). We prove both results in section 4. Finally, in section 5, we prove Theorem 2.1 by showing that the existence result of section 3 applies to the w of Blatz–Ko materials given by (1.4).

2. Notation and description of the problem. We denote the norm of a vector x in $\mathbb{R}^N$ by $|x|$ and the scalar product of x and y by $\langle x, y\rangle$. If A is a subset of $\mathbb{R}^N$, we denote the interior, the closure, and the boundary of A by $\mathrm{int}(A)$, $\overline{A}$, and ∂A, respectively.

As for matrices, let $\mathbb{M}^{N\times N}$ be the set of all $N\times N$ real matrices $A = (A^m_n)_{m,n=1,\dots,N}$ endowed with the euclidean norm denoted by $|A|$. Let $\mathbb{I}_N$ be the identity matrix and let $\mathrm{Adj}\, A$ be the $N\times N$ matrix defined as the transpose of the *cofactors* of A so that Laplace's formula yields

$$\det A = \sum_{1\le n\le N} A^1_n\,(\mathrm{Adj}\, A)^1_n\,.$$

The *singular values* of the matrix A are the eigenvalues $\lambda_1(A),\dots,\lambda_N(A)$ of the positive, symmetric matrix $\sqrt{AA^T}$ so that

$$|A|^2 = \lambda_1^2(A) + \cdots + \lambda_N^2(A) \quad \text{and} \quad |\det A| = \lambda_1(A)\cdots\lambda_N(A).$$

The standard basis of $\mathbb{R}^N$ is denoted by $\{e_1,\dots,e_N\}$, and the *tensor product* of two vectors $a = a^1e_1 + \cdots + a^Ne_N$ and $b = b^1e_1 + \cdots + b^Ne_N$ is the rank-one matrix $a\otimes b$ defined by $(a\otimes b)^m_n = a^mb^n$ for every m and n. Finally, we denote the group of all matrices with positive determinant by $\mathbb{M}^{N\times N}_+$ and its special orthogonal subgroup by $SO(N)$.

With regard to measure and functional theoretic notation, we denote the Lebesgue measure of a measurable subset E in some euclidean space $\mathbb{R}^n$ by $|E|$. We use standard notation for the spaces of continuously differentiable functions as well as for Lebesgue and Sobolev spaces and their norms. In the special case of functions of one variable on a bounded interval I, we let $AC(I)$ and $AC_{\mathrm{loc}}(I)$ be the space of all absolutely continuous functions on I and on all compact subintervals of I, respectively.

In what follows, we shall consider the Jacobian determinant of Sobolev mappings $u\colon \Omega\to\mathbb{R}^N$, $u = (u^1,\dots,u^N)$, where Ω is an open subset of $\mathbb{R}^N$. If $u\in W^{1,p}(\Omega,\mathbb{R}^N)$ for some $p\ge 1$, we denote by $\det Du$ the measurable function defined a.e. as the *pointwise Jacobian determinant* of u which is in $L^1(\Omega)$ for $p\ge N$. If $p<N$, $\det Du$ loses its natural meaning, and for $p\ge N^2/(N+1)$ or when $u\in W^{1,N-1}(\Omega,\mathbb{R}^N)\cap L^\infty(\Omega,\mathbb{R}^N)$, we consider instead the *distributional Jacobian determinant* of u defined as the distributional divergence

$$\mathrm{Det}\, Du = \sum_{1\le n\le N} D_n\left(u^1\,(\mathrm{Adj}\, Du)^1_n\right).$$

In general, $\mathrm{Det}\, Du$ is not a function, and $\mathrm{Det}\, Du \ne \det Du$ in the sense of distribution. However, if u is in $W^{1,p}(\Omega,\mathbb{R}^N)$ with $p\ge N^2/(N+1)$ and $\mathrm{Det}\, Du$ is a Radon measure, then $\det Du$ is the Radon–Nikodym derivative (density) of the absolutely continuous part of $\mathrm{Det}\, Du$ with respect to the Lebesgue measure and we denote the *singular part* of $\mathrm{Det}\, Du$ by $(\mathrm{Det}\, Du)^{\mathrm{s}}$. We refer to [12] for further results on the relation between det and Det and the structure of the distributional Jacobian determinant.

As motivated in the introduction, we are interested in studying the deformations of a hyperelastic, homogeneous, solid body whose reference configuration is the open unit ball B_1 of $\mathbb{R}^N$, the physically interesting case being obviously $N=2$ and $N=3$. We assume that the *stored energy density* of the body is a smooth function $W\in\mathcal{C}^1(\mathbb{M}^{N\times N}_+)$, $W\ge 0$, which can be written as

$$W(A) = W_0(A) + w(\det A), \qquad A\in\mathbb{M}^{N\times N}_+,$$

where the first term $W_0 \in \mathcal{C}^1(\mathbb{M}_+^{N\times N})$, $W_0 \geq 0$, is *frame indifferent* and *isotropic*; i.e.,

$$W_0(QA) = W_0(A) = W_0(AQ), \qquad A \in \mathbb{M}_+^{N\times N}, \quad Q \in SO(N). \tag{2.1}$$

With regard to the behavior of W_0 for large $|A|$, we assume that W_0 has superlinear growth

$$\lim_{|A|\to+\infty} \frac{W_0(A)}{|A|} = +\infty, \tag{2.2}$$

but at the same time, as we are interested in possibly discontinuous deformations, we assume also that W_0 has polynomial growth of order strictly less than N, i.e.,

$$0 \leq W_0(A) \leq C\,(1 + |A|^p)\,, \qquad A \in \mathbb{M}_+^{N\times N}, \tag{2.3}$$

for some $C \geq 0$ and $1 < p < N$. As for the term w depending on the deformation of volume elements, we assume that $w \in \mathcal{C}^1(\mathbb{R}_+)$, $w \geq 0$, is such that

$$\lim_{t\to 0_+} w(t) = +\infty \quad \text{ and } \quad \lim_{t\to+\infty} \frac{w(t)}{t} = w^\infty \in (0\,,+\infty]. \tag{2.4}$$

Hence, linear growth corresponds to $0 < w^\infty < +\infty$ and superlinear growth to $w^\infty = +\infty$. We emphasize again that in both cases w need not be convex and that, for Blatz–Ko materials, (2.4) holds with $w^\infty = w^\infty_{BK} = 2(1-\alpha)$ where $\alpha = -0.19$, i.e., $w^\infty_{BK} = 2.38$.

The properties of frame indifference and isotropy (2.1) of the function W_0 obviously hold for the entire energy W as well, and, as is well known, both W_0 and W can then be written as functions of the singular values $\lambda_1 = \lambda_1(A), \dots, \lambda_N = \lambda_N(A)$ of the gradient matrix A, i.e.,

$$\begin{aligned} W_0(A) &= \Phi_0\,(\lambda_1\,, \dots, \lambda_N)\,, \\ W(A) &= \Phi\,(\lambda_1\,, \dots, \lambda_N) = \Phi_0\,(\lambda_1\,, \dots, \lambda_N) + w\,(\lambda_1 \cdots \lambda_N) \end{aligned} \tag{2.5}$$

for symmetric functions Φ_0, $\Phi\colon \mathbb{R}^N_+ \to [0\,,+\infty)$ satisfying the appropriate growth properties corresponding to (2.2) and (2.3). Moreover, Φ_0 and Φ share the same smoothness with W_0 and W, respectively.

For this energy density W, the total energy associated with a smooth deformation u is given by the integral

$$E_0(u) = \int_{B_1} W(Du(x))\,dx. \tag{2.6}$$

We want to consider bounded deformations which are possibly singular at the origin, i.e., deformations $u\colon B_1 \to \mathbb{R}^N$ corresponding to possibly discontinuous Sobolev functions $u \in L^\infty(B_1, \mathbb{R}^N) \cap W^{1,1}(B_1, \mathbb{R}^N)$ satisfying an appropriate notion of invertibility; see [21] for a general discussion of this issue. For these Sobolev mappings, however, the pointwise Jacobian determinant $\det Du$ loses its natural meaning and has to be replaced by the distributional Jacobian determinant $\mathrm{Det}\, Du$. A general discussion of this question as in Chapter 2 of [13] or [21]—admissible class of deformations and possible definitions of the energy when the pointwise Jacobian determinant is meaningless—goes far beyond the aim of this paper. Indeed, here we shall consider only possibly singular, radial deformations for which invertibility can be stated in elementary terms and for which a reasonable definition of the total energy encompassing the nonregular part of the distributional Jacobian determinant can be easily given. Accordingly, we shall assume that the total energy of a deformation $u \in L^\infty(B_1,\mathbb{R}^N)\cap W^{1,1}(B_1,\mathbb{R}^N)$ such that $\operatorname{Det} Du$ is a nonnegative Radon measure is given by E_0 as in Ball's case when W has superlinear growth with respect to the Jacobian; i.e., the energy depends only on the regular part of the deformation gradient. By contrast, when W is linear with respect to the Jacobian, the energy is the sum of E_0 and a term depending on the total mass of the singular part of $\operatorname{Det} Du$, i.e.,

$$E(u) = E_0(u) + w^\infty \Theta\left((\operatorname{Det} Du)^{\mathrm{s}}(B_1)\right), \tag{2.7}$$

where $\Theta\colon [0\,,+\infty) \to [0\,,+\infty)$ is a continuous, increasing function which vanishes at zero. This special class of deformations is the class of *radial deformations* for which no eversion occurs, i.e., mappings $u \in L^\infty(B_1,\mathbb{R}^N)$ such that

$$u(x) = v(|x|)\frac{x}{|x|} \qquad \text{for a.e. } x \in B_1 \tag{2.8}$$

for some $v \in L^\infty(0\,,1)$ satisfying $v > 0$ a.e. on $(0\,,1)$. It is clear that v is uniquely associated with u up to a null set by (2.8) and vice versa. It is then easy to check (see Lemma 4.1 in [1]) that, whenever the two measurable functions $u\colon B_1 \to \mathbb{R}^N$ and $v\colon (0\,,1] \to [0\,,+\infty)$ are related by (2.8), we have that

$$u \in W^{1,p}(B_1,\mathbb{R}^N) \quad \Longleftrightarrow \quad \begin{cases} v \in AC_{\mathrm{loc}}((0\,,1]) \\ \text{and} \\ \displaystyle\int_0^1 r^{N-1}\left(|v'(r)|^p + \left|\frac{v(r)}{r}\right|^p\right) dr < +\infty \end{cases} \tag{2.9}$$

for every index $1 \le p < +\infty$. In this case, the gradient of u and its singular values are given by

$$Du(x) = \frac{v(|x|)}{|x|}\mathbb{I}_N + \left(v'(|x|) - \frac{v(|x|)}{|x|}\right)\frac{x\otimes x}{|x|^2} \qquad \text{for a.e. } x \in B_1 \tag{2.10}$$

and

$$\begin{cases} \lambda_1(Du(x)) = v'(|x|), \\ \lambda_n(Du(x)) = \dfrac{v(|x|)}{|x|}, \quad n = 2,\dots,N, \end{cases} \qquad \text{for a.e. } x \in B_1. \tag{2.11}$$

It follows from (2.9) that a radial deformation u with a cavity $v(0) > 0$ can be in $W^{1,p}(B_1,\mathbb{R}^N)$ with at most $p < N$.

We shall assume throughout the paper that u defined by (2.8) is such that the corresponding v can be chosen to be strictly increasing. Thus, u is injective and v is actually defined up to a countable set, and we assume also that it is defined by continuity at $r = 0$ and $r = 1$. With this additional assumption, it follows easily that the equivalence (2.9) actually holds with $v \in AC([0\,,1])$, and, moreover, for these mappings u satisfying (2.8) and (2.9) for some $p \ge 1$, the distributional Jacobian determinant is a nonnegative Radon measure whose absolutely continuous part with respect to the Lebesgue measure has density

$$\det Du(x) = v'(|x|)\left(\frac{v(|x|)}{|x|}\right)^{N-1} \qquad \text{for a.e. } x \in B_1 \tag{2.12}$$

and whose singular part is

$$(\operatorname{Det} Du)^{\mathrm{s}} = \frac{\omega_N}{N}[v(0)]^N \delta_0, \tag{2.13}$$

where δ_0 is the Dirac measure at the origin.

For the energies E and E_0 defined by (2.7) and (2.6), respectively, we shall consider the radial displacement boundary value problem in the class of radial deformations, i.e., the variational problem of minimizing E and E_0 among all radial deformations $u \in L^\infty(B_1, \mathbb{R}^N) \cap W^{1,1}(B_1, \mathbb{R}^N)$ satisfying $\det Du > 0$ a.e. on B_1 and the boundary condition $u(x) = \lambda$ for $|x| = 1$ for some $\lambda > 1$. The set of all functions v associated with these radial deformations u by (2.8) is the set

$$\mathcal{A} = \{v \in AC([0\,,1]) :\ v > 0 \text{ on } (0\,,1] \text{ and } v' > 0 \text{ a.e. on } (0\,,1]\}\,, \tag{2.14}$$

and we denote by $\mathcal{A}(\lambda)$ those $v \in \mathcal{A}$ such that $v(1) = \lambda$. Note also that

$$\int_0^1 r^{N-1}\left[v'(r) + \frac{v(r)}{r}\right] dr < +\infty, \qquad v \in \mathcal{A}(\lambda), \tag{2.15}$$

which is the second condition in (2.9). By a change of variables and by (2.5) and (2.11), we have

$$E_0(u) = \omega_N \int_0^1 r^{N-1}\Phi\left(v'(r)\,, \frac{v(r)}{r}\right) dr,$$

where

$$\Phi(\xi\,,\eta) = \Phi_0(\xi\,,\eta) + w(\xi\eta^{N-1}), \qquad \eta\,,\xi > 0,$$

are defined by (2.5) and we have used the shortcuts $\Phi_0(\xi\,,\eta) = \Phi_0(\xi\,,\eta\,,\ldots\,,\eta)$ and $\Phi(\xi\,,\eta) = \Phi(\xi\,,\eta\,,\ldots\,,\eta)$. Similarly, as the singular part of $\operatorname{Det} Du$ for a radial deformation u is a function of $v(0)$ for the corresponding v by (2.13), E defined by (2.7) becomes

$$E(u) = \omega_N \int_0^1 r^{N-1}\Phi\left(v'(r)\,, \frac{v(r)}{r}\right) dr + \omega_N w^\infty \theta\left(v(0)\right),$$

where $\theta\colon [0\,,+\infty) \to [0\,,+\infty)$ is a continuous and increasing function such that $\theta(0) = 0$. Thus, we are led to consider the variational problem

$$\min\left\{J_0(v) :\ v \in \mathcal{A}(\lambda)\right\}, \tag{$\mathcal{P}_0$}$$

where

$$J_0(v) = \omega_N \int_0^1 r^{N-1}\Phi\left(v'(r)\,, \frac{v(r)}{r}\right) dr, \qquad v \in \mathcal{A}(\lambda), \tag{2.16}$$

in the case of superlinear growth and the Bolza problem

$$\min\left\{J(v) :\ v \in \mathcal{A}(\lambda)\right\}, \tag{$\mathcal{P}$}$$

where

$$J(v) = J_0(v) + \omega_N w^\infty \theta\left(v(0)\right), \qquad v \in \mathcal{A}(\lambda), \tag{2.17}$$

in the linear case.

In section 3, we shall prove two general existence results for the Bolza problem $(\mathcal{P})$ (Theorem 3.3) and the variational problem $(\mathcal{P}_0)$ (Theorem 3.5) which apply to stored energy densities Φ which fail to be convex with respect to the derivative v'. As a consequence of Theorem 3.3, we shall prove the following existence result (Theorem 2.1) of optimal, radial solutions for Blatz–Ko materials.

THEOREM 2.1. *Let*

$$W(A) = |A|^2 + w_{BK}(\det A), \qquad A \in \mathbb{M}_+^{N\times N},$$

where $w_{BK}\colon (0\,,+\infty) \to [0\,,+\infty)$ *is given by* (1.4), *and let* $\Theta\colon [0\,,+\infty) \to [0\,,+\infty)$ *be a continuous and increasing function such that*

(a) $\Theta(0) = 0$;

(b) $s \in [0\,,+\infty) \mapsto s - \Theta(s)$ *is increasing.*

Then, the variational integral

$$E(u) = \int_{B_1} W(Du(x))\,dx + w_{BK}^{\infty}\Theta\left((\operatorname{Det} Du)^s(B_1)\right)$$

with $w_{BK}^{\infty} = 2.38$ *has a minimizer among all radial deformations* $u(x) = v(|x|)\frac{x}{|x|}$ *corresponding to* $v \in \mathcal{A}(\lambda)$.

The proof is given in section 5. Here, we point out again that this theorem does not assert that optimal radial deformations for Blatz–Ko materials exhibit cavitation. The issue of the existence of singular radial solutions is more subtle, even for polyconvex energies W, and we refer to [4] for a thorough discussion of this issue in the polyconvex case.

3. The nonconvex, one-dimensional problems. In this section we state two fairly general attainment results for the one-dimensional, one-point boundary value problems $(\mathcal{P}_0)$ and $(\mathcal{P})$ associated with the problem of minimizing, among radial deformations, the energies E_0 and E defined by (2.6) and (2.7), respectively. We shall prove both results in section 4.

We begin by recalling some elementary results from convex analysis. Consider a lower semicontinuous function $\varphi\colon \mathbb{R}_+ \to [0\,,+\infty)$ such that $\varphi(\xi) \to +\infty$ as $\xi \to 0_+$ and $\varphi(\xi)/\xi \to +\infty$ as $\xi \to +\infty$. The *polar function* of φ is the lower semicontinuous, convex function $\varphi^*\colon \mathbb{R} \to \mathbb{R}$ defined by

$$\varphi^*(\xi) = \sup\left\{\xi\zeta - \varphi(\xi) :\ \xi > 0\right\}, \qquad \zeta \in \mathbb{R}$$

(see [11] or [24]), and the *bipolar function* or *convex envelope* of φ is the polar $\varphi^{**}\colon \mathbb{R}_+ \to [0\,,+\infty)$ of φ^*. Thus, φ^{**} is convex, continuous, and such that

(3.1) $\varphi^{**}(\xi) \le \varphi(\xi)$ for every $\xi > 0$;

(3.2) the open set $\{\varphi^{**} < \varphi\}$ has bounded connected components;

(3.3) the closure of each connected component of $\{\varphi^{**} < \varphi\}$ is contained in $\mathbb{R}_+$;

(3.4) φ^{**} is affine on each connected component of $\{\varphi^{**} < \varphi\}$.

Moreover, it is easy to check that whenever $\varphi \in \mathcal{C}^1(\mathbb{R}_+)$, then $\varphi^{**} \in \mathcal{C}^1(\mathbb{R}_+)$ as well and that the values of φ^{**} at $\xi > 0$ and φ^* at $d = (\varphi^{**})'(\xi)$ are related by

$$\varphi^{**}(\xi) + \varphi^*(d) = d\xi \tag{3.5}$$

because of the equality $\varphi^{***} = \varphi^*$ (see [11]). Thus, the convexity inequality

$$\varphi^{**}(\xi) \geq \varphi^{**}(\xi_0) + d(\xi - \xi_0), \qquad \xi\,,\,\xi_0 > 0,$$

where $d = (\varphi^{**})'(\xi_0)$ can be written as

$$\varphi^{**}(\xi) \geq d\xi - \varphi^*(d), \qquad \xi > 0, \tag{3.6}$$

and $-\varphi^*(d)$ yields the value at the origin of the supporting affine function to the graph of φ^{**} at the point ξ. Moreover, because of (3.4), whenever an interval $(\alpha\,,\beta)$ is a connected component of $\{\varphi^{**} < \varphi\}$, (3.6) turns into the equality

$$\varphi^{**}(\xi) = d\xi - \varphi^*(d), \qquad \xi \in (\alpha\,,\beta), \tag{3.7}$$

where $d = (\varphi^{**})'(\xi_0)$ and $\xi_0 \in (\alpha\,,\beta)$.

We can now describe the class of one-dimensional integrands that we shall consider in this section. Recalling the discussion of section 2 on the stored energy W and the formula (2.11) for the singular values of the gradient of radial deformations, we consider a function $\Phi\colon \mathbb{R}^2_+ \to [0\,,+\infty)$ which is the sum of two terms, i.e.,

$$\Phi(\xi\,,\eta) = \Phi_0(\xi\,,\eta) + w(\xi\eta^{N-1}), \qquad \xi\,,\,\eta > 0, \tag{H1}$$

where $\Phi_0\colon \mathbb{R}^2_+ \to [0\,,+\infty)$ and $w\colon \mathbb{R}_+ \to [0\,,+\infty)$. We assume first that both terms are smooth; i.e.,

$$\Phi_0 \in \mathcal{C}^1(\mathbb{R}^2_+) \text{ and } w \in \mathcal{C}^1(\mathbb{R}_+). \tag{H2}$$

Then, recalling the meaning of ξ and η and (2.11), we set

$$\|(\xi\,,\eta)\| = \sqrt{\xi^2 + (N-1)\eta^2}, \qquad \xi\,,\,\eta > 0,$$

for the euclidean norm of the matrix corresponding to the singular values given by $(\xi\,,\eta,\ldots,\eta)$ and, in view of the hypotheses (2.2) and (2.3) on W_0, we assume that

$$\lim_{\|(\xi\,,\eta)\|\to+\infty} \frac{\Phi_0(\xi\,,\eta)}{\|(\xi\,,\eta)\|} = +\infty, \tag{H3}$$

$$0 \leq \Phi_0(\xi\,,\eta) \leq C\,(1 + \|(\xi\,,\eta)\|^p), \qquad \xi,\,\eta > 0, \tag{H4}$$

for some constant $C \geq 0$ and some index $1 < p < N$. Similarly, in view of (2.4), we assume that w is such that either

$$\lim_{t\to 0_+} w(t) = +\infty \text{ and } \lim_{t\to+\infty} \frac{w(t)}{t} = w^\infty \in (0\,,+\infty) \tag{H5$_\mathrm{L}$}$$

(linear case) or

$$\lim_{t\to 0_+} w(t) = +\infty \text{ and } \lim_{t\to+\infty} \frac{w(t)}{t} = w^\infty = +\infty \tag{H5$_\mathrm{S}$}$$

(superlinear case). We remark again that w need not be convex and we note that (H4) and either (H5$_{\mathrm{L}}$) or (H5$_{\mathrm{S}}$) imply that

$$\lim_{\xi\to+\infty}\frac{\Phi(\xi\,,\xi)}{\xi^N}=w^\infty. \tag{3.8}$$

For any such Φ, we consider the variational integrals J_0 and J defined by (2.16) and (2.17), respectively, and the corresponding variational problems $(\mathcal{P}_0)$ and $(\mathcal{P})$ for $\lambda>1$. Moreover, we denote the polar of Φ and the convex envelope of Φ with respect to the first variable ξ by $\Phi^*\colon\mathbb{R}\times\mathbb{R}_+\to\mathbb{R}$ and $\Phi^{**}\colon\mathbb{R}^2_+\to[0\,,+\infty)$, respectively; i.e., $\Phi^*(\xi\,,\eta)$ and $\Phi^{**}(\xi\,,\eta)$ are the polar and the convex envelope of the function $\xi'\to\Phi(\xi'\,,\eta)$ at the point ξ.

In what follows, a special role is played by the *detachment set* $\mathcal{D}$ defined by

$$\mathcal{D}=\{(\xi\,,\eta):\ \Phi^{**}(\xi\,,\eta)<\Phi(\xi\,,\eta)\}\,. \tag{3.9}$$

For every $\xi>0$ and $\eta>0$, its horizontal and vertical sections will be denoted by

$$\mathcal{D}^\eta=\{\xi>0:\ \Phi^{**}(\xi\,,\eta)<\Phi(\xi\,,\eta)\}\quad\text{and}\quad\mathcal{D}_\xi=\{\eta>0:\ \Phi^{**}(\xi\,,\eta)<\Phi(\xi\,,\eta)\}\,,$$

respectively. Also, whenever $(\xi\,,\eta)\in\mathcal{D}$, the connected component of $\mathcal{D}_\xi$ containing η will be denoted by $\mathcal{D}_\xi(\eta)$, and similarly for $\mathcal{D}^\eta(\xi)$. The main properties of the detachment set $\mathcal{D}$ are listed in the following proposition for which we refer to Proposition 3.1 in [5].

PROPOSITION 3.1. *Let $\Phi\colon\mathbb{R}^2_+\to[0\,,+\infty)$ be a continuously differentiable function such that*

$$\lim_{\xi\to+\infty}\frac{\Phi(\xi\,,\eta)}{\xi}=+\infty\ \textit{uniformly with respect to}\ \eta>0; \tag{3.10}$$

$$\lim_{(\xi\,,\eta)\to(\xi_0\,,\eta_0)}\Phi(\xi\,,\eta)=+\infty\ \textit{whenever}\ \xi_0\eta_0=0. \tag{3.11}$$

Then,

(a) *the detachment set $\mathcal{D}$ is open;*

(b) *for every $0<a<b$, the connected components of $\mathcal{D}\cap(\mathbb{R}\times(a\,,b))$ are bounded.*

Moreover, for every $(\xi_0\,,\eta_0)\in\mathcal{D}$, there exist $\delta=\delta(\xi_0\,,\eta_0)>0$ and two functions $d^\pm\colon[\eta_0-\delta\,,\eta_0+\delta]\to\mathbb{R}$ such that

(c) *$d^-(\eta)<\xi_0<d^+(\eta)$ for every $\eta\in[\eta_0-\delta\,,\eta_0+\delta]$;*

(d) *d^+ and d^- are bounded, upper and lower semicontinuous functions, respectively;*

(e) *$\mathcal{D}^\eta(\xi_0)=(d^-(\eta)\,,d^+(\eta))$ for every $\eta\in[\eta_0-\delta\,,\eta_0+\delta]$;*

(f) *$\Phi^{**}(d^\pm(\eta)\,,\eta)=\Phi(d^\pm(\eta)\,,\eta)$ for every $\eta\in[\eta_0-\delta\,,\eta_0+\delta]$.*

Note that this result applies to every Φ satisfying (H1), ..., (H4) and either (H5$_{\mathrm{L}}$) or (H5$_{\mathrm{S}}$). As for the properties of $\mathcal{D}$ listed above, note that, with different words, (b) states that, in the $\xi\eta$ plane, every connected component of every horizontal strip of $\mathcal{D}$ is bounded, and (e) states that, provided that the strip is narrow enough, every such connected component is the plane set contained between the graphs of two functions $d^\pm$ satisfying (d).

In the following proposition, we describe the properties of the convex envelope Φ^{**} of Φ that will be used in what follows. We refer to [2] and [16] for a more detailed discussion of the regularity properties of convex envelopes. We are also indebted to J. Kristensen for pointing us to the proof of (a) in [2].

Proposition 3.2. *Let*

$$\Phi(\xi\,,\eta) = \Phi_0(\xi\,,\eta) + w(\xi\eta^{N-1}), \qquad \xi\,,\,\eta > 0,$$

satisfy (H1), ..., (H4) *and either* ($\mathrm{H5_L}$) *or* ($\mathrm{H5_S}$). *Then,*

(a) Φ^{**} *is continuous;*

(b) $\lim_{\|(\xi\,,\eta)\|\to+\infty} \frac{\Phi^{**}(\xi\,,\eta)}{\|(\xi\,,\eta)\|} = +\infty$*;*

(c) $(w^{**})^\infty = w^\infty$ *and* $\lim_{\xi\to+\infty} \frac{\Phi^{**}(\xi\,,\xi)}{\xi^N} = w^\infty$*;*

(d) *the partial derivative* $m(\xi\,,\eta) = \Phi^{**}_\xi(\xi\,,\eta)$ *exists at every point* $(\xi\,,\eta) \in \mathcal{D}$ *and is continuous on* $\mathcal{D}$*;*

(e) *the function* $q(\xi\,,\eta) = -\Phi^*(m(\xi\,,\eta)\,,\eta)$, $(\xi\,,\eta) \in \mathcal{D}$, *is continuous on* $\mathcal{D}$*;*

(f) *the restrictions* $\xi \in \mathcal{D}^\eta \mapsto m(\xi\,,\eta)$ *and* $\xi \in \mathcal{D}^\eta \mapsto q(\xi\,,\eta)$ *are constant on each connected component of* $\mathcal{D}^\eta$*;*

(g) *if each connected component of* $\mathcal{D}$ *is simply connected, there exists a continuous function* $M\colon \mathcal{D} \to \mathbb{R}$ *such that the partial derivative* M_η *exists in* $\mathcal{D}$ *and*

$$M_\eta(\xi\,,\eta) = m(\xi\,,\eta), \qquad (\xi\,,\eta) \in \mathcal{D}. \tag{3.12}$$

Moreover, $\xi \in \mathcal{D}^\eta \mapsto M(\xi\,,\eta)$ *is locally constant on* $\mathcal{D}^\eta$*.*

It follows from (b), (d), and (e), that m and q are locally constant with respect to ξ on $\mathcal{D}$; i.e., for every $(\xi_0\,,\eta_0) \in \mathcal{D}$ we have

$$\begin{cases} m(\xi\,,\eta) = m(\xi_0\,,\eta), \\ q(\xi\,,\eta) = q(\xi_0\,,\eta), \end{cases} \qquad \xi \in \mathcal{D}^\eta(\xi_0), \quad \eta \in \mathcal{D}_{\xi_0}(\eta_0). \tag{3.13}$$

Moreover, it follows also from (e) and (3.7) that, on the connected component of $\mathcal{D}^\eta$ containing ξ_0, the convex envelope Φ^{**} can be written as an affine function of ξ, i.e.,

$$\Phi^{**}(\xi\,,\eta) = m(\xi_0\,,\eta)\xi + q(\xi_0\,,\eta), \qquad \xi \in \mathcal{D}^\eta(\xi_0), \quad \eta \in \mathcal{D}_{\xi_0}(\eta_0). \tag{3.14}$$

Finally, as for (g), it is clear that, even if $\mathcal{D}$ has multiply connected components, a continuous function M satisfying (3.12) can be locally defined anyway.

Proof of Proposition 3.2. Property (a) is essentially known (see [2] for instance), and (b) follows from (H3). As for (c), the first equality is easy, and hence (H4) and

$$w^{**}(\xi\eta^{N-1}) \le \Phi^{**}(\xi\,,\eta) \le C\,(1 + \|(\xi\,,\eta)\|^p) + w(\xi\eta^{N-1}), \qquad \xi\,,\,\eta > 0,$$

yield the conclusion in both cases of ($\mathrm{H5_L}$) and ($\mathrm{H5_S}$) because $1 < p < N$. The remaining properties (d), (e), and (f) are proved in Proposition 2.1 in [5], and, finally, (g) is obvious. □

After these preliminaries, we can turn to the existence result for the nonconvex Bolza problem ($\mathcal{P}$). To this aim, we consider the relaxed problem

$$\min\left\{J^{**}(v)\colon\ v \in \mathcal{A}(\lambda)\right\}, \tag{$\mathcal{P}^{**}$}$$

where

$$J^{**}(v) = \omega_N \int_0^1 r^{N-1}\Phi^{**}\left(v'(r)\,,\frac{v(r)}{r}\right)dr + \omega_N w^\infty \theta\left(v(0)\right), \qquad v \in \mathcal{A}(\lambda),$$

and, for the sake of simplicity, we state the attainment result for $(\mathcal{P})$ with the following additional hypothesis:

(H6) each connected component of $\mathcal{D}$ is simply connected,

so that (g) of Proposition 3.2 holds. As we shall see at point (c) after Theorem 3.3 below, this additional hypothesis gives a simpler statement but does not affect the scope of application of the existence result.

THEOREM 3.3. *Let*

$$\Phi(\xi,\eta) = \Phi_0(\xi,\eta) + w(\xi\eta^{N-1}), \qquad \xi,\eta > 0,$$

be such that (H1), ..., ($H5_L$) *and* (H6) *hold and let* $\theta\colon [0,+\infty) \to [0,+\infty)$ *be a continuous and increasing function such that* $\theta(0) = 0$. *Set*

$$\psi(\xi,\eta) = -NM(\xi,\eta) + m(\xi,\eta)\eta + q(\xi,\eta), \qquad (\xi,\eta) \in \mathcal{D}, \tag{3.15}$$

and assume also that the following properties of ψ *hold at every point* $(\xi_0,\eta_0) \in \mathcal{D}$:

(3.16) *there exists* $\delta = \delta(\xi_0,\eta_0) > 0$ *such that* $[\eta_0 - \delta, \eta_0 + \delta] \subset \mathcal{D}_{\xi_0}(\eta_0)$ *and such that the section* $\eta \in \mathcal{D}_{\xi_0} \to \psi(\xi_0,\eta)$ *is monotone on both intervals* $[\eta_0 - \delta, \eta_0]$ *and* $[\eta_0, \eta_0 + \delta]$;

(3.17) *if* $\xi_0 = \eta_0$, *then the section* $\eta \in \mathcal{D}_{\xi_0} \to \psi(\xi_0,\eta)$ *has no strict local minima at* $\eta = \eta_0$.

Then, the minimum problem $(\mathcal{P})$ *admits a solution whenever* $(\mathcal{P}^{**})$ *has a solution.*

Before giving sufficient conditions for the existence of solutions to the relaxed problem $(\mathcal{P}^{**})$ (see Theorem 3.4 below), we want to make a few remarks about the hypotheses and the proof of Theorem 3.3.

(a) If the detachment set $\mathcal{D}$ does not meet the line $\xi = \eta$, attainment for the variational problem $(\mathcal{P})$ holds with (3.16) only, which is a mild assumption on the behavior of the convex envelope Φ^{**}. As we shall see in detail in section 5, this is the case of Blatz–Ko materials where $w = w_{BK}$.

(b) By (3.13), ψ is locally constant with respect to ξ; i.e., $\psi(\xi_1,\eta) = \psi(\xi_2,\eta)$ for every $(\xi_i,\eta) \in [\xi_0 - \rho, \xi_0 + \rho] \times [\eta_0 - \delta, \eta_0 + \delta] \subset \mathcal{D}$. Moreover, it follows from (3.12) and (3.14) that in the simplest case when all sections $\mathcal{D}_\xi$ of $\mathcal{D}$ are intervals, (H6) obviously holds and ψ can be explicitly written as

$$\psi(\xi,\eta) = -N\int \Phi^{**}_\xi(\xi,\eta)\,d\eta + \Phi^{**}_\xi(\xi,\eta)(\eta - \xi) + \Phi^{**}(\xi,\eta), \qquad (\xi,\eta) \in \mathcal{D}.$$

(c) Theorem 3.3 remains valid even if $\mathcal{D}$ has multiply connected components. Indeed, in that case, ψ can be defined only locally by (3.15), on every convex open subset of $\mathcal{D}$, for instance, and any two such locally defined functions ψ differ only by a constant on the intersections of their domains. As the other hypotheses (3.16) and (3.17) on ψ have local nature, the theorem remains true with the very same proof even if (H6) fails.

(d) The proof of Theorem 3.3 follows a somewhat standard path for this kind of variational problem. We start with a solution $\overline{v}$ to the relaxed problem $(\mathcal{P}^{**})$ which exists by assumption and we show that $\overline{v}$ can be modified so as to find a new solution v to $(\mathcal{P}^{**})$ satisfying $v(0) = \overline{v}(0)$ and the differential relation

$$\left(v'(r), \frac{v(r)}{r}\right) \notin \mathcal{D} \qquad \text{for a.e. } r \in (0,1). \tag{3.18}$$

Thus, $\Phi^{**}(v'(r)\,,v(r)/r)=\Phi(v'(r)\,,v(r)/r)$ for a.e. $r\in(0\,,1)$, and the minimality of v for J follows staightforwardly from the corresponding property for J^{**} and the convexity inequality $\Phi^{**}\le\Phi$ on $\mathbb{R}^2_+$. As is to be expected, the most technically difficult part of this program is the definition of new solutions v to $(\mathcal{P}^{**})$ satisfying the differential relation (3.18). For this part of the proof, we shall adapt the ideas developed in [5].

We now proceed to give sufficient conditions for attainment for the relaxed problem $(\mathcal{P}^{**})$. The proof of the following result is essentially the same as Marcellini's relaxation result; see Theorem 1 in [19].

THEOREM 3.4. *Let*

$$\Phi(\xi\,,\eta)=\Phi_0(\xi\,,\eta)+w(\xi\eta^{N-1}),\qquad \xi\,,\eta>0,$$

be such that (H1)$,\ldots,$(H5$_\mathrm{L}$) *hold and let* $\theta\colon[0\,,+\infty)\to[0\,,+\infty)$ *be a continuous and increasing function such that*

(a) $\theta(0)=0$*;*

(b) $\Delta(t)=t^N/N-\theta(t)$*,* $t\ge0$*, is increasing.*

Then, the minimum problem $(\mathcal{P}^{**})$ *admits a solution.*

Proof. First, set $v_\lambda(r)=\lambda r$ for every $r\in(0\,,1]$ and note that J^{**} is finite at v_λ. Then, as Φ^{**} is now convex with respect to ξ, attainment for $(\mathcal{P}^{**})$ can be proved by Tonelli's direct method, and the proof follows immediately from the following two claims.

Claim 1. For every minimizing sequence $\{v_k\}_k\subset\mathcal{A}(\lambda)$, there exists $v\in\mathcal{A}(\lambda)$ and a subsequence (not relabeled) $\{v_k\}_k$ such that $v_k\rightharpoonup v$ weakly in $\mathcal{A}(\lambda)$; i.e.,

$$\begin{cases} v_k\to v & \text{pointwise on } (0\,,1];\\ v_k'\rightharpoonup v' & \text{weakly in } L^1(\varepsilon\,,1) \text{ for every } 0<\varepsilon<1.\end{cases}\tag{3.19}$$

Claim 2. Let v_k, $v\in\mathcal{A}(\lambda)$ be such that $v_k\rightharpoonup v$ as in (3.19). Then,

$$J^{**}(v)\le\liminf_{k\to+\infty}J^{**}(v_k).\tag{3.20}$$

Proof of Claim 1. Set

$$J_0^{**}(z)=\omega_N\int_0^1 r^{N-1}\Phi^{**}\left(z'(r)\,,\frac{z(r)}{r}\right)dr,\qquad z\in\mathcal{A}(\lambda).\tag{3.21}$$

As $\{v_k\}_k$ is a minimizing sequence for J^{**} and $\theta\ge0$, we have

$$J_0^{**}(v_k)\le J^{**}(v_k)\le C\tag{3.22}$$

for every k and for some $C>0$. Thus, the growth assumption (b) of Proposition 3.2 implies that the sequence is sequentially weakly compact in $AC([\varepsilon\,,1])$ for every $0<\varepsilon<1$, and the usual diagonal argument yields a subsequence $\{v_h\}_h$ converging to a function $v\in AC_{\mathrm{loc}}((0\,,1])$ in the sense of (3.19). The pointwise convergence implies that $v(1)=\lambda$ and that v is increasing on $[0\,,1]$ because every v_h enjoys the same properties. Hence, $v'\ge0$ a.e. on $(0\,,1)$, and we set $v(0)=\lim_{r\to0_+}v(r)$. To complete the proof of the claim, we only have to check that $v\in\mathcal{A}(\lambda)$, i.e., that (2.14) holds.

To see this, note that the first assertion of ($\mathrm{H5_L}$) implies that Φ^{**} can be extended to a lower semicontinuous function defined on $\mathbb{R}^2$ by setting $\Phi^{**}(\xi\,,\eta)=+\infty$ whenever either $\xi\le 0$ or $\eta\le 0$. Therefore, for every $0<\varepsilon<1$, the integral

$$z\in AC_{\mathrm{loc}}((0\,,1])\mapsto \omega_N\int_\varepsilon^1 r^{N-1}\Phi^{**}\left(z'(r)\,,\frac{z(r)}{r}\right)\,dr$$

is sequentially lower semicontinuous along sequences converging as in (3.19) by classical results (see [10] or [11]) whence

$$\int_\varepsilon^1 r^{N-1}\Phi^{**}\left(v'(r)\,,\frac{v(r)}{r}\right)\,dr\le\liminf_h\int_\varepsilon^1 r^{N-1}\Phi^{**}\left(v_h'(r)\,,\frac{v_h(r)}{r}\right)\,dr$$

follows for every $0<\varepsilon<1$. As $\phi^{**}\ge 0$, the right-hand side of the estimate above is obviously bounded by $\liminf_h J_0^{**}(v_h)$ and hence, letting $\varepsilon\to 0_+$ on the left-hand side, we conclude that

$$J_0^{**}(v)\le\liminf_h J_0^{**}(v_h)\le C, \tag{3.23}$$

where C is as in (3.22). Thus, $v'>0$ a.e. on $(0\,,1)$, and this gives also that $v>0$ on $(0\,,1]$. Thus, $v\in\mathcal{A}(\lambda)$ and the claim is proved.

Proof of Claim 2. First, set

$$J_V^{**}(z)=J_0^{**}(z)+w^\infty\frac{\omega_N}{N}\left[z(0)\right]^N\,,\qquad z\in\mathcal{A}(\lambda),$$

and recall that Theorem 1 in [19] shows that

$$J_V^{**}(v)=\inf\left\{\liminf_k J_0^{**}(v_k):\ v_k\in\mathcal{A}(\lambda),\,v_k(0)=0\text{ and }v_k\rightharpoonup v\right\} \tag{3.24}$$

for every $v\in\mathcal{A}(\lambda)$ and that

$$J_V^{**}(v)\le\liminf_k J_V^{**}(v_k); \tag{3.25}$$

i.e., J_V^{**} is lower semicontinuous along sequences $v_k\rightharpoonup v$ as in (3.19). Moreover,

$$J_0^{**}(v)\le J^{**}(v)\le J_V^{**}(v),\qquad v\in\mathcal{A}(\lambda), \tag{3.26}$$

because of (a) and (b). Then, without loss of generality, we can assume that both sequences $\{J_0^{**}(v_k)\}_k$ and $\{v_k(0)\}_k$ are convergent so that

$$\liminf_k J^{**}(v_k)=\lim_k J_0^{**}(v_k)+\omega_N w^\infty\lim_k\theta(v_k(0)).$$

If there is a subsequence $\{v_h\}_h$ such that $v_h(0)=0$ for every h, then $J^{**}(v_h)=J_0^{**}(v_h)$ and (3.24) and (3.26) yield that

$$\lim_k J^{**}(v_k)=\lim_h J_0^{**}(v_h)\ge J_V^{**}(v)\ge J^{**}(v).$$

Otherwise, we can assume that $v_k(0)>0$ for every k and either $v(0)\le\lim_k v_k(0)$ or $v(0)>\lim_k v_k(0)$. In the former case, the conclusion follows from (3.23) and the monotonicity of θ. Otherwise, we have

$$J^{**}(v_k)=J_V^{**}(v_k)-w^\infty\omega_N\Delta(v_k(0))$$

for every k where Δ is defined in (b), and then (3.25) yields

$$\begin{aligned}\liminf_k J^{**}(v_k) &= \liminf_k J_V^{**}(v_k) - w^\infty \omega_N \lim_k \Delta(v_k(0)) \\ &\geq J_V^{**}(v) - w^\infty \omega_N \lim_k \Delta(v_k(0)) \\ &= J^{**}(v) + w^\infty \omega_N \left[\Delta(v(0)) - \lim_k \Delta(v_k(0))\right].\end{aligned}$$

As $v(0) > \lim_k v_k(0)$ and Δ is increasing by assumption, the conclusion follows. ◻

We conclude this section with the existence result for the nonconvex variational problem $(\mathcal{P}_0)$ corresponding to the case of functions w having superlinear growth at infinity. As already explained, this is the situation of nearly incompressible elastomers which can possibly exhibit phase transition. In this superlinear case, the existence result of Theorem 3.3 turns into the following theorem.

THEOREM 3.5. *Let the hypotheses of Theorem* 3.3 *hold with* ($\mathrm{H5_L}$) *replaced by* ($\mathrm{H5_S}$). *Then, the minimum problem* $(\mathcal{P}_0)$ *admits a solution.*

4. Attainment for the nonconvex, one-dimensional problems. In this section, we prove Theorems 3.3 and 3.5. Our starting point is the following result which is proved as Lemma 3.2 in [5]. It describes a procedure to define local, piecewise linear approximations of absolutely continuous functions subject to the constraint of using only two given values of the derivative.

LEMMA 4.1. *Let* $z \in AC([0,1])$ *be differentiable at some point* $0 < r_0 < 1$ *with* $m = z(r_0)$ *and* $\xi_0 = z'(r_0)$ *and let* $\alpha,\ \beta \in \mathbb{R}$ *be such that*

$$\alpha < \xi_0 < \beta.$$

Then, for every $\delta > 0$, *there exist two families of compact subintervals* $\{H_\varepsilon^\pm\}_\varepsilon$ *of* $(0,1)$ *and two families of functions* $\{z_\varepsilon^\pm\}_\varepsilon$ *in* $AC([0,1])$ *such that, setting*

$$\begin{cases} I_\varepsilon^+ = \left(r_0 - \dfrac{\varepsilon}{\beta - \xi_0}, r_0 + \dfrac{\varepsilon}{\xi_0 - \alpha}\right), \\ I_\varepsilon^- = \left(r_0 - \dfrac{\varepsilon}{\xi_0 - \alpha}, r_0 + \dfrac{\varepsilon}{\beta - \xi_0}\right), \end{cases}$$

the following properties hold for every $\varepsilon > 0$ *small enough:*

(4.1) $\quad I_{\varepsilon/2}^\pm \subset H_\varepsilon^\pm \subset I_{2\varepsilon}^\pm \subset (0,1)$;

(4.2) $\quad z_\varepsilon^\pm = z$ *on* $[0,1] \setminus \operatorname{int}\left(H_\varepsilon^\pm\right)$;

(4.3+) $\quad z(r) < z_\varepsilon^+(r) \leq z(r) + \delta$ *for every* $r \in \operatorname{int}\left(H_\varepsilon^+\right)$;

(4.3−) $\quad z(r) - \delta \leq z_\varepsilon^-(r) < z(r)$ *for every* $r \in \operatorname{int}\left(H_\varepsilon^-\right)$;

(4.4+) $\quad \varepsilon \geq z_\varepsilon^+(r) - [m + \xi_0(r - r_0)] \geq \varepsilon/2$ *for every* $r \in I_{\varepsilon/2}^+$;

(4.4−) $\quad -\varepsilon/2 \geq z_\varepsilon^-(r) - [m + \xi_0(r - r_0)] \geq -\varepsilon$ *for every* $r \in I_{\varepsilon/2}^-$;

(4.5) $\quad \left(z_\varepsilon^\pm(r)\right)' \in \{\alpha, \beta\}$ *for a.e.* $r \in H_\varepsilon^\pm$.

We can then exploit the construction of the previous lemma to find comparison functions that decrease the value of the integral

$$\int_I r^{N-1} f\left(\frac{v(r)}{r}\right) dr$$

when $v \in \mathcal{A}(\lambda)$ and the compact interval $I \subset (0,1)$ have the property that $v(r)/r$ stays on a strict, local maximum point of f on a subset of I of positive measure.

PROPOSITION 4.2. *Let $0 < \delta < m$ and let $f\colon [m-\delta, m+\delta] \to \mathbb{R}$ be a continuous function such that*

(a) *f has a strict, local maximum point at m;*

(b) *f is increasing on the interval $[m-\delta, m]$ and decreasing on $[m, m+\delta]$.*

Assume also that there exist a compact interval $I \subset (0,1)$ and a function $v \in \mathcal{A}(\lambda)$ such that

(c) *$|v(r)/r - m| \le \delta/2$ for every $r \in I$;*

(d) *$|I \cap \{v(r)/r = m\}| > 0$.*

Then, $v' = m$ a.e. on $I \cap \{v(r)/r = m\}$, and there exists $u \in \mathcal{A}(\lambda)$ such that

$$\{u \ne v\} \subset \operatorname{int}(I); \tag{4.6}$$

$$|u(r)/r - m| \le \delta \text{ for every } r \in I; \tag{4.7}$$

$$|u'(r) - m| \le \delta \text{ for a.e. } r \in \{u \ne v\}; \tag{4.8}$$

$$\int_I r^{N-1} f\left(\frac{u(r)}{r}\right) dr < \int_I r^{N-1} f\left(\frac{v(r)}{r}\right) dr. \tag{4.9}$$

Proof. The proof is based on the same idea of Step 3 of Theorem 2.2 in [5]. Set $z(r) = v(r)/r$ for $0 < r \le 1$ and $E = \{z = m\}$. Then, $z \in AC_{\mathrm{loc}}((0,1])$ and $z' = 0$ a.e. on $I \cap E$ for it is a level set of positive measure; i.e., $v' = m$ a.e. on $I \cap E$. We choose a density point $r_0 \in \operatorname{int}(I) \cap E$ where v and z are both differentiable and $z'(r_0) = 0$ so that $v'(r_0) = v(r_0)/r_0 = m$. Then, we apply Lemma 4.1 to the function z choosing $\alpha < 0 < \beta$ such that $\max\{-\alpha, \beta\} < \delta/4$ and with δ replaced by $\delta/4$ in (4.3$-$) and (4.3+). We thus find compact intervals $\{H_\varepsilon^\pm\}_\varepsilon$ of $(0,1)$ and functions $\{z_\varepsilon^\pm\}_\varepsilon$ in $AC_{\mathrm{loc}}((0,1])$ such that (4.1), …, (4.5) hold. Moreover, choosing $0 < \gamma < \min\{-\alpha, \beta\}$, we can assume that

$$H_\varepsilon^\pm \subset I_{2\varepsilon}^\pm \subset [r_0 - 2\varepsilon/\gamma, r_0 - 2\varepsilon/\gamma] \subset \operatorname{int}(I) \tag{4.10}$$

holds for every ε small enough because of (4.1), $z'(r_0) = 0$, and the choice of α and β. Then, set

$$v_\varepsilon^\pm(r) = r z_\varepsilon^\pm(r), \qquad 0 < r \le 1.$$

We shall prove that either $u = v_\varepsilon^+$ or $u = v_\varepsilon^-$ has the required properties for ε small enough.

First, it is clear that $v_\varepsilon^\pm \in AC([0,1])$ and that $\{v_\varepsilon^\pm \ne v\} \subset \operatorname{int}(H_\varepsilon^\pm) \subset \operatorname{int}(I)$ because of (4.2). Moreover, (c) and either (4.3+) or (4.3$-$) with $\delta/4$ yield

$$\left|v_\varepsilon^\pm(r)/r - m\right| \le \left|v_\varepsilon^\pm(r)/r - v(r)/r\right| + |v(r)/r - m| \le \delta/4 + \delta/2 = 3\delta/4 < \delta \tag{4.11}$$

which will be (4.7), and we claim that

$$\left|\left(v_\varepsilon^\pm\right)'(r) - m\right| \le \delta \qquad \text{for a.e. } r \in H_\varepsilon^\pm$$

which will be (4.8). In particular, this implies that $\left(v_\varepsilon^\pm\right)' > 0$ a.e. on $H_\varepsilon^\pm$ and hence that $v_\varepsilon^\pm \in \mathcal{A}(\lambda)$. Indeed, we have $\left(z_\varepsilon^\pm\right)' \in \{\alpha, \beta\}$ a.e. on $H_\varepsilon^\pm$ by (4.5), and hence

$$\left|\left(v_\varepsilon^\pm\right)' - m\right| \le \left|\left(z_\varepsilon^\pm\right)'\right| + \left|v_\varepsilon^\pm(r)/r - m\right| \le \max\{-\alpha, \beta\} + 3\delta/4 < \delta$$

for a.e. $r \in H_\varepsilon^\pm$ by (4.11) and the choice of α and β. At last, we prove that either v_ε^+ or v_ε^- is such that

$$\int_I r^{N-1} f\left(\frac{v_\varepsilon^\pm(r)}{r}\right) dr < \int_I r^{N-1} f\left(\frac{v(r)}{r}\right) dr$$

eventually as $\varepsilon \to 0_+$ so that also (4.9) holds for u chosen accordingly.

To see this, note first that it is enough to compare the integrals over the intervals $H_\varepsilon^\pm$ only because of (4.2). Then, we choose a sequence $\varepsilon_j \to 0_+$ and we set

$$\eta_j = \frac{1}{\varepsilon_j} \sup\left\{|v(r)/r - m| : |r - r_0| \le 2\varepsilon_j/\gamma\right\}.$$

We note that $\eta_j\varepsilon_j \le \delta/2$ by (c) and that $\eta_j \to 0_+$ because $z = v/r$ is differentiable at r_0 with $z(r_0) = m$ and $z'(r_0) = 0$. Then, possibly passing to a subsequence still denoted by ε_j, we can assume that the minimum between $f(m-\eta_j\varepsilon_j)$ and $f(m+\eta_j\varepsilon_j)$ is actually achieved for every j by terms having always the same sign inside, say $f(m + \eta_j\varepsilon_j)$, so that

$$0 < f(m) - f(m + \eta_j\varepsilon_j) = \max\left\{f(m) - f(m \pm \eta_j\varepsilon_j)\right\} \tag{4.12}$$

holds for every j. According to this assumption, we choose the + functions and, to simplify the notation, we set $v_j = v_{\varepsilon_j}^+$ and $H_j = H_{\varepsilon_j}^+$. Finally, set $I_j^1 = I_{\varepsilon_j/2}^+$ and $I_j^2 = I_{2\varepsilon_j}^+$ so that (4.1) turns into

$$I_j^1 \subset H_j \subset I_j^2 \tag{4.13}$$

and set also

$$A_j^1 = \frac{1}{|H_j|}\int_{H_j} r^{N-1}\left[f(m) - f\left(\frac{v_j(r)}{r}\right)\right] dr,$$
$$A_j^2 = \frac{1}{|H_j|}\int_{H_j} r^{N-1}\left[f(m) - f\left(\frac{v(r)}{r}\right)\right] dr.$$

We shall prove that $A_j^1 - A_j^2 > 0$ eventually. In fact, note first that (4.4+) reduces to $\varepsilon_j/2 \le v_j(r)/r - m \le \varepsilon_j$ for every $r \in I_j^1$. Hence, (4.13) and (a) and (b) of Proposition 4.2 yield that

$$\begin{aligned} A_j^1 &\ge \frac{1}{|I_j^2|}\int_{I_j^1} r^{N-1}\left[f(m) - f\left(\frac{v_j(r)}{r}\right)\right] dr \\ &\ge \frac{1}{|I_j^2|}\int_{I_j^1} r^{N-1}\left[f(m) - f\left(m + \frac{\varepsilon_j}{2}\right)\right] dr \ge \frac{c^{N-1}}{4}\left[f(m) - f\left(m + \frac{\varepsilon_j}{2}\right)\right] \end{aligned}$$

because $|I_j^2|/|I_j^1| = 4$ with $c > 0$ such that $c < \inf I_j^1$ for every j. As for A_j^2, we have

$$A_j^2 = \frac{1}{|H_j|}\int_{H_j\setminus E} r^{N-1}\left[f(m) - f\left(\frac{v(r)}{r}\right)\right] dr$$

and $|v(r)/r - m| \le \eta_j\varepsilon_j$ for every $r \in H_j \subset I_j^2$ by (4.13), the definition of η_j, and (4.10). Hence,

$$0 \le f(m) - f(v(r)/r) \le \max\left\{f(m) - f(m \pm \eta_j\varepsilon_j)\right\} = f(m) - f(m + \eta_j\varepsilon_j)$$

for every $r \in H_j$ because of (a) and (b) whence

$$A_j^2 \leq 4 \frac{|I_j^2 \setminus E|}{|I_j^2|} \left[f(m) - f(m + \eta_j \varepsilon_j) \right].$$

Since $\eta_j \to 0_+$ and f is decreasing on the interval $[m\,, m+\delta]$, we conclude that $f(m) - f(m + \varepsilon_j/2) \geq f(m) - f(m + \eta_j \varepsilon_j) > 0$ eventually. Finally, as r_0 is a density point of E by assumption, the ratio $|I_j^2 \setminus E|/|I_j^2|$ goes to zero and the conclusion follows. $\square$

The next step is the construction of a further family of comparison functions which are obtained as solutions to partial differential relations. This construction is an instance of the convex integration of partial differential relations developed in this variational framework by Müller and Šverák in [22]. Again, apart from minor and obvious changes, this construction is similar to the one described in Proposition 3.3 in [5], and we refer to this paper for the proof.

Proposition 4.3. *Let $d^{\pm} \colon [\eta_0 - \delta\,, \eta_0 + \delta] \to \mathbb{R}$ be two bounded, upper and lower semicontinuous functions, respectively, and let $u \in \mathcal{A}(\lambda)$ be such that*

(a) *u is differentiable at $0 < r_0 < 1$;*

(b) *$d^-(\eta) < u'(r_0) < d^+(\eta)$ for every $\eta \in [\eta_0 - \delta\,, \eta_0 + \delta]$, where $\eta_0 = u(r_0)/r_0$.*

Then, there exist $\varepsilon_0 = \varepsilon_0(r_0\,, \delta) > 0$, two families of compact subintervals $\mathcal{K}^{\pm} = \{K_\varepsilon^{\pm}\}_\varepsilon$ of $(0\,,1)$ such that

$$\text{(4.14)} \qquad \textit{each set } K_\varepsilon^{\pm} \textit{ is a neighborhood of } r_0 \textit{ and } |K_\varepsilon^{\pm}| \to 0 \textit{ as } \varepsilon \to 0^+,$$

and two families of functions $\mathcal{U}^{\pm} = \{u_\varepsilon^{\pm}\}_\varepsilon$ in $\mathcal{A}(\lambda)$ such that the following properties hold for every $0 < \varepsilon < \varepsilon_0$:

(4.15) $\quad u_\varepsilon^{\pm} = u$ *on* $(0\,,1] \setminus \operatorname{int}\left(K_\varepsilon^{\pm}\right)$;

(4.16^+) $\quad u(r)/r < u_\varepsilon^+(r)/r \leq u(r)/r + \varepsilon$ *for every* $r \in \operatorname{int}\left(K_\varepsilon^+\right)$;

(4.16^-) $\quad u(r)/r - \varepsilon \leq u_\varepsilon^-(r)/r < u(r)/r$ *for every* $r \in \operatorname{int}\left(K_\varepsilon^-\right)$;

(4.17) $\quad |u_\varepsilon^{\pm}(r)/r - \eta_0| \leq \delta$ *for every* $r \in K_\varepsilon^{\pm}$;

$$\text{(4.18)} \qquad \left(u_\varepsilon^{\pm}\right)'(r) \in \left\{ d^- \left(\frac{u_\varepsilon^{\pm}(r)}{r} \right),\ d^+ \left(\frac{u_\varepsilon^{\pm}(r)}{r} \right) \right\} \textit{ for a.e. } r \in K_\varepsilon^{\pm}.$$

After these technical preliminaries, we can turn to the proof of Theorem 3.3.

Proof of Theorem 3.3. Let $v \in \mathcal{A}(\lambda)$ be a solution to the relaxed Bolza problem $(\mathcal{P}^{**})$ which exists by assumption and set

$$\text{(4.19)} \qquad E(v) = \left\{ r \in (0\,,1) : \ v \text{ is differentiable at } r \text{ and } \left(v'(r)\,, \frac{v(r)}{r} \right) \in \mathcal{D} \right\},$$

where $\mathcal{D}$ is the detachment set defined by (3.9).

We are going to prove the theorem by showing that, among all solutions $v \in \mathcal{A}(\lambda)$ to $(\mathcal{P}^{**})$, there is one such that $|E(v)| = 0$ so that

$$\Phi^{**} \left(v'(r)\,, \frac{v(r)}{r} \right) = \Phi \left(v'(r)\,, \frac{v(r)}{r} \right) \qquad \text{for a.e. } r \in (0\,,1),$$

which implies immediately that v is a solution to $(\mathcal{P})$ as well.

To this aim, note first that by Proposition 3.2, the function $\xi \in \mathcal{D}^\eta \mapsto \psi(\xi\,,\eta)$ defined by (3.15) is constant on each connected component of $\mathcal{D}^\eta$ and hence (3.16) implies the existence of a countable family of subsets $\{m_i\}\times L_i$ such that every open interval L_i is a connected component of $\mathcal{D}^{m_i}$ and, for every $\xi \in L_i$, m_i is a strict, local extremum point of the mapping $\eta \mapsto \psi(\xi\,,\eta)$. Moreover, setting $\mathcal{M} = \cup_i (\{m_i\}\times L_i)$, the set $\mathcal{D}\setminus\mathcal{M}$ is open, and we emphasize that ψ is locally constant with respect to ξ and locally monotone with respect to η on the set $\mathcal{D}\setminus\mathcal{M}$.

Then, when v is a solution to $(\mathcal{P}^{**})$, we write

$$E(v) = E_0(v) \cup \left[\bigcup_i E_i(v)\right],$$

where

$$\begin{aligned} E_0(v) &= \left\{ r \in E(v):\ \left(v'(r)\,, \frac{v(r)}{r}\right) \in \mathcal{D}\setminus\mathcal{M} \right\}, \\ E_i(v) &= \left\{ r \in E(v):\ \frac{v(r)}{r} = m_i \right\}. \end{aligned} \tag{4.20}$$

We shall prove the theorem by proving the following two claims.

Claim 1. There exists a solution $v \in \mathcal{A}(\lambda)$ to $(\mathcal{P}^{**})$ such that $|E_0(v)| = 0$.

Claim 2. For the solution v of Claim 1, we have $|E_i(v)| = 0$ for every i.

Proof of Claim 1. Let $u \in \mathcal{A}(\lambda)$ be a solution to $(\mathcal{P}^{**})$ and assume that the corresponding set $E_0(u)$ defined by (4.20) has positive measure. Choose $r_0 \in E_0(u)$ and set $\eta_0 = u(r_0)/r_0$ and $\xi_0 = u'(r_0)$ so that $(\xi_0\,,\eta_0) \in \mathcal{D}\setminus\mathcal{M}$. Then, we find $\delta > 0$ and two functions $d^\pm$ satisfying (c), (d), (e), and (f) of Proposition 3.1 and we set

$$\mathcal{D}' = \left\{(\xi\,,\eta):\ |\eta-\eta_0| < \delta \text{ and } d^-(\eta) < \xi < d^+(\eta)\right\}$$

so that, recalling (d), (e), and (f) of Proposition 3.2, we can write the convex envelope Φ^{**} (with respect to ξ) of Φ on the set $\mathcal{D}'$ as in (3.14); i.e.,

$$\Phi^{**}(\xi\,,\eta) = m(\eta)\xi + q(\eta), \qquad (\xi\,,\eta) \in \mathcal{D}', \tag{4.21}$$

where the continuous functions $m,\ q\colon [\eta_0-\delta\,,\eta_0+\delta] \to \mathbb{R}$ are defined by $m(\eta) = m(\xi\,,\eta)$ and $q(\eta) = q(\xi\,,\eta)$ for every $(\xi\,,\eta) \in \mathcal{D}'$ as in (3.13). Similarly, relying again on (d), (e), (f), and (g) of the same proposition, we can write

$$\psi(\eta) = \psi(\xi\,,\eta), \qquad (\xi\,,\eta) \in \mathcal{D}', \tag{4.22}$$

and, recalling that $(\xi_0\,,\eta_0) \in \mathcal{D}\setminus\mathcal{M}$ and possibly choosing a smaller value of δ, we can assume also that (3.16) holds for δ; i.e., ψ is monotone on the interval $[\eta_0-\delta\,,\eta_0+\delta]$. Then, going back to the properties of Φ^{**}, we note also that

$$\Phi^{**}(\xi\,,\eta) \geq m(\eta)\xi + q(\eta), \qquad \eta \in [\eta_0-\delta\,,\eta_0+\delta], \quad \xi \in \mathbb{R}, \tag{4.23}$$

holds because of (3.6) and that the equalities

$$\Phi^{**}(d^\pm(\eta)\,,\eta) = \Phi(d^\pm(\eta)\,,\eta), \qquad \eta \in [\eta_0-\delta\,,\eta_0+\delta], \tag{4.24}$$

follow from (f) of Proposition 3.1.

Then, we apply Proposition 4.3 and we let $\mathcal{K}^{\pm}_{r_0} = \{K^{\pm}_{r_0,\varepsilon}\}_\varepsilon$ and $\mathcal{V}^{\pm}_{r_0} = \{u^{\pm}_{r_0,\varepsilon}\}_\varepsilon$ be the corresponding intervals and functions. We assume in addition that $\varepsilon = \varepsilon_0(r_0)$ is small enough so as to have

$$\begin{cases} |u(r)/r - \eta_0| \le \delta, \\ |u^{\pm}_{r_0,\varepsilon}(r)/r - \eta_0| \le \delta, \end{cases} \qquad r \in K^{\pm}_{r_0,\varepsilon}, \quad 0 < \varepsilon \le \varepsilon_0. \tag{4.25}$$

Now, we wish to compare $J^{**}(u^{\pm}_{r_0,\varepsilon})$ and $J^{**}(u)$. Since $u(0) = u^{\pm}_{r_0,\varepsilon}(0)$, we have

$$\begin{aligned} &J^{**}(u^{\pm}_{r_0,\varepsilon}) - J^{**}(u) \\ &\qquad = \omega_N \int_{K^{\pm}_{r_0,\varepsilon}} r^{N-1} \left[\Phi^{**}\left((u^{\pm}_{r_0,\varepsilon})'(r)\, , \frac{u^{\pm}_{r_0,\varepsilon}(r)}{r} \right) - \Phi^{**}\left(u'(r)\, , \frac{u(r)}{r} \right) \right] dr, \end{aligned}$$

and, because of (4.21) and (4.23), the right-hand side is less than or equal to

$$\begin{aligned} \omega_N \int_{K^{\pm}_{r_0,\varepsilon}} r^{N-1} \Bigg\{ &\left[m\left(\frac{u^{\pm}_{r_0,\varepsilon}(r)}{r} \right) (u^{\pm}_{r_0,\varepsilon})'(r) - m\left(\frac{u(r)}{r} \right) u'(r) \right] \\ &\qquad + \left[q\left(\frac{u^{\pm}_{r_0,\varepsilon}(r)}{r} \right) - q\left(\frac{u(r)}{r} \right) \right] \Bigg\}\, dr. \end{aligned}$$

To evaluate this, let $z \in AC([\alpha\, , \beta])$ be any function such that $|z(r)/r - \eta_0| \le \delta$ for $0 < \alpha \le r \le \beta$. Then, recalling the definition of M in (g) of Proposition 3.2 and (H6) and integrating by parts, we obtain that

$$\begin{aligned} \int_\alpha^\beta r^{N-1} m\left(\frac{z(r)}{r} \right) z'(r)\, dr &= \int_\alpha^\beta r^N \left\{ \left[M\left(\frac{z(r)}{r} \right) \right]' + m\left(\frac{z(r)}{r} \right) \frac{z(r)}{r^2} \right\} dr \\ = r^N M\left(\frac{z(r)}{r} \right) \Bigg|_\alpha^\beta &+ \int_\alpha^\beta r^{N-1} \left[-NM\left(\frac{z(r)}{r} \right) + m\left(\frac{z(r)}{r} \right) \frac{z(r)}{r} \right] dr. \end{aligned}$$

We now exploit this equality with both $z = u^{\pm}_{r_0,\varepsilon}$ and $z = u$. Since (4.25) holds and $u^{\pm}_{r_0,\varepsilon}$ and u coincide at the endpoints of the interval $K^{\pm}_{r_0,\varepsilon}$, we finally get

$$J^{**}(u^{\pm}_{r_0,\varepsilon}) - J^{**}(u) \le \omega_N \int_{K^{\pm}_{r_0,\varepsilon}} r^{N-1} \left[\psi\left(\frac{u^{\pm}_{r_0,\varepsilon}(r)}{r} \right) - \psi\left(\frac{u(r)}{r} \right) \right] dr. \tag{4.26}$$

Thus, recalling again (4.25), the fact that ψ is monotone on the interval $[\eta_0 - \delta\, , \eta_0 + \delta]$ by (3.16), and the assumption that $(\xi_0\, , \eta_0) \in \mathcal{D} \setminus \mathcal{M}$, we conclude that, choosing $\sigma = \sigma(r_0)$ to be either $+$ or $-$ according to the monotonicity of ψ on the 2δ long interval centered at η_0, we can make the right-hand side of the previous estimate ≤ 0, thus showing that all modified functions $u^{\sigma}_{r_0,\varepsilon}$ are solutions to $(\mathcal{P}^{**})$ satisfying

$$\Phi^{**}\left(\left(u^{\sigma}_{r_0,\varepsilon}\right)'(r)\, , \frac{u^{\sigma}_{r_0,\varepsilon}(r)}{r} \right) = \Phi\left(\left(u^{\sigma}_{r_0,\varepsilon}\right)'(r)\, , \frac{u^{\sigma}_{r_0,\varepsilon}(r)}{r} \right) \qquad \text{for a.e. } r \in K^{\pm}_{r_0,\varepsilon}$$

because of (4.18) and (4.24).

At last, a new solution $v \in \mathcal{A}(\lambda)$ to $(\mathcal{P}^{**})$ satisfying $|E_0(v)| = 0$ can be easily defined by exploiting a standard covering argument. In fact, Vitali's covering theorem applies because of (4.14) and yields countably many points $r_j \in E_0(u)$, positive numbers ε_j, and symbols $\sigma_j \in \{+\,,-\}$ such that the corresponding sets K_j are pairwise disjoint and cover $E_0(u)$ up to a null set, i.e.,

$$|E_0(u) \setminus (\cup_j K_j)\,| = 0,$$

while the corresponding functions u_j are solutions to $(\mathcal{P}^{**})$ satisfying

$$\Phi^{**}\left(u_j'(r)\,,\,\frac{u_j(r)}{r}\right) = \Phi\left(u_j'(r)\,,\,\frac{u_j(r)}{r}\right) \qquad \text{for a.e. } r \in K_j.$$

It is then easy to check that the series $v = u + \sum_j (u_j - u)$ (there is only one nonvanishing summand at every point r) defines a function in $\mathcal{A}(\lambda)$ as required.

Proof of Claim 2. Let v be the solution to $(\mathcal{P}^{**})$ of Claim 1, assume that $|E_i(v)| > 0$ for some i and, to simplify the notation, write $m = m_i$ and $E = E_i(v)$. As E is a level set of $v(r)/r$ of positive measure, $v'(r) = v(r)/r = m$ for a.e. $r \in E$, and hence we can choose $\delta > 0$ small enough so that the square $Q = [m-\delta\,, m+\delta]\times[m-\delta\,, m+\delta]$ is contained in $\mathcal{D}$ and then, arguing again as in Claim 1, we can assume that the functions $m(\eta) = m(\xi\,,\eta)$, $q(\eta) = q(\xi\,,\eta)$, and $\psi(\eta) = \psi(\xi\,,\eta)$ are well defined and continuous for every $(\xi\,,\eta) \in Q$ and that (4.21) and (4.23) hold for every $(\eta\,,\xi) \in Q$ and $(\eta\,,\xi) \in [m-\delta\,, m+\delta]\times\mathbb{R}$, respectively. Moreover, the very definition of E implies that m is a strict, local maximum point of ψ by (3.17), and, finally, we can assume that δ is small enough to have ψ increasing on the interval $[m-\delta\,, m]$ and decreasing on the interval $[m\,, m+\delta]$ by (3.16). Then, we choose a compact interval $I \subset (0\,,1)$ such that (c) and (d) of Proposition 4.2 hold and we exploit this result with $f = \psi$. We thus find a function $u \in \mathcal{A}(\lambda)$ satisfying (4.6), …, (4.9). Relying on (4.6), (4.7), and (4.8) which says that the derivative of the modified function u' remains in the interval where $\Phi^{**}(\xi\,,\eta)$ is affine with respect to ξ, we can repeat the argument of Claim 1 leading to (4.26) to find that

$$J^{**}(u) - J^{**}(v) \leq \omega_N \int_I r^{N-1}\left[\psi\left(\frac{u(r)}{r}\right) - \psi\left(\frac{v(r)}{r}\right)\right] dr.$$

This gives a contradiction to the minimality of v as the right-hand side is negative by (4.9). Thus, $|E| = 0$, and this completes the proof. ∎

Finally, the very same proof of Theorem 3.3 provides attainment in the superlinear case as well.

Proof of Theorem 3.5. The convexified integral J_0^{**} defined by (3.21) has a minimizer over $\mathcal{A}(\lambda)$, say u, because of Claim 1 of Theorem 3.4 and (3.23). Then, u can be modified to find a solution v to $(\mathcal{P}_0)$ by exploiting the very same arguments of Theorem 3.3. In fact, the proofs of Claims 1 and 2 in that theorem are of local nature so that the behavior of w at infinity is irrelevant and, moreover, the proofs involve only the absolutely continuous part of J^{**} because the modified functions out of which v is constructed in Claim 1 and the comparison functions with which v is compared in Claim 2 share the same value at $r = 0$ with the original u so that comparing J_0^{**} is the same as comparing J^{**}. ∎

5. The special case of Blatz–Ko materials. In this section we prove the existence of optimal radial solutions for Blatz–Ko materials (Theorem 2.1) by showing that the related Lagrangian satisfies all the hypotheses of Theorem 3.3.

Proof of Theorem 2.1. We want to check that the function

$$\Phi(\xi\,,\eta) = \xi^2 + 2\eta^2 + \frac{2\alpha}{\xi\eta^2} + (1-\alpha)\left(2\xi\eta^2 + \frac{1}{\xi^2\eta^4}\right), \qquad \xi,\,\eta > 0,$$

satisfies all the hypotheses of Theorem 3.3. Indeed, the hypotheses (H1), ..., ($\mathrm{H5_L}$), (H6), and (3.16) obviously hold and we shall prove that $\xi < \eta$ for every $(\xi\,,\eta) \in \mathcal{D}$ so that (3.17) is actually empty. In order to draw the graph of the function Φ, we set $\theta = \xi\eta^2$ and consider the new function Ψ defined by

$$\Psi(\theta\,,\eta) = \Phi\left(\frac{\theta}{\eta^2}\,,\eta\right) = \frac{\theta^2}{\eta^4} + 2\eta^2 + \frac{2\alpha}{\theta} + (1-\alpha)\left(2\theta + \frac{1}{\theta^2}\right), \qquad \theta,\,\eta > 0.$$

We want to estimate the set $I_\eta = \{\theta :\ \Psi^{**}(\theta\,,\eta) < \Psi(\theta\,,\eta)\}$ for every η and prove that $\theta < \eta^3$ for every $\theta \in I_\eta$.

To this aim, consider first the second derivative

$$\Psi''_{\theta\theta}(\theta\,,\eta) = \frac{2}{\eta^4} + \frac{4\alpha}{\theta^3} + (1-\alpha)\frac{6}{\theta^4}, \qquad \theta,\,\eta > 0.$$

Then, $\Psi''_{\theta\theta}(\theta\,,\eta) \to +\infty$ as $\theta \to 0_+$ and $\Psi''_{\theta\theta}(\theta\,,\eta) \to 2/\eta^4$ as $\theta \to +\infty$, and, moreover, setting

$$\theta_0 = \frac{\alpha - 1}{\alpha} > 0,$$

we find that the minimum of $\theta \mapsto \Psi''_{\theta\theta}(\theta\,,\eta)$ is achieved at $2\theta_0$ and is given by $2/\eta^4 + (\alpha/8)\theta_0^3$ for every $\eta > 0$. Thus, for $\eta > \eta^* = 2\sqrt[4]{\theta_0^3/(-\alpha)}$, the set I_η is nonempty and the set where $\theta \mapsto \Psi''_{\theta\theta}(\theta\,,\eta)$ is ≤ 0 is an interval, say $[\theta_1(\eta)\,,\theta_2(\eta)]$. Since, for every $\eta > \eta^*$, $\Psi''_{\theta\theta}(2\theta_0\,,\eta) < 0$ and $\Psi''_{\theta\theta}(3\theta_0/2\,,\eta) = 2/\eta^4 > 0$, we conclude that

$$\frac{3}{2}\theta_0 < \theta_1(\eta) < 2\theta_0 < \theta_2(\eta). \tag{5.1}$$

Therefore, the first derivative $\Psi'_\theta(\cdot\,,\eta)$ has a local maximum at $\theta = \theta_1(\eta)$ and a local minimum at $\theta = \theta_2(\eta)$, and goes to $-\infty$ as $\theta \to 0_+$ and to $+\infty$ as $\theta \to +\infty$. Moreover, it is easy to check that $\Psi'_\theta(\theta_2(\eta)\,,\eta) > 0$ for $\eta > \eta^*$, so that $\Psi'_\theta(\theta\,,\eta)$ vanishes at one point only, say $\theta = \theta_3(\eta) < \theta_1(\eta)$.

Finally, the function $\theta \mapsto \Psi(\theta, \eta)$ is decreasing on the interval $(0\,,\theta_3(\eta)]$, increasing on $[\theta_3(\eta)\,,+\infty)$, and convex on $(0\,,\theta_1(\eta)]$ and on $[\theta_2(\eta)\,,+\infty)$. Thus, the set I_η is a bounded, open interval containing $\theta_1(\eta)$ and $\theta_2(\eta)$, say $I_\eta = (\overline{\theta}_1(\eta)\,,\overline{\theta}_2(\eta))$.

In order to estimate $\overline{\theta}_2(\eta)$, we note that the tangent lines to the graph of $\theta \mapsto \Psi(\theta\,,\eta)$ at the points $\overline{\theta}_1$ and $\overline{\theta}_2$ are the same so that the following two equalities hold:

(a) $\Psi'_\theta(\overline{\theta}_1\,,\eta) = \Psi'_\theta(\overline{\theta}_2\,,\eta)$;

(b) $\Psi(\overline{\theta}_1\,,\eta) - \overline{\theta}_1\Psi'_\theta(\overline{\theta}_1\,,\eta) = \Psi(\overline{\theta}_2\,,\eta) - \overline{\theta}_2\Psi'_\theta(\overline{\theta}_2\,,\eta)$.

It follows from (b) that, for every $\eta > \eta^*$, the interval I_η is contained in the complement of the set $\{\overline{\theta} :\ q(\overline{\theta}\,,\eta) \neq q(\theta\,,\eta)$ for every $\theta > 0$ and $\theta \neq \overline{\theta}\}$, where q is the function $q(\theta\,,\eta) = \Psi(\theta\,,\eta) - \theta\Psi'_\theta(\theta\,,\eta)$. It is then easy to check that, for $\eta > \eta^*$, q has the following properties:

(5.2) $q(\theta\,,\eta) \to +\infty$ as $\theta \to 0_+$ and $q(\theta\,,\eta) \to -\infty$ as $\theta \to +\infty$;

(5.3) $\theta \mapsto q(\theta\,,\eta)$ has a local minimum at $\theta = \theta_1(\eta)$ and a local maximum at $\theta = \theta_2(\eta)$;

(5.4) $\theta \mapsto q(\theta\,,\eta)$ is decreasing on the intervals $(0\,,\theta_1(\eta)]$ and $[\theta_2(\eta)\,,+\infty)$ and increasing on the interval $[\theta_1(\eta)\,,\theta_2(\eta)]$.

For $\eta > \eta^*$, the interval I_η is a subset of $[\tilde{\theta}_2(\eta)\,,\tilde{\theta}_1(\eta)]$ where the $\tilde{\theta}_i(\eta)$'s are defined by

$$\tilde{\theta}_2 < \theta_2 \text{ and } q(\tilde{\theta}_2\,,\eta) = q(\theta_2\,,\eta),$$
$$\tilde{\theta}_1 > \theta_1 \text{ and } q(\tilde{\theta}_1\,,\eta) = q(\theta_1\,,\eta).$$

Therefore, if we show that $\tilde{\theta}_1(\eta) < \eta^3$ for every $\eta > \eta^*$, we conclude that $\theta < \eta^3$ for every $\theta \in I_\eta$.

To see this, note that (5.1), (5.3), and (5.4) yield

$$q(\tilde{\theta}_1\,,\eta) = q(\theta_1\,,\eta) = \min_{1\le s\le 2} q(s\theta_0\,,\eta) \ge -\frac{4}{\eta^4}\frac{(1-\alpha)^2}{\alpha^2} + 2\eta^2 - \frac{13}{4}\frac{\alpha^2}{1-\alpha}.$$

Moreover,

$$q\left(\eta^2\theta_0\,,\eta\right) = -\frac{(1-\alpha)^2}{\alpha^2} + 2\eta^2 + \frac{\alpha^2}{1-\alpha}\left(-\frac{4}{\eta^2} + \frac{3}{\eta^4}\right).$$

Thus, for $\eta > \eta^*$, we have

$$q(\tilde{\theta}_1\,,\eta) - q\left(\eta^2\theta_0\,,\eta\right) > -\frac{7}{2}\frac{\alpha^2}{1-\alpha} - \frac{3}{16}\frac{\alpha^6}{(1-\alpha)^4} + \frac{(1-\alpha)^2}{\alpha^2} > 0.$$

Since $\theta \mapsto q(\theta\,,\eta)$ is decreasing on $[\theta_2\,,+\infty)$ and $(\eta^*)^2\theta_0 < (\eta^*)^3$, we conclude that $\tilde{\theta}_1(\eta) < \eta^2\theta_0 < \eta^3$ for every $\eta > \eta^*$. Thus, $\theta < \eta^3$ for every $\theta \in I_\eta$ for $\eta > \eta^*$, and this concludes the proof. $\square$

REFERENCES

[1] J. M. Ball, *Discontinuous equilibrium solutions and cavitation in nonlinear elasticity*, Philos. Trans. Roy. Soc. London Ser. A, 306 (1982), pp. 557–611.

[2] J. M. Ball, B. Kirchheim, and J. Kristensen, *Regularity of quasiconvex envelopes*, Calc. Var. Partial Differential Equations, 11 (2000), pp. 333–359.

[3] P. J. Blatz and W. L. Ko, *Application of finite elastic theory to the deformation of rubbery materials*, Trans. Soc. Rheology, 6 (1962), pp. 223–251.

[4] P. Celada, D. Mucci, P. Pedregal, and S. Perrotta, *A Radial Variational Model Allowing for Cavitation*, in preparation.

[5] P. Celada and S. Perrotta, *Existence of minimizers for nonconvex, noncoercive simple integrals*, SIAM J. Control Optim., 41 (2002), pp. 1118–1140.

[6] A. Cellina and S. Perrotta, *On minima of radially symmetric functionals of the gradient*, Nonlinear Anal., 23 (1994), pp. 239–249.

[7] S. Conti and C. De Lellis, *Some remarks on the theory of elasticity for Neohookean materials*, Ann. Sc. Norm. Super. Pisa Cl. Sci. (5), 2 (2003), pp. 521–549.

[8] G. Crasta, *Existence, uniqueness and qualitative properties of minima to radially symmetric, non-coercive non-convex variational problems*, Math. Z., 235 (2000), pp. 569–589.

[9] G. Crasta, *On a class of nonconvex noncoercive Bolza problems with constraints on the derivatives*, J. Optim. Theory Appl., 118 (2003), pp. 295–325.

[10] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Appl. Math. Sci. 78, Springer, Berlin, 1989.

[11] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, Stud. Math. Appl. 1, North–Holland, Amsterdam, 1976.

[12] M. Giaquinta, G. Modica, and J. Souček, *Cartesian Currents in the Calculus of Variations I*, Ergeb. Math. Grenzgeb. (3) 37, Springer, Berlin, 1998.

[13] M. Giaquinta, G. Modica, and J. Souček, *Cartesian Currents in the Calculus of Variations II*, Ergeb. Math. Grenzgeb. (3) 38, Springer, Berlin, 1998.

[14] C. O. Horgan and R. Abeyaratne, *A bifurcation problem for a compressible nonlinearly elastic medium: Growth of a microvoid*, J. Elasticity, 16 (1986), pp. 189–200.

[15] C. O. Horgan and D. A. Polignone, *Cavitation in nonlinearly elastic solids: A review*, Appl. Mech. Rev., 48 (1995), pp. 471–485.

[16] B. Kirchheim and J. Kristensen, *Differentiability of convex envelopes*, C. R. Acad. Sci. Paris Sér. I Math., 333 (2001), pp. 725–728.

[17] R. J. Knops and C. A. Stuart, *Quasiconvexity and uniqueness of equilibrium solutions in nonlinear elasticty*, Arch. Rational Mech. Anal., 86 (1984), pp. 233–249.

[18] J. K. Knowles and E. Sternberg, *On the ellipticity of the equations of nonlinear elastostatics for a special material*, J. Elasticity, 5 (1975), pp. 341–361.

[19] P. Marcellini, *The stored-energy for some discontinuous deformations in nonlinear elasticity*, in Partial Differential Equations and the Calculus of Variations, Vol. II, Progr. Nonlinear Differential Equations Appl. 2, F. Colombini, A. Marino, L. Modica, and S. Spagnolo, eds., Birkäuser Boston, Boston, 1989, pp. 767–786.

[20] P. Marcellini, *Nonconvex integrals of the calculus of variations*, in Methods of Nonconvex Analysis (Varenna, 1989), Lecture Notes in Math. 1446, A. Cellina, ed., Springer, Berlin, 1990, pp. 16–57.

[21] S. Müller and S. J. Spector, *An existence theory for nonlinear elasticity that allows for cavitation*, Arch. Rational Mech. Anal., 131 (1995), pp. 1–66.

[22] S. Müller and V. Šverák, *Attainment results for the two-well problem by convex integration*, in Geometric Analysis and the Calculus of Variations, J. Jost, ed., International Press, Cambridge, MA, 1996, pp. 239–251.

[23] P. Podio-Guidugli, G. Vergara Caffarelli, and E. G. Virga, *Discontinuous energy minimizers in nonlinear elastostatics: An example of J. Ball revisited*, J. Elasticity, 16 (1986), pp. 75–96.

[24] T. Rockafeller, *Convex analysis*, Princeton University Press, Princeton, NJ, 1972.

[25] J. Sivaloganathan, *A field theory approach to stability of radial equilibria in nonlinear elasticity*, Math. Proc. Cambridge Philos. Soc., 99 (1986), pp. 589–604.

[26] J. Sivaloganathan, *Uniqueness of regular and singular equilibria for spherically symmetric problems of nonlinear elasticity*, Arch. Rational Mech. Anal., 96 (1986), pp. 97–136.

[27] J. Sivaloganathan, *On the stability of cavitating equilibria*, Quart. Appl. Math., 53 (1995), pp. 301–313.

[28] J. Sivaloganathan and S. J. Spector, *On the existence of minimizers with prescribed singular points in nonlinear elasticity*, J. Elasticity, 59 (2000), pp. 83–113.

[29] J. Sivaloganathan and S. J. Spector, *On the optimal location of singularities arising in variational problems of nonlinear elasticity*, J. Elasticity, 58 (2000), pp. 191–224.

[30] S. J. Spector, *Linear deformations as global minimizers in nonlinear elasticty*, Quart. Appl. Math., 52 (1994), pp. 59–64.

[31] C. A. Stuart, *Radially symmetric cavitation for hyperelastic materials*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 2 (1985), pp. 33–66.

SIAM J. Math. Anal.
Vol. 39, No. 3, pp. 863–890

INCOMPRESSIBLE IONIZED NON-NEWTONIAN FLUID MIXTURES*

TOMÁŠ ROUBÍČEK†

Abstract. The model combining Navier–Stokes equations in a non-Newtonian p-power-law modification for barycentric velocity together with the Nernst–Planck equation for concentrations of particular mutually reacting constituents, the heat equation, and the Poisson equation for a self-induced quasi-static electric field is formulated, existence of its (very) weak solutions is proved for $p > 11/5$, and its thermodynamics is discussed.

Key words. chemically reacting fluids, Eckart–Prigogine concept, Navier–Stokes equation, Nernst–Planck equation, Poisson equation, heat equation

AMS subject classifications. 35Q35, 76T30, 80A20, 92C05

DOI. 10.1137/060667335

1. Introduction.

Chemically reacting mixtures represent a framework for modeling various complicated processes in biology and chemistry. The research in this area, resulting in a model [36], was initiated by Nečas who, for many years before he passed away, spoke about "living fluids," although he never elaborated on any concept of such fluids. The model proposed in [36] uses an incompressible Newtonian framework with the barycentric impulse balance. This "barycentric" approach is called the Eckart–Prigogine [13, 29] concept, phenomenologically simplifying the description by considering only one temperature and one velocity of the whole mixture and being awarded in the context of nonequilibrium thermodynamics of dissipative structures the Nobel prize in chemistry in 1977; for the compressible case, see also [3, 5, 12, 17]. The incompressibility refers here to each particular constituent and, through a volume-additivity hypothesis (i.e., Amagat's law) as in, e.g., [22, 31, 40], also to the overall mixture. To cover biological applications on a cellular or subcellular level where the intensity of the electric field on cell membranes is very high, the self-induced electrostatic field must be considered; recall that the intracellular electric potential usually ranges between 60 and 100 mV, while the thickness of cell membranes is on the order of 10–100 nm, which results in an intensity of the electric field on the order of 1 MV/m.

In comparison with [33, 34, 36] or [35, sect. 12.6], we consider here a more general model exploiting the non-Newtonian concept with a (possibly temperature-dependent) shear-thickening p-power-law stress tensor and admitting diffusive fluxes with different mobilities, and we prove existence of its solution in a fully coupled and fully nonlinear case. The key mathematical tool is a nonvariational technique for the heat equation based on integrability of temperature gradient observed in [6, 7, 8] combined with a regularization of the Navier–Stokes equation and a sophisticated limit passage. Finally, in section 4, thermodynamics of a specific model is discussed.

*Received by the editors August 11, 2006; accepted for publication (in revised form) February 23, 2007; published electronically September 5, 2007. This work was created as a research activity of "Nečas Center for Mathematical Modeling" LC 06052 (MŠMT ČR), partly supported also by grants 201/06/0352 (GA ČR) and MSM 21620839 (MŠMT ČR).

http://www.siam.org/journals/sima/39-3/66733.html

†Mathematical Institute, Charles University, Sokolovská 83, CZ-186 75 Prague 8, and Institute of Information Theory and Automation, Academy of Sciences, Pod vodárenskou věží 4, CZ-182 08 Prague 8, Czech Republic (tomas.roubicek@mff.cuni.cz).

2. The model: A general framework. We consider a three-dimensional incompressible flow of a mixture of L mutually reacting chemical ionic constituents, the ℓth constitutent having a specific charge z_ℓ, $\ell = 1, \dots, L$. Our model consists in a system of 4+L+1+1 differential equations combining the *non-Newtonian* modification of the *Navier–Stokes equation* (balancing the barycentric momentum ϱv) with the incompressibility constraint $\mathrm{div}(v) = 0$, the *Nernst–Planck equation* modified for moving media (balancing the mass of particular constituents), the *heat equation* (balancing the heat part $c_\mathrm{v}\theta$ of the internal energy u; cf. (4.17) below), and the quasi-static *Poisson equation* for the electrostatic field (balancing the electric induction $\varepsilon\nabla\phi$):

$$\varrho\frac{\partial v}{\partial t} - \mathrm{div}\big(\tau(\mathrm{D}v, c, \theta) - \varrho v \otimes v\big) + \nabla\pi = -q\nabla\phi, \qquad q = c\cdot z, \tag{2.1a}$$

$$\mathrm{div}(v) = 0, \tag{2.1b}$$

$$\frac{\partial c}{\partial t} - \mathrm{div}\big(\mathfrak{D}(c,\theta)\nabla c + \mathfrak{m}(c,\theta)\otimes\nabla\phi - c\otimes v\big) = r(c,\theta)\,, \tag{2.1c}$$

$$\begin{aligned} c_\mathrm{v}\frac{\partial\theta}{\partial t} - \mathrm{div}\big(\kappa\nabla\theta - c_\mathrm{v} v\,\theta\big) &= \tau(\mathrm{D}v, c, \theta){:}\mathrm{D}v \\ &\quad + \big(\mathfrak{D}(c,\theta)\nabla c + \mathfrak{m}(c,\theta)\otimes\nabla\phi\big) : (z\otimes\nabla\phi) + h(c,\theta), \end{aligned} \tag{2.1d}$$

$$-\,\mathrm{div}(\varepsilon\nabla\phi) = q. \tag{2.1e}$$

The variables v, π, c, θ, and ϕ have the following meanings:

$v = (v_1, v_2, v_3)$: barycenter velocity,
π: pressure,
$c = (c_1, \dots, c_L)$: the vector of concentrations of particular constituents,
ϕ: electrostatic potential,
θ: temperature,

where the concentration vector c is to satisfy the constraint

$$\forall \ell = 1, \dots, L: \quad c_\ell(t,x) \ge 0 \quad \text{and} \quad \sum_{\ell-1}^{L} c_\ell(t,x) = 1 \quad \text{for a.a. } (t,x). \tag{2.2}$$

We will write (2.2) briefly as $c(t,x)\in G_1^+$ for a.a. (t,x), where we denote

$$G_1^+ := \{c\in\mathbb{R}^L; \quad c\cdot\mathbf{1} = 1, \quad \forall\ell = 1,\dots,L{:} \quad c_\ell \ge 0\} \tag{2.3}$$

with $\mathbf{1} \in \mathbb{R}^L$ denoting the "unit" vector $(1, \dots, 1)$; usually G_1^+ is called the *Gibbs simplex*. The meaning of the scalar or tensorial products (denoted by "$\cdot$" and "$\otimes$," respectively) is standard, while "$:$" means $[\tau_{ij}]{:}[e_{ij}] = \sum_{i=1}^n \sum_{j=1}^n \tau_{ij}e_{ij}$; i.e., (2.1a), (2.1b) means

$$\varrho\frac{\partial v_i}{\partial t} - \sum_{j=1}^{3}\frac{\partial}{\partial x_j}\big(\tau_{ij}(\mathrm{Dv}, c, \theta) - \varrho v_i v_j\big) + \frac{\partial\pi}{\partial x_i} = -q\frac{\partial\phi}{\partial x_i}, \quad q = \sum_{\ell=1}^{L} c_\ell z_\ell, \tag{2.4a}$$

$$\sum_{j=1}^{3}\frac{\partial v_j}{\partial x_j} = 0, \tag{2.4b}$$

while (2.1c) means

$$\frac{\partial c_\ell}{\partial t} - \sum_{i=1}^{3}\frac{\partial}{\partial x_i}\left(\sum_{k=1}^{L}\mathfrak{D}_{k\ell}(c,\theta)\frac{\partial c_k}{\partial x_i} + \mathfrak{m}_\ell(c,\theta)\frac{\partial\phi}{\partial x_i} - v_i c_\ell\right) = r_\ell(c,\theta) \tag{2.5}$$

for any $\ell = 1, \ldots, L$, and (2.1d) means

$$c_{\mathrm{v}}\frac{\partial\theta}{\partial t} - \sum_{i=1}^{3}\frac{\partial}{\partial x_i}\Big(\kappa\frac{\partial\theta}{\partial x_i} - c_{\mathrm{v}}v_i\,\theta\Big) = \sum_{i=1}^{3}\sum_{j=1}^{3}\tau_{ij}(\mathrm{D}v,c,\theta)\Big(\frac{1}{2}\frac{\partial v_j}{\partial x_i} + \frac{1}{2}\frac{\partial v_i}{\partial x_j}\Big) + \sum_{\ell=1}^{L}\sum_{i=1}^{3}\left(\sum_{k=1}^{L}\mathfrak{D}_{k\ell}(c,\theta)\frac{\partial c_k}{\partial x_i} + \mathfrak{m}_\ell(c,\theta)\frac{\partial\phi}{\partial x_i}\right)z_\ell\frac{\partial\phi}{\partial x_i} + h(c,\theta). \tag{2.6}$$

We use the following notation:
$\tau = [\tau_{ij}]_{i,j=1}^3$: the stress tensor, depending on $(\mathrm{D}v, c, \theta)$,
$\mathrm{D}v = \frac{1}{2}(\nabla v)^\top + \frac{1}{2}\nabla v$: the symmetrized velocity gradient,
$\varrho > 0$: mass density (assumed to be equal to 1 in what follows),
$z = (z_1, \ldots, z_L)$: the vector of specific charges of particular constituents,
$q = c \cdot z$: the total charge, depending on time t and space x,
$\varepsilon > 0$: permitivity,
$r = (r_1, \ldots, r_L)$: the vector of chemical production rates, depending on (c, θ),
h: the heat production rate due to all chemical reactions, depending on (c, θ),
$\mathfrak{D} = [\mathfrak{D}_{kl}]_{k,l=1}^L$: the matrix of diffusion coefficients, depending on (c, θ),
$\mathfrak{m} = (\mathfrak{m}_1, \ldots, \mathfrak{m}_L)$: the vector of effective mobilities, depending on (c, θ),
$\kappa > 0$: thermal conductivity, and
$c_{\mathrm{v}} > 0$: heat capacity.
The system (2.1) is to be completed by the initial conditions

$$v(0,\cdot) = v_0\,, \qquad c(0,\cdot) = c_0\,, \qquad \theta(0,\cdot) = \theta_0 \tag{2.7}$$

on a bounded C^2-domain under consideration, and by the boundary conditions corresponding, e.g., to a closed container, which, in some simplified version, leads, respectively, to

$$v = 0, \tag{2.8a}$$

$$\big(\mathfrak{D}(c,\theta)\nabla c + \mathfrak{m}(c,\theta)\otimes\nabla\phi\big)\nu = 0, \tag{2.8b}$$

$$\varepsilon\frac{\partial\phi}{\partial\nu} = \alpha(\phi_\Sigma - \phi), \tag{2.8c}$$

$$\kappa\frac{\partial\theta}{\partial\nu} = 0 \tag{2.8d}$$

on $(0, T) \times \partial\Omega$, where ν is the unit outward normal to the boundary $\partial\Omega$, ϕ_Σ is a prescribed external electric potential, and $\alpha > 0$ is a "lumped capacity" of the boundary $\partial\Omega$.

Remark 2.1 (right-hand sides of (2.1)). The right-hand side of (2.1a) represents the *Lorenz force* $q\nabla\phi$ due to Coulomb electrostatic interactions. The particular terms on the right-hand side of (2.1d) represent, respectively, the production rate of the *dissipative heat* due to friction in the fluid $\tau(\mathrm{D}v, c, \theta){:}\mathrm{D}v$, the power $(\mathfrak{D}(c,\theta)\nabla c) : (z \otimes \nabla\phi)$ of the electric current arising by the diffusion flux $z^\top(\mathfrak{D}(c,\theta)\nabla c)$ in the electric field gradient $\nabla\phi$ (so-called *Peltier effect*), the power of the *Joule heat* produced by the electric current $(\mathfrak{m} \otimes \nabla\phi) : (z \otimes \nabla\phi) = (z{\cdot}\mathfrak{m})|\nabla\phi|^2$, and, as already stated, the heat production rate due to all chemical reactions h; see also Remark 4.6 below.

Remark 2.2 (Fourier's, Fick's, Ohm's laws). The model (2.1) involves various phenomenological laws. Certainly, (2.1d) relies on the conventional *Fourier law* in linear isotropic homogenous medium; i.e., the heat flux $-\kappa\nabla\theta$ is proportional to the negative temperature gradient. Further, (2.1c) involves a certain generalization of *Fick's law* stating that diffusive fluxes are proportional to negative concentration gradients; here, however, cross-effects make it more complicated, and a nonconstant diffusivity matrix $\mathfrak{D}$ occurs instead of a single constant; see also (3.8) below. In view of Remark 2.1, the effective *electric conductivity* is $\sigma := z{\cdot}\mathfrak{m}$, and we can identify *Ohm's law* that the electric current $(z{\cdot}\mathfrak{m})\nabla\phi$ is proportional to the gradient $\nabla\phi$ of the electric field just via σ. Naturally, σ now depends through $\mathfrak{m} = \mathfrak{m}(c,\theta)$ on the ion concentrations c.

Remark 2.3 (simplifying assumptions). It should be emphasized that many simplifications are adopted in the presented model. In particular, we have considered small electrical currents (i.e., the magnetic field is neglected), we have adopted the mentioned volume-additivity and incompressibility assumption, we have further assumed mass densities equal for all constituents and diffusion fluxes independent of the temperature gradient (i.e., Soret's effect is neglected), and then, in agreement with Onsager's reciprocity principle, we also have considered the heat flux independent of the concentration gradients (i.e., Dufour's effect is neglected). Detailed identification of simplifying assumptions related to (2.1) in comparison with the rational Truesdell concept [24, 30, 38, 39, 41, 42] was made by Samohýl [40].

3. Analysis of the model: Existence of a solution. We will prove the existence of a very weak solution, defined in section 3.1, in several steps. First, in section 3.2 we treat an auxiliary, so-called *multipolar* regularization of the Navier–Stokes equation and prove existence of its solution by Schauder's fixed point technique in a way similar to how it was done in [33] for the spatial regular case (except that [33] had thus assumed composition/temperature-independent potential stress-tensor τ). Then, in section 3.3 we pass this regularization to zero. This two-step approach allows us to avoid any regularity results for p-power-law non-Newtonian fluids (and thus any qualifications of data related to them) and to admit a temperature-dependent stress tensor and rather low exponent $p > 11/5$; this bound even improves some particular known results; cf. Remark 3.4. The advantage of the smoothing is that it avoids difficult (or even unrealistic) requirements of uniqueness or convexity of the set of solutions of decoupled systems needed for Schauder's or Kakutani's fixed point theorems. It is also particularly important for the multicomponent fluids to have $\frac{\partial}{\partial t}c$ in duality with the negative part of c to get $c \geq 0$; cf. (3.57). As a side effect, it simplifies some other arguments; e.g., it ensures the sum-equals-one property $\sum_{\ell=1}^{L} c_\ell = 1$ in (3.47) for smooth velocity field v, and the unique response (see, e.g., (3.50)) for the fixed-point mapping that is used in Schauder's fixed point theorem.

3.1. Definition of a very weak solution and data qualification. We consider an evolution of (2.1) on a fixed time interval $(0,T)$. We use a standard notation $C^1(\cdot;\mathbb{R}^n)$ of continuously differentiable $\mathbb{R}^n$-valued functions, and $L^p(\cdot;\mathbb{R}^n)$ for Lebesgue L^p-spaces as well as $W^{k,p}(\cdot;\mathbb{R}^n)$ for the Sobolev spaces on the domain indicated. Let us abbreviate $I := (0,T)$, $Q := I\times\Omega$, $\Sigma := I\times\Gamma$, $\Gamma := \partial\Omega$, and let $W^{k,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)$ denote the space of functions from the zero-trace Sobolev space $W^{k,p}_0(\Omega;\mathbb{R}^3)$ but with zero divergence (in the distributional sense), and later we will also use $L^2_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)$ to denote the closure of $W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)$ in $L^2(\Omega;\mathbb{R}^3)$. Usage of such divergence-free test functions has the usual effect that the pressure π disappears from the (very) weak formulation. If $\mathbb{R}^n$ is replaced by a Banach space X, then $L^p(I;X)$ refers to the L^p-Bochner space of Banach-space-valued functions, while $W^{k,p}(I;X)$ is a Sobolev–Bochner space. We also denote standardly

$W^{-k,2}(\Omega) = W_0^{k,2}(\Omega)^*$.

The adjective "very weak" is used to emphasize that, contrary to conventional weak solutions, the very weak solutions have less regularity than possible test functions, which particularly concerns the temperature.

DEFINITION 3.1 (very weak solution). *We will call*

$$v \in L^p(I; W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)) \cap W^{1,p/(p-1)}(I; W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)^*), \tag{3.1a}$$

$$c \in L^\infty(I; L^2(\Omega;\mathbb{R}^L)) \cap L^2(I; W^{1,2}(\Omega;\mathbb{R}^L)) \cap W^{1,r/(r-1)}(I; W^{1,r}(\Omega;\mathbb{R}^L)^*), \tag{3.1b}$$

$$\theta \in L^\infty(I; L^1(\Omega)) \cap L^{5/4-\xi}(I; W^{1,5/4-\xi}(\Omega)) \cap W^{1,1}(I; W^{-3,2}(\Omega)), \tag{3.1c}$$

$$\phi \in L^\infty(I; W^{1,2}(\Omega)) \tag{3.1d}$$

with any $\xi > 0$ *and* $r = \max(2, 10p/(7p-6))$ *a very weak solution to the system* (2.1)–(2.2) *with the initial and boundary conditions* (2.7) *and* (2.8) *if*

$$\int_Q \tau(\mathrm{D}v, c, \theta):\mathrm{D}w - (v\otimes v):\nabla w + (z\cdot c)(\nabla\phi\cdot w) - v\frac{\partial w}{\partial t}\,\mathrm{d}x\mathrm{d}t = \int_\Omega v_0(x)\cdot w(0,x)\,\mathrm{d}x \tag{3.2}$$

for any $w \in C^1(Q;\mathbb{R}^3)$ *with* $\operatorname{div} w = 0$, $w|_\Sigma = 0$, *and* $w(T,\cdot) = 0$,

$$\int_Q \big(\mathfrak{D}(c,\theta)\nabla c + \mathfrak{m}(c,\theta)\otimes\nabla\phi - c\otimes v\big):\nabla w + r(c,\theta)w - c\cdot\frac{\partial w}{\partial t}\,\mathrm{d}x\,\mathrm{d}t = \int_\Omega c_0(x)\cdot w(0,x)\,\mathrm{d}x \tag{3.3}$$

with the test-function $w \in C^1(Q;\mathbb{R}^L)$ *arbitrary with* $w(T,\cdot) = 0$,

$$\int_Q \varepsilon\nabla\phi\cdot\nabla w - qw\,\mathrm{d}x\,\mathrm{d}t + \int_\Sigma \phi\alpha w\,\mathrm{d}S\,\mathrm{d}t = \int_\Sigma \alpha\phi_\Sigma w\,\mathrm{d}S\,\mathrm{d}t \tag{3.4}$$

for any $w \in C^1(Q)$, *and*

$$\int_Q \big(\kappa\nabla\theta - c_{\mathrm{v}}v\theta\big)\cdot\nabla w - \Big(\tau(\mathrm{D}v,c,\theta):\mathrm{D}v + \big(\mathfrak{D}(c,\theta)\nabla c + \mathfrak{m}(c,\theta)\otimes\nabla\phi\big) : (z\otimes\nabla\phi) + h\Big)w - c_{\mathrm{v}}\theta\frac{\partial w}{\partial t}\,\mathrm{d}x\,\mathrm{d}t = c_{\mathrm{v}}\int_\Omega \theta_0(x)w(0,x)\,\mathrm{d}x \tag{3.5}$$

for any $w \in C^1(Q)$ *with* $w(T,\cdot) = 0$ *on* Ω. *Finally,* (2.2) *is to be satisfied, too.*

We naturally assume the mass conservation in all chemical reactions, and the volume-additivity constraint holding for the initial conditions c_0, i.e.,

$$r(c,\theta)\cdot\mathbf{1} = 0, \tag{3.6a}$$

$$c_0\cdot\mathbf{1} = 1, \qquad [c_0]_\ell \geq 0 \quad \forall \ell = 1,\dots,L. \tag{3.6b}$$

Other important qualifications concern the diffusion matrix $\mathfrak{D}$ and the effective-mobility vector $\mathfrak{m}$:

(3.7a) $\mathfrak{D}: G_1^+ \times \mathbb{R} \to \mathbb{R}^{L\times L}$, $\mathfrak{m}: G_1^+ \times \mathbb{R} \to \mathbb{R}^L$ continuous and bounded,

(3.7b) $\exists \eta_0 > 0\ \forall c \in G_1^+,\ \theta \in \mathbb{R},\ d \in \mathbb{R}^L: \quad d^\top \mathfrak{D}(c,\theta) d := \sum_{\ell=1}^L \sum_{k=1}^L \mathfrak{D}_{k\ell}(c,\theta) d_k d_\ell \geq \eta_0 |d|^2,$

(3.7c) $\exists \beta \geq 0\ \forall k = 1, \ldots, L: \quad \sum_{\ell=1}^L \mathfrak{D}_{k\ell}(c,\theta) = \beta,$

(3.7d) $\sum_{\ell=1}^L \mathfrak{m}_\ell(c,\theta) = 0,$

where ${G_1}^+$ is from (2.3). Note that, if $\beta = 0$, then (3.7c), (3.7d) means that the sum $\sum_{\ell=1}^L j_\ell$ of the diffusive fluxes

$$j_\ell := \sum_{k=1}^L \mathfrak{D}_{k\ell}(c,\theta) \nabla c_k + \mathfrak{m}_\ell(c,\theta) \nabla \phi \tag{3.8}$$

is identically zero, which is to hold the equality constraint in (2.2). Essentially the same effect is made by (3.7c), (3.7d) also if $\beta > 0$; cf. the arguments around (3.47). In fact, (3.7b) suffices to hold only for d with $\sum_{\ell=1}^L d_\ell = 0$ if later (3.47) is used simultaneously when testing (3.18d) by c in the proof of (3.22e), e.g., in an additional auxiliary Galerkin approximation just for (3.18d) itself. This makes (3.7c) indeed consistent with (3.7b) even for $\beta = 0$.

As for the stress tensor $\tau: \mathbb{R}^{3\times3}_{\rm sym} \times G_1^+ \times \mathbb{R} \to \mathbb{R}^{3\times3}_{\rm sym}$, where $\mathbb{R}^{3\times3}_{\rm sym}$ denotes the set of symmetric 3×3-matrices, we assume that, for some $\eta_1 > 0$, $C \in \mathbb{R}$, it satisfies

(3.9a) $\tau: \mathbb{R}^{3\times3}_{\rm sym} \times G_1^+ \times \mathbb{R} \to \mathbb{R}^{3\times3}_{\rm sym}$ continuously differentiable,

(3.9b) $\forall D_1, D_2 \in \mathbb{R}^{3\times3}_{\rm sym},\ c \in G_1^+,\ \theta \in \mathbb{R}: \quad \big(\tau(D_1,c,\theta) - \tau(D_2,c,\theta)\big):(D_1 - D_2) \geq \eta_1 \big|D_1 - D_2\big|^p,$

(3.9c) $\forall D \in \mathbb{R}^{3\times3}_{\rm sym},\ c \in G_1^+,\ \theta \in \mathbb{R}: \quad |\tau(D,c,\theta)| \leq C\big(1 + |D|^{p-1}\big),$

(3.9d) $\big|\tau'_D(D,c,\theta)\big| \leq C\big(1 + |D|^{p-2}\big),$

(3.9e) $\big|\tau'_{(c,\theta)}(D,c,\theta)\big| \leq C,$

(3.9f) $\tau(0,c,\theta) = 0.$

Note that (3.9b), (3.9f) yields the coercivity $\tau(D,c,\theta) : D \geq \eta_1 |D|^p$. Other important assumptions ensure nonnegativity of concentrations during their evolution, namely by nonnegative production rate and by a natural direction of the flux j_ℓ of the ℓth constituent from (3.8) if the concentration of this particular constituent vanishes:

(3.10a) $\mathfrak{D}_{k\ell}(c_1, \ldots, c_{\ell-1}, 0, c_{\ell+1}, \ldots, c_L, \theta) \begin{cases} \geq 0 & \text{for } k = \ell, \\ = 0 & \text{for } k \neq \ell, \end{cases}$

(3.10b) $\mathfrak{m}_\ell(c_1, \ldots, c_{\ell-1},\ 0,\ c_{\ell+1}, \ldots, c_L, \theta) = 0\,,$

(3.10c) $r_\ell\,(c_1, \ldots, c_{\ell-1},\ 0,\ c_{\ell+1}, \ldots, c_L, \theta) \geq 0$

for each $\ell = 1, \ldots, L$. Eventually, we still assume

$$r_\ell, h: G_1^+ \times \mathbb{R} \to \mathbb{R} \ \text{ continuous and bounded.} \tag{3.11}$$

Remark 3.2 (extension convention). For the purpose of the proof of Proposition 3.9, we consider $\mathfrak{D}$, $\mathfrak{m}$, τ, r_ℓ, and h extended suitably from the Gibbs' simplex G_1^+ defined by (2.3) on the affine manifold

$$G_1 := \left\{ c \in \mathbb{R}^L;\ \sum_{\ell=1}^{L} c_\ell = 1 \right\}. \tag{3.12}$$

We assume continuous and bounded extension so that (3.7b)–(3.7d) and (3.9b)–(3.9e) hold even for $c \in G_1 \setminus G_1^+$. Moreover, (3.10) allows us to consider nonnegative extensions of r_ℓ and zero-extension of $\mathfrak{m}_\ell$ if $c_\ell \le 0$.

Remark 3.3 (data qualification versus reality). The assumption (3.11) represents a rather drastic mathematical simplification contrasting with the usual feature that the rate of chemical reactions r and the corresponding heat production h depend rather exponentially on the temperature θ. In fact, making the estimates in section 3.2 in a still more complicated (and less lucid) way, a certain (although only sublinear) growth of $r(c,\cdot)$ and $h(c,\cdot)$ may be admitted, too; cf. also [36]. The mentioned exponential growth would allow for "explosive" blow-ups which we do not have in mind, especially in the context of usual biological applications. Also, (3.7b) is not directly relevant and contradicts an Einstein law if $\theta \searrow 0$; cf. also the arguments in Remark 4.6. Yet, considering $\mathfrak{D}(c,\theta)$ approaching zero if $\theta \searrow 0$ would inevitably make the analysis of the problem extremely difficult, if possible at all. Anyhow, the model of fluid mixtures loses its validity much earlier than the absolute temperature θ approaches zero because of ultimate phase transition to solid state.

Remark 3.4 (special case: single-component fluids). A subsystem (2.1a), (2.1b), and (2.1d) with $\mathfrak{D}$ and $\mathfrak{m}$ vanishing and with a general heat flux $j_0(\theta, \nabla\theta)$ instead of $\kappa\nabla\theta$ together with a fixed right-hand side instead of $q\nabla\phi$ was considered in [9]. Assuming monotonicity and p_0-polynomial structure of $j_0(\theta,\cdot) : \mathbb{R}^3 \to \mathbb{R}^3$, existence of a weak solution was proved for $p \ge 5/2$ and $p_0 \ge 10p/(5p-1) > 2$. Physically, the heat flux j_0 may depend substantially on θ, but the dependence of $\nabla\theta$ is rather linear, which corresponds to the case $p_0 = 2$ not covered by [9]. Our results cover, in particular, the case $11/5 < p < 5/2$ and enable us to treat the physically more relevant case $p_0 = 2$ for such a subsystem (2.1a), (2.1b), (2.1d). Also, [9] assumed $\tau(\cdot,\theta)$ to have a potential, even with a special structure $\sum_{l=1}^{L} \mu_l(\theta) F_l(|\cdot|)$ (which we do not need at all) but, on the other hand, allowed for a temperature dependence of $c_{\rm v}$ and κ.

3.2. Auxiliary multipolar regularization. We will regularize (2.1a) by a $2k$th-order term $(-1)^k \epsilon \Delta^k v$ with a regularization parameter $\epsilon > 0$ and with an integer $k \ge 5$ specified later (see (3.21) with (3.29)) as follows:

$$\frac{\partial v}{\partial t} - \mathrm{div}\big(\tau(\mathrm{D}v, c, \theta) - v \otimes v\big) + \nabla\pi + (-1)^k \epsilon \Delta^k v = -c{\cdot}z\nabla\phi, \tag{3.13a}$$

$$\mathrm{div}(v) = 0. \tag{3.13b}$$

Such a "multipolar" regularization is even physically motivated; cf. [25]. Let us emphasize that we distinguish ε (the permitivity) from ϵ (the regularizing parameter). The boundary conditions (2.8) are now to be completed by another higher-order condition for the Δ^k-operator. In fact, its choice is not important as this term has only an auxiliary character; let us choose, say, the homogeneous Dirichlet condition

$$\frac{\partial^l v}{\partial \nu^l} = 0, \qquad l = 1, \ldots, k-1. \tag{3.14}$$

We modify Definition 3.1 for a weak solution to the system (2.1c)–(2.1e) and (3.13) with the initial and boundary conditions (2.7) and (2.8) and (3.14).

DEFINITION 3.5 (weak solution to (2.1c)–(2.1e) and (3.13)). *We will call* (v, c, θ, ϕ) *satisfying*

$$v \in L^\infty(I; W^{k,2}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)) \cap W^{1,2}(I; L^2(\Omega;\mathbb{R}^3)), \tag{3.15a}$$

$$c \in L^2(I; W^{1,2}(\Omega;\mathbb{R}^L)) \cap W^{1,2}(I; W^{1,2}(\Omega;\mathbb{R}^L)^*), \tag{3.15b}$$

$$\theta \in L^\infty(I; W^{1,2}(\Omega)) \cap W^{1,2}(I; L^2(\Omega)), \tag{3.15c}$$

$$\phi \in L^\infty(I; W^{1,2}(\Omega)) \tag{3.15d}$$

a weak solution to the system (2.1c)–(2.1e) *and* (3.13) *with the initial and boundary conditions* (2.7), (2.8), *and* (3.14) *if* (3.3), (3.4), (3.5), *and* (2.2) *hold, while, instead of* (3.2), *we require the following identity to hold for any* $w \in C^1(Q;\mathbb{R}^3)$ *with* $\mathrm{div}(w) = 0$, $\frac{\partial^l}{\partial\nu^l} w|_\Sigma = 0$ *for* $l = 0, \dots, k-1$, *and* $w(T, \cdot) = 0$:

$$\int_Q \tau(\mathrm{D}v, c, \theta) : \mathrm{D}w - (v \otimes v) : \nabla w + z \cdot c \nabla\phi \cdot w + \epsilon \nabla^k v \,\vdots\, \nabla^k w - v \frac{\partial w}{\partial t} \,\mathrm{d}x\,\mathrm{d}t = \int_\Omega v_0(x) \cdot w(0, x) \,\mathrm{d}x, \tag{3.16}$$

where " $\vdots$ *" denotes the scalar product of kth-order tensors; for* $k = 1$ *or* $k = 2$ *we already used "* $\cdot$ *" or "* $:$*," respectively.*

To correct the concentrations that satisfy the constraint $\sum_{\ell=1}^L c_\ell = 1$ but that may possibly be negative, we define a retract $K : G_1 \to G_1^+$ by

$$K_\ell(c) := \frac{c_\ell^+}{\sum_{l=1}^L c_l^+}, \qquad c_\ell^+ := \max(c_\ell, 0), \tag{3.17}$$

where G_1 is from (3.12). Let us note that K is continuous and bounded on G_1 and leaves G_1^+ fixed, and even $K_\ell(c) = 0$ if $c_\ell \le 0$. Further, we consider r, h, $\mathfrak{D}$, and $\mathfrak{m}$ continuously and boundedly extended on G_1. Considering $\gamma = (\gamma_1, \dots, \gamma_L) =$ "old" concentrations and $\vartheta =$ an "old" temperature field, we define the quadruple (v, c, θ, ϕ) as the weak solution to the *decoupled regularized system*:

$$-\,\mathrm{div}(\varepsilon \nabla\phi) = q, \qquad q = z \cdot K(\gamma), \tag{3.18a}$$

$$\frac{\partial v}{\partial t} - \mathrm{div}\big(\tau(\mathrm{D}v, \gamma, \vartheta) - v \otimes v\big) + \nabla\pi + (-1)^k \epsilon \Delta^k v = -q \nabla\phi, \tag{3.18b}$$

$$\mathrm{div}(v) = 0\,, \tag{3.18c}$$

$$\frac{\partial c}{\partial t} - \mathrm{div}\big(\mathfrak{D}(\gamma, \vartheta) \nabla c + \mathfrak{m}(\gamma, \vartheta) \otimes \nabla\phi - c \otimes v\big) = r(\gamma, \vartheta)\,, \tag{3.18d}$$

$$c_\mathrm{v} \frac{\partial\theta}{\partial t} - \mathrm{div}\big(\kappa \nabla\theta - c_\mathrm{v} v\, \theta\big) = \tau(\mathrm{D}v, \gamma, \vartheta) : \mathrm{D}v + \big(\mathfrak{D}(\gamma, \vartheta) \nabla c + \mathfrak{m}(\gamma, \vartheta) \otimes \nabla\phi\big) : (z \otimes \nabla\phi) + h(\gamma, \vartheta), \tag{3.18e}$$

$$c \cdot \mathbf{1} := \sum_{\ell=1}^L c_\ell = 1 \tag{3.18f}$$

with the boundary conditions (2.8) and (3.14) and with the initial conditions

$$v(0, \cdot) = v_{0\epsilon}\,, \qquad c(0, \cdot) = c_0\,, \qquad \theta(0, \cdot) = \theta_{0\epsilon}. \tag{3.19}$$

Obviously, given (γ,ϑ), we are to solve first (3.18a), and after knowing also ϕ we can solve (3.18b), (3.18c) to get v, and then we can solve (3.18d) to obtain c, and finally (3.18e) to obtain also θ. In (3.19), we have made a regularization of the original initial conditions $v_0\in L^2_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)$ and $\theta_0\in L^1(\Omega)$, respectively, by $v_{0\epsilon}\in W^{k,2}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)$ and $\theta_{0\epsilon}\in W^{1,2}(\Omega)$ in such a way that

$$\text{(3.20a)}\qquad \big\|v_{0\epsilon}\big\|_{W^{k,2}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)}\le\frac{C}{\epsilon},\qquad \big\|\theta_{0\epsilon}\big\|_{W^{1,2}(\Omega)}\le\frac{C}{\epsilon},\quad\text{and}$$

$$\text{(3.20b)}\qquad \big\|v_{0\epsilon}\big\|_{L^2(\Omega;\mathbb{R}^3)}\le C,\qquad \big\|\theta_{0\epsilon}\big\|_{L^1(\Omega)}\le C.$$

PROPOSITION 3.6 (a priori estimates for (3.18)). *Let the assumptions* (3.6), (3.7), (3.9), (3.10), (3.11) *hold, let* $v_{0\epsilon}\in W^{k,2}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)$, $c_0\in L^\infty(\Omega;\mathbb{R}^L)$, $\theta_{0\epsilon}\in W^{1,2}(\Omega)$ *satisfy* (3.20), *let* Ω *be a bounded* C^2*-domain, and let* $p\in\mathbb{R}$ *and* $k\in\mathbb{N}$ *satisfy*

$$\text{(3.21)}\qquad p>\frac{11}{5}\quad\textit{and}\quad k\ge\frac{5p-3}{2}.$$

Further, let $(\gamma,\vartheta)\in L^2(I;W^{1,2}(\Omega))^{L+1}$ *be given such that* $\sum_{\ell=1}^L\gamma_\ell=1$ *a.e. on* Q*. Then,* (3.18) *with the boundary condition* (2.8) *and the initial condition* (3.19) *has a weak solution (which need not satisfy* $c_\ell\ge 0$*, however) which satisfies, for any* $\xi>0$ *and some* $C_0,\dots,C_{11}<+\infty$ *independent of* ϵ*, the following a priori estimates:*

$$\text{(3.22a)}\qquad \big\|\phi\big\|_{L^\infty(I;W^{2,r}(\Omega))}\le C_1\quad\textit{with}\quad r<+\infty,$$

$$\text{(3.22b)}\qquad \big\|v\big\|_{L^\infty(I;L^2(\Omega;\mathbb{R}^3))\,\cap\,L^p(I;W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))}\le C_2,$$

$$\text{(3.22c)}\qquad \big\|v\big\|_{L^\infty(I;W^{k,2}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))\,\cap\,W^{1,2}(I;L^2(\Omega;;\mathbb{R}^3))}\le\frac{C_3\mathrm{e}^{C_0/\epsilon^2}}{\epsilon},$$

$$\text{(3.22d)}\qquad \Big\|\frac{\partial v}{\partial t}+(-1)^k\epsilon\Delta^k v\Big\|_{L^{p/(p-1)}(I;W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)^*)}\le C_4,$$

$$\text{(3.22e)}\qquad \big\|c\big\|_{L^\infty(I;L^2(\Omega;\mathbb{R}^L))\,\cap\,L^2(I;W^{1,2}(\Omega;\mathbb{R}^L))}\le C_5,$$

$$\text{(3.22f)}\qquad \Big\|\frac{\partial c}{\partial t}\Big\|_{L^2(I;W^{1,2}(\Omega;\mathbb{R}^L)^*)}\le\frac{C_6}{\sqrt{\epsilon}},$$

$$\text{(3.22g)}\qquad \Big\|\frac{\partial c}{\partial t}\Big\|_{L^{r/(r-1)}(I;W^{1,r}(\Omega;\mathbb{R}^L)^*)}\le C_7\quad\textit{with}\quad r=\max\Big(2,\frac{10p}{7p-6}\Big),$$

$$\text{(3.22h)}\qquad \big\|\theta\big\|_{L^2(I;W^{1,2}(\Omega))}\le\frac{C_8}{\sqrt{\epsilon}},$$

$$\text{(3.22i)}\qquad \big\|\theta\big\|_{L^\infty(I;W^{1,2}(\Omega))\,\cap\,W^{1,2}(I;L^2(Q))}\le\frac{C_9\mathrm{e}^{C_0/\epsilon^2}}{\epsilon},$$

$$\text{(3.22j)}\qquad \big\|\theta\big\|_{L^\infty(I;L^1(\Omega))\,\cap\,L^{5/4-\xi}(I;W^{1,5/4-\xi}(\Omega))}\le C_{10},$$

$$\text{(3.22k)}\qquad \Big\|\frac{\partial\theta}{\partial t}\Big\|_{L^1(I;W^{-3,2}(\Omega))}\le C_{11}.$$

Moreover, except for C_0*,* C_3*, and* C_9*, the constants* C *are independent of* (γ,ϑ)*, while* C_0*,* C_3*, and* C_9 *depend on* $\|(\nabla\gamma,\nabla\vartheta)\|_{L^2(Q;\mathbb{R}^3))^{L+1}}$ *due to* (3.32) *below. The meaning of* $\frac{\partial v}{\partial t}+(-1)^k\epsilon\Delta^k v$ *in* (3.22d) *as a linear continuous functional on* $L^p(I;W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))$ *is a continuous extension of the weak form of* $\mathrm{div}(\tau(\mathrm{D}v,\gamma,\vartheta)-v\otimes v)-q\nabla\phi$*, which has a good sense for smooth functions; cf.* (3.26).

Proof. First, we realize that the total charge $z{\cdot}K(\gamma)$ in (3.18a) is always bounded, namely, $\|z{\cdot}K(\gamma)\|_{L^\infty(Q)} \le \max_{\ell=1,\dots,L}|z_\ell|$, and then (3.22a) follows by usual $W^{2,r}$-regularity of the Δ-operator with (2.8) for any $r<+\infty$; cf. [1, 2]. Then also the driving force $q\nabla\phi=(z\cdot K(\gamma))\nabla\phi$ in (3.18b) is bounded in $L^\infty(Q;\mathbb{R}^3)$, and hence certainly in $L^1(I;L^2(\Omega;\mathbb{R}^3))$. Then, by a test of (3.13) by v itself and by using the Korn inequality

$$\exists\eta_2>0\quad \forall v\in W^{1,p}_0(\Omega;\mathbb{R}^3):\qquad \eta_2\big\|v\big\|_{W^{1,p}_0(\Omega;\mathbb{R}^3)}\le\big\|\mathrm{D}v\big\|_{L^p(\Omega;\mathbb{R}^{3\times3})} \tag{3.23}$$

with $\eta_2>0$ depending on the Lipschitz domain Ω, and by using the usual trick that $\int_\Omega\nabla\pi\cdot v\,\mathrm{d}x=-\int_\Omega\pi\,\mathrm{div}(v)\,\mathrm{d}x=0$ as well as $\int_\Omega(v\otimes v){:}\nabla v\,\mathrm{d}x=0$, and by using also (3.9b), we obtain the estimate

$$\begin{aligned}\frac12\frac{\mathrm{d}}{\mathrm{d}t}\big\|v\big\|^2_{L^2(\Omega;\mathbb{R}^3)}&+\eta_1\eta_2^p\big\|v\big\|^p_{W^{1,p}_0(\Omega;\mathbb{R}^3)}+\epsilon\big\|\nabla^m v\big\|^2_{L^2(\Omega;\mathbb{R}^{3^{k+1}})}\\ &\le\int_\Omega\frac{\partial v}{\partial t}\cdot v+\tau(\mathrm{D}v,\gamma,\theta){:}\mathrm{D}v+\epsilon\nabla^m v\,\vdots\,\nabla^m v\ \mathrm{d}x\\ &=-\int_\Omega q\nabla\phi{\cdot}v\ \mathrm{d}x\le\max_{\ell=1,\dots,L}|z_\ell|\ \big\|\nabla\phi\big\|_{L^2(\Omega;\mathbb{R}^3)}\big\|v\big\|_{L^2(\Omega;\mathbb{R}^3)};\end{aligned} \tag{3.24}$$

let us recall the convention pronounced in Remark 3.2 so that $\tau(\mathrm{D}v,\gamma,\theta)$ behaves well even if some γ's are negative. By Young's and Gronwall's inequalities, we obtain (3.22b) and

$$\big\|v\big\|_{L^2(I;W^{k,2}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))}\le\frac{C_3}{\sqrt{\epsilon}} \tag{3.25}$$

by a test of (3.13) by v itself and using the usual trick that $\int_\Omega\nabla\pi{\cdot}v\,\mathrm{d}x=-\int_\Omega\pi\,\mathrm{div}(v)\,\mathrm{d}x=0$ as well as $\int_\Omega(v\otimes v){:}\nabla v\,\mathrm{d}x=0$. Note that, because of the retract K used in (3.18a), the bounds in (3.22b) and (3.25) are completely independent of γ.

The estimate (3.22d) can be obtained by testing (3.18b) by $w\in L^p(I;W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))$ as follows:

$$\begin{aligned}&\Big\|\frac{\partial v}{\partial t}+(-1)^k\epsilon\Delta^k v\Big\|_{L^{p/(p-1)}(I;W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)^*)}\\ &\quad:=\sup_{\|w\|_{L^p(I;W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))}\le1}\Big\langle\frac{\partial v}{\partial t}+(-1)^k\epsilon\Delta^k v,w\Big\rangle\\ &\quad=\sup_{\|w\|_{L^p(I;W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))}\le1}\int_Q\tau(\mathrm{D}v,\gamma,\vartheta){:}\mathrm{D}w-(v\otimes v){:}\nabla w+q\,\nabla\phi{\cdot}w\,\mathrm{d}x\,\mathrm{d}t.\end{aligned} \tag{3.26}$$

The boundedness of $\int_Q(v\otimes v){:}\nabla w\ \mathrm{d}x\,\mathrm{d}t$ just requires $p\ge11/5$ because, by interpolation, (3.22b) guarantees v bounded in $L^{11/3}(Q;\mathbb{R}^3)$ so that $(v\otimes v){:}\nabla w\in L^1(Q)$ if $\nabla w\in L^{11/5}(Q;\mathbb{R}^3)$; see also, e.g., [21, Chap. 5, Lemma 2.44(iii)].

To get (3.22e), we test (3.18d) by c. We realize that

$$\int_\Omega c_\ell v{\cdot}\nabla c_\ell\ \mathrm{d}x=\frac12\int_\Omega v{\cdot}\nabla c_\ell^2\ \mathrm{d}x=-\frac12\int_\Omega(\mathrm{div}\,v)c_\ell^2\ \mathrm{d}x=0 \tag{3.27}$$

for each $\ell = 1, \dots, L$; i.e., $\int_\Omega (c\otimes v){:}\nabla c \,\mathrm{d}x = 0$. By (3.27), we obtain

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega |c|^2\,\mathrm{d}x + \eta_0\int_\Omega |\nabla c|^2\,\mathrm{d}x &\le \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega |c|^2\,\mathrm{d}x + \int_\Omega \nabla c^\top \mathfrak{D}(\gamma,\vartheta)\nabla c\,\mathrm{d}x\\ &= \int_\Omega (c\otimes v){:}\nabla c - \big(\mathfrak{m}(\gamma,\vartheta)\otimes\nabla\phi\big){:}\nabla c + r(\gamma,\vartheta)\,\mathrm{d}x\\ &\le \frac{\eta_0}{2}\|\nabla c\|^2_{L^2(\Omega;\mathbb{R}^{3\times L})} + \frac{1}{2\eta_0}\|\mathfrak{m}\|_{L^\infty(\mathbb{R}^2;\mathbb{R}^L)}\|\nabla\phi\|^2_{L^2(\Omega;\mathbb{R}^3)} + \|r\|_{L^1(\mathbb{R}^2;\mathbb{R}^L)}.\end{aligned}$$

Then (3.22e) follows by Young's and Gronwall's inequalities when using also (3.11) and (3.22a). Again, Remark 3.2 applies, of course.

As for (3.22f), let us realize that, by (3.22e), c is bounded in $L^\infty(I;L^2(\Omega;\mathbb{R}^L))$ and, by (3.25), $\sqrt{\epsilon}v$ is bounded in $L^2(I;L^\infty(\Omega;\mathbb{R}^3))$ so that $\sqrt{\epsilon}\,c\otimes v$ is certainly bounded in $L^2(Q;\mathbb{R}^{3\times L})$. Then we obtain (3.22f) by testing (3.18d) by an arbitrary w from $L^2(I;W^{1,2}(\Omega;\mathbb{R}^L))$ when using (3.11) and (3.22a).

The estimate (3.22g) can be obtained in the same way as (3.22d) by testing (3.18d) by arbitrary $w \in L^r(I;W^{1,r}(\Omega;\mathbb{R}^L))$ with a suitable r. The resulting term $\int_Q (c\otimes v){:}\nabla w \,\mathrm{d}x\,\mathrm{d}t$ is now to be estimated as

$$\int_Q (c\otimes v) : \nabla w\,\mathrm{d}x\,\mathrm{d}t \le C\|v\|_{L^{5p/3}(Q;\mathbb{R}^3)}\|c\|_{L^{10/3}(Q;\mathbb{R}^L)}\|\nabla w\|_{L^r(Q;\mathbb{R}^{3\times L})}, \tag{3.28}$$

provided that $r \ge 10p/(7p-6)$. The other resulting term, $\int_Q \nabla w^\top\mathfrak{D}(\gamma,\vartheta)\nabla c\,\mathrm{d}x\,\mathrm{d}t$, requires $r \ge 2$, which eventually gives the restriction (3.22g) on r.

We now want to show boundedness of ∇v in $L^{2p}(Q;\mathbb{R}^{3\times 3})$, which will guarantee the dissipative heat $\tau(\mathrm{D}v,\gamma,\vartheta){:}\mathrm{D}v$ bounded in $L^2(Q)$ to allow for a test of (3.18e) by $\frac{\partial\theta}{\partial t}$. We get it by the Gagliardo–Nirenberg inequality $\|w\|_{W^{1,2p}(\Omega)} \le C\|w\|^\lambda_{W^{k,2}(\Omega)}\|w\|^{1-\lambda}_{L^2(\Omega)}$, which holds for $1/(2p) + \lambda k/3 \ge 5/6$. To also have the interpolation $\|w\|_{L^{2p}(I)} \le C\|w\|^\lambda_{L^2(I)}\|w\|^{1-\lambda}_{L^\infty(I)}$, which holds for $0 \le \lambda \le 1/p$, we put $\lambda = 1/p$. Choosing k large enough, namely, as specified in (3.21), we obtain the desired interpolation

$$\|v\|_{L^{2p}(I;W^{1,2p}(\Omega;\mathbb{R}^3))} \le \|v\|^\lambda_{L^2(I;W^{k,2}(\Omega;\mathbb{R}^3))}\|v\|^{1-\lambda}_{L^\infty(I;L^2(\Omega;\mathbb{R}^3))} = \mathcal{O}\left(\frac{1}{\epsilon^{\frac{\lambda}{2}}}\right) = \mathcal{O}\left(\frac{1}{\epsilon^{\frac{1}{2p}}}\right), \tag{3.29}$$

where the order with respect to the parameter ϵ comes from (3.22b) and (3.25). Thus

$$\|\tau(\mathrm{D}v,\gamma,\vartheta){:}\mathrm{D}v\|_{L^2(Q)} = \mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right). \tag{3.30}$$

The other terms on the right-hand side of (3.18e) are bounded in $L^2(Q)$, too; note that $\nabla c_\ell\cdot\nabla\phi \in L^2(Q)$ because $\nabla c_\ell \in L^2(Q;\mathbb{R}^3)$ due to (3.22e), while $\nabla\phi \in L^\infty(Q;\mathbb{R}^3)$ due to (3.22a). Hence the total right-hand side of (3.18e), let us denote it by h_{tot}, is bounded in $L^2(Q)$. Then the test of (3.18e) by θ gives (3.22h) with the order $\mathcal{O}(1/\sqrt{\epsilon})$ coming from (3.30), the constant C_8 being still independent of γ and ϑ.

Now we simultaneously test (3.18b) by $\frac{\partial v}{\partial t}$ and (3.18e) by $\frac{\partial\theta}{\partial t}$. (Rigorously, this step is not legal unless we have L^2-information about $\frac{\partial v}{\partial t}$ and $\frac{\partial\theta}{\partial t}$ which we want just to derive, but one can, for a moment, imagine, e.g., a Galerkin approximation of (3.18b) and (3.18e) to make these tests and a subsequent limit passage.) We sum them to obtain, for a.a. $t \in I$,

$$
\begin{aligned}
(3.31)\quad & \Big\|\frac{\partial v}{\partial t}\Big\|^2_{L^2(\Omega;\mathbb{R}^3)} + c_{\rm v}\Big\|\frac{\partial \theta}{\partial t}\Big\|^2_{L^2(\Omega)} + \frac{\rm d}{{\rm d}t}\Big(\frac{\epsilon}{2}\big\|\nabla^k v\big\|^2_{L^2(\Omega;\mathbb{R}^{3^{k+1}})} + \frac{\kappa}{2}\big\|\nabla\theta\big\|^2_{L^2(\Omega;\mathbb{R}^3)}\Big) \\
&= \int_\Omega {\rm div}\big(\tau({\rm D}v,\gamma,\vartheta)\big)\cdot\frac{\partial v}{\partial t} - (v\cdot\nabla)v\cdot\frac{\partial v}{\partial t} + q\nabla\phi\cdot\frac{\partial v}{\partial t} - c_{\rm v}(v\cdot\nabla)\theta\frac{\partial\theta}{\partial t} + h_{\rm tot}\frac{\partial\theta}{\partial t} \\
&\le 2\big\|{\rm div}\big(\tau({\rm D}v,\gamma,\vartheta)\big)\big\|^2_{L^2(\Omega;\mathbb{R}^3)} + 2\big\|(v\cdot\nabla)v\big\|^2_{L^2(\Omega;\mathbb{R}^3)} \\
&\quad + 2\big\|q\nabla\phi\big\|^2_{L^2(\Omega;\mathbb{R}^3)} + \frac{1}{2}\Big\|\frac{\partial v}{\partial t}\Big\|^2_{L^2(\Omega;\mathbb{R}^3)} \\
&\quad + \frac{1}{c_{\rm v}}\big\|(v\cdot\nabla)\theta\big\|^2_{L^2(\Omega)} + \frac{1}{c_{\rm v}}\big\|h_{\rm tot}\big\|^2_{L^2(\Omega)} + \frac{c_{\rm v}}{2}\Big\|\frac{\partial\theta}{\partial t}\Big\|^2_{L^2(\Omega)}.
\end{aligned}
$$

The particular right-hand side terms can be estimated as follows. The first one allows for the estimate

$$
\begin{aligned}
(3.32)\ & \big\|{\rm div}\big(\tau({\rm D}v,\gamma,\vartheta)\big)\big\|^2_{L^2(\Omega;\mathbb{R}^3)} \le \big\|\tau_D'({\rm D}v,\gamma,\vartheta)\big\|^2_{L^2(\Omega;\mathbb{R}^{3\times3\times3})}\big\|\nabla^2 v\big\|^2_{L^\infty(\Omega;\mathbb{R}^{3^{k+1}})} \\
&\quad + \big\|\tau_\gamma'({\rm D}v,\gamma,\vartheta)\big\|^2_{L^\infty(\Omega;\mathbb{R}^{3\times3})}\big\|\nabla\gamma\big\|^2_{L^2(\Omega;\mathbb{R}^3)} \\
&\quad + \big\|\tau_\theta'({\rm D}v,\gamma,\vartheta)\big\|^2_{L^\infty(\Omega;\mathbb{R}^{3\times3})}\big\|\nabla\vartheta\big\|^2_{L^2(\Omega;\mathbb{R}^3)} \\
&\le \tilde C\Big(1+\big\|\nabla v\big\|^{2p}_{L^{2p}(\Omega;\mathbb{R}^{3\times3})}\Big)\big\|\nabla^k v\big\|^2_{L^2(\Omega;\mathbb{R}^{3^{k+1}})} + \tilde C\big\|\nabla\gamma\big\|^2_{L^2(\Omega;\mathbb{R}^3)} + \tilde C\big\|\nabla\vartheta\big\|^2_{L^2(\Omega;\mathbb{R}^3)},
\end{aligned}
$$

where (3.9e) and the embedding $W^{k,2}(\Omega) \subset W^{2,\infty}(\Omega)$ have been used as well as the growth condition (3.9d) to estimate

$$
\begin{aligned}
\big\|\tau_D'({\rm D}v,\gamma,\vartheta)\big\|^2_{L^2(\Omega;\mathbb{R}^{3\times3\times3})} &\le \int_\Omega C^2\big(1+|\nabla v|^{p-2}\big)^2\,{\rm d}x \\
&\le 2C^2\Big({\rm meas}(\Omega) + \int_\Omega |\nabla v|^{2p-4}\,{\rm d}x\Big) \le \tilde C\Big(1+\int_\Omega |\nabla v|^{2p}\,{\rm d}x\Big)
\end{aligned}
$$

with C from (3.9d). In view of (3.22b), this term can then be handled by Gronwall's inequality because, due to (3.29), $t \mapsto \|\nabla v(t,\cdot)\|^{2p}_{L^{2p}(\Omega;\mathbb{R}^{3\times3})}$ is integrable; note that (3.29) implies that the $L^1(0,T)$-norm of this function is of the order $\mathcal{O}(1/\epsilon)$, which gives the factors ${\rm e}^{C_0/\epsilon}$ in (3.22c), (3.22i). The second term in the right-hand side of (3.31) can be estimated as

$$
\big\|(v\cdot\nabla)v\big\|^2_{L^2(\Omega;\mathbb{R}^3)} \le \big\|v\big\|^2_{L^2(\Omega;\mathbb{R}^3)}\big\|\nabla v\big\|^2_{L^\infty(\Omega;\mathbb{R}^{3\times3})} \le C_2^2N^2\big\|\nabla^k v\big\|^2_{L^2(\Omega;\mathbb{R}^{3^{k+1}})},
$$

where C_2 is from (3.22b) and N is the norm of the embedding $W^{k,2}(\Omega) \subset W^{1,\infty}(\Omega)$, so that this term also can be handled by Gronwall's inequality. The term $q\nabla\phi$ is already estimated in (3.22a). The term $(v\cdot\nabla)\theta$ is to be estimated as

$$
\big\|(v\cdot\nabla)\theta\big\|^2_{L^2(\Omega)} \le \big\|v\big\|^2_{L^\infty(\Omega;\mathbb{R}^3)}\big\|\nabla\theta\big\|^2_{L^2(\Omega;\mathbb{R}^3)} \le N^2\big\|\nabla^k v\big\|^2_{L^2(\Omega;\mathbb{R}^{3^{k+1}})}\big\|\nabla\theta\big\|^2_{L^2(\Omega;\mathbb{R}^3)},
$$

where N is the norm of the embedding $W^{k,2}(\Omega) \subset L^\infty(\Omega)$, and again we can treat it by Gronwall's inequality if (3.25) is taken into account. The boundedness of $h_{\rm tot}$ in $L^2(Q)$ has already been mentioned. Therefore, (3.31) yields both (3.22c) and (3.22i); note that (3.25) gives $\|\nabla^k v\|^2_{L^2(\Omega;\mathbb{R}^{3^{k+1}})} = \mathcal{O}(1/\epsilon)$ and (3.30) gives $\|h_{\rm tot}\|^2_{L^2(\Omega)} = \mathcal{O}(1/\epsilon)$, which (together with the already mentioned factor $\mathrm{e}^{C_0/\epsilon}$) eventually determines the order both in (3.22c) and in (3.22i).

Having ∇v bounded in $L^p(Q;\mathbb{R}^{3\times3})$) (see (3.22b)), $\tau(\mathrm{D}v,\gamma,\vartheta){:}\mathrm{D}v$ is then certainly bounded in $L^1(Q)$ (independently of ϵ) while the other right-hand side terms of (3.18e) are bounded in this space, too, because of (3.22a), (3.22e), and (3.11). This allows us to use fine results about integrability of temperature gradient [6, 7, 8] modified for the initial-boundary-value problem

$$c_\mathrm{v}\frac{\partial\theta}{\partial t}-\mathrm{div}\big(\kappa\nabla\theta-c_\mathrm{v}v\,\theta\big)=h_\mathrm{tot}\ \text{on } Q,\qquad \kappa\frac{\partial\theta}{\partial\nu}=0\ \text{on } \Sigma,\qquad \theta(0,\cdot)=\theta_0\ \text{on } \Omega; \tag{3.33}$$

recall that $h_\mathrm{tot}\in L^1(Q)$ denotes the total right-hand side of (3.18e). First, let us test (3.33) by $\mathrm{sign}(\theta)$ or, more rigorously, by a regularization of it, say $\max(-1,\min(1,n\theta))$, and then make a limit passage with $n\to\infty$, which gives the first part of the estimate (3.22j), i.e., a bound for θ in $L^\infty(I;L^1(\Omega))$; for more details about this rather standard technique, see, e.g., [35, sect. 9.4 with sect. 3.2.3]. The second part of (3.22j) is more involved. Following [7, 8], we test (3.33) by $\psi_n(\theta)$ with $\psi_n:\mathbb{R}\to[-1,1]$ a bounded Lipschitz function defined, for $n\in\mathbb{N}$, by

$$\psi_n(\theta):=\begin{cases}0 & \text{if } |\theta|\le n,\\ \mathrm{sign}(\theta)(|\theta|-n) & \text{if } n\le|\theta|\le n+1,\\ \mathrm{sign}(\theta) & \text{if } |\theta|\ge n+1.\end{cases} \tag{3.34}$$

We use

$$\begin{aligned}\int_\Omega\theta\,v\cdot\nabla\psi_n(\theta)\,\mathrm{d}x&=\int_\Omega\theta\,\psi_n'(\theta)\,v\cdot\nabla\theta\,\mathrm{d}x\\&=\int_\Omega v\cdot\nabla\hat\phi_n(\theta)\,\mathrm{d}x=-\int_\Omega\mathrm{div}(v)\hat\phi_n(\theta)\,\mathrm{d}x=0,\end{aligned} \tag{3.35}$$

where $\hat\phi_n:\mathbb{R}\to\mathbb{R}$ denotes a primitive function of $\phi_n:\theta\mapsto\theta\,\psi_n'(\theta)$. Further, we denote by $\hat\psi_n$ the primitive function of ψ_n such that $\hat\psi_n(0)=0$; note that $0\le\hat\psi_n(\theta)\le|\theta|$. Testing (3.33) by $\psi_n(\theta)$ and denoting $B_n:=\{(t,x)\in Q:\ n\le|\theta(t,x)|\le n{+}1\}$ then gives

$$\begin{aligned}\kappa\int_{B_n}|\nabla\theta|^2\,\mathrm{d}x\,\mathrm{d}t&=\kappa\int_Q\psi_n'(\theta)|\nabla\theta|^2\,\mathrm{d}x\,\mathrm{d}t=\int_Q\kappa\nabla\theta{\cdot}\nabla\psi_n(\theta)\,\mathrm{d}x\,\mathrm{d}t\\&\le\int_Q\kappa\nabla\theta\cdot\nabla\psi_n(\theta)\,\mathrm{d}x\,\mathrm{d}t+\int_\Omega c_\mathrm{v}\hat\psi_n(\theta(T,\cdot))\,\mathrm{d}x\\&=\int_\Omega c_\mathrm{v}\hat\psi_n(\theta_{0\epsilon})\,\mathrm{d}x+\int_Q h_\mathrm{tot}\psi_n(\theta)\,\mathrm{d}x\,\mathrm{d}t\\&\le c_\mathrm{v}\big\|\theta_{0\epsilon}\big\|_{L^1(\Omega)}+\big\|h_\mathrm{tot}\big\|_{L^1(Q)}.\end{aligned} \tag{3.36}$$

For $\mu>0$ fixed, we get

$$\begin{aligned}\int_Q\frac{|\nabla\theta|^2}{(1+\theta)^{1+\mu}}\,\mathrm{d}x\,\mathrm{d}t&=\sum_{n=0}^\infty\int_{B_n}\frac{|\nabla\theta|^2}{(1+\theta)^{1+\mu}}\,\mathrm{d}x\,\mathrm{d}t\\&\le\sum_{n=0}^\infty\frac{1}{(1+n)^{1+\mu}}\int_{B_n}|\nabla\theta|^2\,\mathrm{d}x\,\mathrm{d}t\\&\le\frac{c_\mathrm{v}\big\|\theta_{0\epsilon}\big\|_{L^1(\Omega)}+\big\|h_\mathrm{tot}\big\|_{L^1(Q)}}{\kappa}\sum_{n=0}^\infty\frac{1}{(1+n)^{1+\mu}}\le C_\mu\end{aligned} \tag{3.37}$$

with some C_μ. Further, we simplify [7, 8] which estimate $\nabla\theta$ in an anisotropic space. For our purposes, an estimate of $\nabla\theta$ in an "isotropic" space $L^\zeta(Q;\mathbb{R}^3)$ will suffice. For this, let us take $1 \le \zeta < 2$. By Hölder's inequality,

$$
\begin{aligned}
(3.38)\qquad \int_Q |\nabla\theta|^\zeta \,\mathrm{d}x\,\mathrm{d}t &= \int_Q \frac{|\nabla\theta|^\zeta}{(1+\theta)^{(1+\mu)\zeta/2}}(1+\theta)^{(1+\mu)\zeta/2}\,\mathrm{d}x\,\mathrm{d}t \\
&\le \left(\int_Q \frac{|\nabla\theta|^2}{(1+\theta)^{1+\mu}}\,\mathrm{d}x\,\mathrm{d}t\right)^{\zeta/2}\left(\int_Q (1+\theta)^{(1+\mu)\zeta/(2-\zeta)}\,\mathrm{d}x\,\mathrm{d}t\right)^{(2-\zeta)/2} \\
&\le C_\mu^{\zeta/2}\left(\int_0^T \|1+\theta(t,\cdot)\|_{L^{(1+\mu)\zeta/(2-\zeta)}(\Omega)}^{(1+\mu)\zeta/(2-\zeta)}\,\mathrm{d}t\right)^{(2-\zeta)/2}.
\end{aligned}
$$

The proven first part of (3.22j), i.e., $\|\theta\|_{L^\infty(I;L^1(\Omega))} \le C_{10}$, allows us further to estimate, by using the Gagliardo–Nirenberg inequality,

$$
\begin{aligned}
(3.39)\qquad \|1+\theta(t,\cdot)\|_{L^{(1+\mu)\zeta/(2-\zeta)}(\Omega)} &\le C_{\mathrm{GN}}\|\nabla\theta(t,\cdot)\|_{L^\zeta(\Omega)}^\lambda \|1+\theta(t,\cdot)\|_{L^1(\Omega)}^{1-\lambda} \\
&\le C_{\mathrm{GN}}\big(|\Omega|+C_{10}\big)^{1-\lambda}\|\nabla\theta(t,\cdot)\|_{L^\zeta(\Omega)}^\lambda
\end{aligned}
$$

for some $C_{\mathrm{GN}} \in \mathbb{R}$, provided that

$$
(3.40)\qquad \frac{2-\zeta}{(1+\mu)\zeta} \ge \lambda\left(\frac{1}{\zeta}-\frac{1}{3}\right)+1-\lambda.
$$

We raise (3.39) to the power $(1+\mu)\zeta/(2-\zeta)$, exploit it for (3.38), and choose $\lambda := (2-\zeta)/(1+\mu)$, which yields

$$
\begin{aligned}
(3.41)\qquad &\left(\int_0^T \|1+\theta(t,\cdot)\|_{L^{(1+\mu)\zeta/(2-\zeta)}(\Omega)}^{(1+\mu)\zeta/(2-\zeta)}\,\mathrm{d}t\right)^{(2-\zeta)/2} \\
&\le \left(\int_0^T C_{\mathrm{GN}}^{\frac{(1+\mu)\zeta}{2-\zeta}}\big(|\Omega|+C_{10}\big)^{\frac{(1-\lambda)(1+\mu)\zeta}{2-\zeta}}\|\nabla\theta(t,\cdot)\|_{L^\zeta(\Omega)}^{\frac{\lambda(1+\mu)\zeta}{2-\zeta}}\,\mathrm{d}t\right)^{\frac{2-\zeta}{2}} \\
&\le \left(\int_0^T C_{\mathrm{GN}}^{\frac{(1+\mu)\zeta}{2-\zeta}}\big(|\Omega|+C_{10}\big)^{\frac{(1-\lambda)(1+\mu)\zeta}{2-\zeta}}\|\nabla\theta(t,\cdot)\|_{L^\zeta(\Omega)}^{\zeta}\,\mathrm{d}t\right)^{\frac{2-\zeta}{2}} \\
&= C_{\mathrm{GN}}^{(1+\mu)\zeta/2}\big(|\Omega|+C_{10}\big)^{\zeta(\zeta-1+\mu)/2}\left(\int_Q |\nabla\theta|^\zeta\,\mathrm{d}x\,\mathrm{d}t\right)^{(2-\zeta)/2}.
\end{aligned}
$$

Merging (3.38) with (3.41) gives the estimate

$$
(3.42)\qquad \|\nabla u\|_{L^\zeta(Q;\mathbb{R}^3)}^\zeta \le C_\mu C_{\mathrm{GN}}^{1+\mu}\big(|\Omega|+C_{10}\big)^{\zeta-1+\mu}.
$$

Putting our choice of $\lambda := (2-\zeta)/(1+\mu)$ into (3.40), one obtains, after some algebra, the conditions $\zeta \le (5-3\mu)/4$ so that (3.42) gives just the second part of the estimate (3.22j) with $\xi := \frac{3}{4}\mu$.

To prove (3.22k), we must, in particular, estimate the term $\mathrm{div}(v\theta)$ in the following way:

$$\begin{aligned}(3.43)\qquad \big\|\mathrm{div}(v\theta)\big\|_{L^1(I;W^{-3,2}(\Omega))} &= \sup_{\|w\|_{L^\infty(I;W_0^{3,2}(\Omega))}\le 1} \int_Q v\theta\cdot\nabla w \,\mathrm{d}x\,\mathrm{d}t \\ &\le \sup_{\|w\|_{L^\infty(I;W_0^{3,2}(\Omega))}\le 1} C\big\|v\big\|_{L^{5p/3}(Q;\mathbb{R}^3)}\big\|\theta\big\|_{L^{5/3-\delta}(Q)}\big\|\nabla w\|_{L^\infty(Q;\mathbb{R}^3)} \\ &= C\big\|v\big\|_{L^{5p/3}(Q;\mathbb{R}^3)}\big\|\theta\big\|_{L^{5/3-\delta}(Q)}\end{aligned}$$

with a sufficiently small $\delta > 0$ and with a suitable constant C, where we used the embedding $W^{3,2}(\Omega) \subset W^{1,\infty}(\Omega)$ and, by the Gagliardo–Nirenberg inequality, also the embedding

$$(3.44)\qquad L^p(I;W^{1,p}(\Omega)) \cap L^\infty(I;L^2(\Omega)) \subset L^{5p/3}(Q)$$

(cf. [11, sect. I.3]), and finally also the embedding

$$(3.45)\qquad L^\infty(I;L^1(\Omega)) \,\cap\, L^{5/4-\xi}(I;W^{1,5/4-\xi}(\Omega)) \subset L^{5/3-\delta}(Q),$$

again by the Gagliardo–Nirenberg inequality. Alternatively, we could use here Sobolev embeddings and usual interpolation of Lebesgue spaces; note that (3.43) works even for $p > 3/2$.

Eventually, we prove (3.18f). Let us abbreviate $\sigma := \sum_{\ell=1}^L c_\ell = 1$. By summing (3.18e) for $\ell = 1,\dots,L$ and by (3.6a) and (3.7c,d), one gets

$$\begin{aligned}(3.46)\qquad \frac{\partial\sigma}{\partial t} &= \sum_{\ell=1}^L r_\ell\big(\gamma,\vartheta\big) + \mathrm{div}\left(\sum_{\ell=1}^L\sum_{k=1}^L \mathfrak{D}_{k\ell}(\gamma,\vartheta)\nabla c_k + \mathfrak{m}_\ell(\gamma,\vartheta)\nabla\phi - vc_\ell\right) \\ &= 0 + \mathrm{div}\left(\beta\sum_{k=1}^L\nabla c_k + \left(\sum_{\ell=1}^L \mathfrak{m}_\ell(\gamma,\vartheta)\right)\nabla\phi - v\left(\sum_{\ell=1}^L c_\ell\right)\right) \\ &= \mathrm{div}(\beta\nabla\sigma) - v\cdot\nabla\sigma.\end{aligned}$$

Due to (3.6b) and (2.8b), a solution to the thus obtained initial-boundary-value problem for a parabolic (if $\beta > 0$) or hyperbolic (if $\beta = 0$) equation, i.e.,

$$(3.47)\qquad \frac{\partial\sigma}{\partial t} - \mathrm{div}(\beta\nabla\sigma) + v\cdot\nabla\sigma = 0 \text{ on } Q, \qquad \frac{\partial\sigma}{\partial\nu} = 0 \text{ on } \Sigma, \qquad \sigma(0,\cdot) = 1 \text{ on } \Omega,$$

is $\sigma \equiv 1$. This solution is unique, which can be proved by testing the difference of (3.47) for two solutions σ_1 and σ_2 by $\sigma_1 - \sigma_2$. The important fact is that the resulting term $\int_\Omega (v\cdot\nabla(\sigma_1-\sigma_2))(\sigma_1-\sigma_2)\,\mathrm{d}x$ vanishes as in (3.27); note that our estimates (3.25) and (3.22e) ensure integrability of all integrands occurring in (3.27) with $\sigma_1-\sigma_2$ in place of c_ℓ. ∎

Remark 3.7. The parabolic/hyperbolic equation (3.47) can be found in the literature in this context; cf. [17, sect. 7.3.5].

PROPOSITION 3.8 (continuity). *Let the assumptions of Proposition* 3.6 *hold. Then the weak solution to* (3.18)–(3.19) *with the boundary conditions* (2.8) *and* (3.14) *is determined uniquely, and the mapping*

$$(3.48)\qquad (\gamma,\vartheta) \mapsto \big\{(v,c,\theta,\phi) \textit{ is a weak solution to } (3.18),(3.19),(2.8),(3.14)\big\}$$

with $\sum_{\ell=1}^{L}\gamma_\ell=1$ *is continuous from the weak topology on* $\mathcal{W}^L\times\mathcal{W}$ *with*

$$\mathcal{W}:=L^2(I;W^{1,2}(\Omega))\cap W^{1,2}(I;W^{1,2}(\Omega)^*) \tag{3.49}$$

to the weak topology related to the spaces from the estimates* (3.22a), (3.22c), (3.22f), (3.22i).

Proof. The uniqueness of the solution to (3.18a) follows in a standard way because of linearity and because $\varepsilon>0$ and $\alpha>0$ are assumed. As for (3.18b), (3.18c), the uniqueness is due to the monotonicity (3.9b) of $\tau(\cdot,\gamma,\vartheta)$ and because the term $\mathrm{div}(v\otimes v)$ can be estimated on the right-hand side: indeed, considering two solutions v_1 and v_2, by the difference of the weak formulations of (3.18b), (3.18c) for v_1 and v_2 tested by v_1-v_2 and using Green's formula several times, we obtain

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\big\|v_1-v_2\big\|^2_{L^2(\Omega;\mathbb{R}^3)}&+\epsilon\big\|\nabla^k v_1-\nabla^k v_2\big\|^2_{L^2(\Omega;\mathbb{R}^{3^{k+1}})}\\ &\le\int_\Omega\big(v_1\otimes v_1-v_2\otimes v_2\big):\nabla(v_1-v_2)\,\mathrm{d}x\\ &=\int_\Omega\big((v_1\cdot\nabla)v_1-(v_2\cdot\nabla)v_2\big)\cdot(v_1-v_2)\,\mathrm{d}x\\ &=\int_\Omega\big(((v_1-v_2)\cdot\nabla)v_1\big)\cdot(v_1-v_2)\,\mathrm{d}x\le\big\|\nabla v_1\big\|_{L^\infty(\Omega;\mathbb{R}^{3\times3})}\big\|v_1-v_2\big\|^2_{L^2(\Omega;\mathbb{R}^3)}\end{aligned} \tag{3.50}$$

from which $v_1=v_2$ follows by Gronwall's inequality when taking into account also the estimates $\|\nabla v_1\|_{L^\infty(\Omega;\mathbb{R}^{3\times3})}\le N\|\nabla^k v_1\|_{L^2(\Omega;\mathbb{R}^{3^{k+1}})}$ and (3.22c). The uniqueness of solutions to (3.18d), (3.18e) then follows in a standard way because these equations are decoupled and linear and all time derivatives are in duality with the corresponding solutions.

Take a sequence $\{(\gamma_n,\vartheta_n)\}_{n\in\mathbb{N}}$ converging weakly to some (γ,ϑ) in $\mathcal{W}^L\times\mathcal{W}$. Take the corresponding $(v_n,c_n,\theta_n,\phi_n)$ and choose a subsequence converging weakly* in the spaces specified in the estimates (3.22). By the Aubin–Lions compact-embedding theorem [4, 19] (see also, e.g., [35, Lemma 7.7]), the estimates (3.22e) and (3.22f) imply that

$$\gamma_n\to\gamma\ \text{ in }\ L^2(I;L^{6-\xi}(\Omega;\mathbb{R}^L)) \tag{3.51}$$

in the norm topology with any $\xi>0$. This allows us to pass to the limit $K(\gamma_n)\to K(\gamma)$ and also ensures $\nabla\phi_n\to\nabla\phi$ strongly in $L^r(Q;\mathbb{R}^3)$ for any $r<+\infty$ to be exploited for (3.18e). Using again the Aubin–Lions theorem, we obtain $\vartheta_n\to\vartheta$ strongly in $L^2(I;L^{6-\xi}(\Omega))$, which allows us to pass to the limit $h(\gamma_n,\vartheta_n)\to h(\gamma,\vartheta)$ and $r_\ell(\gamma_n,\vartheta_n)\to r_\ell(\gamma,\vartheta)$. Moreover, again by the Aubin–Lions theorem and by interpolation as in (3.29) in the proof of Proposition 3.6,

$$\nabla v_n\to\nabla v\ \text{ in }\ L^{2p}(Q;\mathbb{R}^3) \tag{3.52}$$

in the norm topology; hence

$$\tau(\mathrm{D}v_n,\gamma_n,\vartheta_n){:}\mathrm{D}v_n\to\tau(\mathrm{D}v,\gamma,\vartheta){:}\mathrm{D}v\ \text{ in }\ L^2(Q), \tag{3.53}$$

which is essential for the limit passage in (3.18e) to obtain a weak solution. For the convective term in (3.18e), let us realize that $v_n\to v$ weakly* in $L^\infty(Q;\mathbb{R}^3)$ and, due to (3.22i), $\theta_n\to\theta$ weakly in $W^{1,2}(Q)$ and hence strongly in $L^2(Q)$ just by Rellich's theorem, which easily implies $v_n\theta_n\to v\theta$ weakly in $L^2(Q;\mathbb{R}^3)\subset L^1(Q;\mathbb{R}^3)$.

The limit passage in (3.18) is then routine. The uniqueness already proved above eventually ensures the convergence of the whole sequence. ∎

PROPOSITION 3.9 (existence of a weak solution to (2.1c)–(2.1e) and (3.13)). *Let again the assumptions of Proposition* 3.6 *hold; then the mapping* $\mathcal{F}:(\gamma,\vartheta)\mapsto(c,\theta)$, *where* (c,θ) *is uniquely determined by* (3.18)*, maps the set*

$$\text{(3.54)}\qquad \mathcal{S}:=\bigg\{(c,\theta)\in\mathcal{W}^L\times\mathcal{W}:\ \big\|c\big\|_{\mathcal{W}^L}\le\max\Big(C_5,\frac{C_6}{\sqrt{\epsilon}}\Big),\\ \big\|\theta\big\|_{\mathcal{W}}\le\max\bigg(\frac{C_8}{\sqrt{\epsilon}},\frac{C_9\mathrm{e}^{C_0/\epsilon^2}}{\epsilon}\bigg),\ \ \sum_{\ell=1}^L c_\ell=1\bigg\},$$

where C_0,C_5,C_6,C_8,C_9 *are from* (3.22e), (3.22f), (3.22h), (3.22i) *with* C_0 *and* C_9 *depending on* C_5 *and* C_8*, into itself and has a fixed point* $(c,\theta)\in\mathcal{S}$*. Moreover, every such fixed point also satisfies* $c_\ell\ge 0$ *for each* $\ell=1,\dots,L$*, and, considering the corresponding* ϕ *and* v*, the quadruple* (v,c,θ,ϕ) *is a weak solution to* (2.1c)–(2.1e) *and* (3.13) *with* (2.8), (3.14) and (3.19).

Proof. The fact that $\mathcal{F}:\mathcal{S}\to\mathcal{S}$ follows from Proposition 3.6 because C_5, C_6, and C_8 from (3.22e), (3.22f), (3.22h) do not depend on (γ,ϑ) at all, while C_0 and C_9 from (3.22i) are fixed when C_5 and C_8 are fixed; hence $\mathcal{F}$ indeed maps $\mathcal{S}$ into itself. We use $\mathcal{S}$ equipped with the weak topology $\mathcal{W}^{L+1}$. The continuity of $\mathcal{F}$ in this topology was proved in Proposition 3.8. The fixed point then exists by Schauder's theorem (in Tikhonov's modification).

Although we cannot prove $c_\ell\ge0$ if $c\ne\gamma$, in the fixed point we have $c=\gamma$ and we can prove $c_\ell\ge0$ for each $\ell=1,\dots,L$ by testing (3.18d) by the negative part c^- with $c_\ell^-:=\min(c_\ell,0)$. It is important that c is a conventional weak solution so that $\frac{\partial}{\partial t}c_\ell$ is in duality with c_ℓ and also with $c_\ell^-\in L^2(I;W^{1,2}(\Omega))$. For any $\ell=1,\dots,L$, by (3.10a), we use

$$\text{(3.55)}\qquad \sum_{k=1}^3\mathfrak{D}_{k\ell}(c,\theta)\nabla c_k\cdot\nabla c_\ell^-=\begin{cases}\mathfrak{D}_{\ell\ell}(c,\theta)\nabla c_\ell\cdot\nabla c_\ell^-\ge0 & \text{if } c_\ell(t,x)<0,\\ 0\quad\text{ as just }\ \nabla c_\ell^-=0 & \text{if } c_\ell(t,x)\ge0,\end{cases}$$

which holds for a.a. $(t,x)\in Q$; recall that $\mathfrak{D}$ is considered as extended continuously (cf. Remark 3.2), so that (3.10a) holds for c_ℓ negative, too. For the convective term, we use

$$\text{(3.56)}\quad \int_\Omega c_\ell v\cdot\nabla c_\ell^-\,\mathrm{d}x=\int_\Omega c_\ell^- v\cdot\nabla c_\ell^-\,\mathrm{d}x=\int_\Omega v\cdot\nabla\frac{(c_\ell^-)^2}{2}\,\mathrm{d}x=\int_\Omega(-\mathrm{div}\,v)\frac{(c_\ell^-)^2}{2}\,\mathrm{d}x=0.$$

Recall also (3.10c) which allows us to consider $r_\ell(\cdot,\theta)$ extended continuously and nonnegatively for $c_\ell\le0$ (cf. Remark 3.2), so that $r_\ell(\cdot,\theta)c_\ell^-\le0$. By (3.10b), similar extension can be assumed for $\mathfrak{m}_\ell(\cdot,\theta)$ so that $\mathfrak{m}_\ell(\cdot,\theta)\nabla\phi\cdot\nabla c_\ell^-=0$ a.e. on Q.

Hence the suggested test of the Nernst–Planck equation (3.18d) in the weak formulation by $c^-(t,\cdot)$ yields

$$\text{(3.57)}\qquad \frac12\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega|c_\ell^-|^2\,\mathrm{d}x\le\frac12\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega|c_\ell^-|^2\,\mathrm{d}x+\int_\Omega\sum_{k=1}^L\mathfrak{D}_{k\ell}(c,\theta)\nabla c_k\cdot\nabla c_\ell^-\,\mathrm{d}x\\ +\int_\Omega\mathfrak{m}_\ell(c,\theta)\nabla\phi\cdot\nabla c_\ell^-\,\mathrm{d}x-\int_\Omega c_\ell v\cdot\nabla c_\ell^-\,\mathrm{d}x=\int_\Omega r_\ell(c,\theta)c_\ell^-\,\mathrm{d}x\le0$$

for a.a. $t \in (0,T)$, so that $c_\ell^- = 0$ a.e. on Q provided $c_\ell|_{t=0} \geq 0$ for any $\ell = 1,\ldots,L$, as indeed assumed in (3.6a). Therefore $c = K(c)$ and the retract K occurring in (3.18b) can eventually be "forgotten" in the fixed point. ∎

3.3. Limit passage for $\epsilon \to 0$. In this section we will make a limit passage for $\epsilon \to 0$ in the weak solution to (2.1c)–(2.1e) and (3.13), denoted in this section by $(v_\epsilon, c_\epsilon, \theta_\epsilon, \phi_\epsilon)$, whose existence was proved in Proposition 3.9. This means that $(v_\epsilon, c_\epsilon, \theta_\epsilon, \phi_\epsilon)$ together with some π_ϵ solves (in the weak sense) the system

$$\frac{\partial v_\epsilon}{\partial t} - \operatorname{div}\big(\tau(\mathrm{D}v_\epsilon, c_\epsilon, \theta_\epsilon) - v_\epsilon \otimes v_\epsilon\big) + \nabla\pi_\epsilon + (-1)^k \epsilon \Delta^k v_\epsilon = -c_\epsilon{\cdot}z\nabla\phi_\epsilon, \tag{3.58a}$$

$$\operatorname{div}(v_\epsilon) = 0, \tag{3.58b}$$

$$\frac{\partial c_\epsilon}{\partial t} - \operatorname{div}\big(\mathfrak{D}(c_\epsilon,\theta_\epsilon)\nabla c_\epsilon + \mathfrak{m}(c_\epsilon,\theta_\epsilon) \otimes \nabla\phi_\epsilon - c_\epsilon \otimes v_\epsilon\big) = r(c_\epsilon,\theta_\epsilon)\,, \tag{3.58c}$$

$$\begin{aligned} c_{\rm v}\frac{\partial \theta_\epsilon}{\partial t} - \operatorname{div}\big(\kappa\nabla\theta_\epsilon - c_{\rm v} v_\epsilon\,\theta_\epsilon\big) &= \tau(\mathrm{D}v_\epsilon, c_\epsilon, \theta_\epsilon){:}\mathrm{D}v_\epsilon \\ &\quad + \big(\mathfrak{D}(c_\epsilon,\theta_\epsilon)\nabla c_\epsilon + \mathfrak{m}(c_\epsilon,\theta_\epsilon) \otimes \nabla\phi_\epsilon\big){:}(z \otimes \nabla\phi_\epsilon) + h(c_\epsilon,\theta_\epsilon), \end{aligned} \tag{3.58d}$$

$$-\operatorname{div}(\varepsilon\nabla\phi_\epsilon) = c_\epsilon{\cdot}z\,, \tag{3.58e}$$

$$c_\epsilon \cdot \mathbf{1} := \sum_{\ell=1}^{L} [c_\epsilon]_\ell = 1 \quad \text{ and } \quad [c_\epsilon]_\ell \geq 0 \quad \text{for } \ell = 1,\ldots,L, \tag{3.58f}$$

together with the initial conditions,

$$v_\epsilon(0,\cdot) = v_{0\epsilon}\,, \qquad c_\epsilon(0,\cdot) = c_0\,, \qquad \theta_\epsilon(0,\cdot) = \theta_{0\epsilon}, \tag{3.59}$$

and the boundary conditions on Σ,

$$\frac{\partial^l v_\epsilon}{\partial \nu^l} = 0, \qquad l = 0,\ldots,k{-}1, \tag{3.60a}$$

$$\big(\mathfrak{D}(c_\epsilon,\theta_\epsilon)\nabla c_\epsilon + \mathfrak{m}(c_\epsilon,\theta_\epsilon) \otimes \nabla\phi_\epsilon\big)\nu = 0, \tag{3.60b}$$

$$\varepsilon\frac{\partial \phi_\epsilon}{\partial \nu} = \alpha(\phi_\Sigma{-}\phi_\epsilon), \tag{3.60c}$$

$$\kappa\frac{\partial \theta_\epsilon}{\partial \nu} = 0. \tag{3.60d}$$

Proposition 3.10 (existence of a very weak solution to (2.1)–(2.3)). *Let the assumptions of Proposition* 3.6 *be satisfied, let $v_0 \in L^2_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)$, $c_0 \in L^\infty(\Omega;\mathbb{R}^L)$, $\theta_0 \in L^1(\Omega)$, and let $v_{0\epsilon} \to v_0$ in $L^2(\Omega;\mathbb{R}^3)$ and $\theta_{0\epsilon} \to \theta_0$ in $L^1(\Omega)$. Then any sequence $\{(v_\epsilon, c_\epsilon, \theta_\epsilon, \phi_\epsilon)\}_{\epsilon>0}$ of weak solutions obtained in Proposition* 3.9 *contains a subsequence converging weakly* in spaces involved in* (3.22a), (3.22b), (3.22d), (3.22e), (3.22g), (3.22j), (3.22k)*; let us denote by (v,c,θ,ϕ) its limit, and every (v,c,θ,ϕ) obtained in this way is a very weak solution due to Definition* 3.1.

Proof. We choose a subsequence that converges weakly* as claimed. Without confusion, let us denote it briefly again by $\{(v_\epsilon, c_\epsilon, \theta_\epsilon, \phi_\epsilon)\}_{\epsilon>0}$.

First, by the Aubin–Lions theorem [4, 19] and by (3.22e), (3.22g) together with the L^∞-information from Proposition 3.9, $c_\epsilon \to c$ in $L^r(Q;\mathbb{R}^L)$ for any $r < +\infty$, and from (3.58e) together with the already used $W^{2,2}$-regularity of Δ-operator, $\phi_\epsilon \to \phi$ strongly in $L^s(I;W^{2,2}(\Omega))$.

Let us prove that the weak* limit v is the very weak solution to (2.1a), (2.1b) that we seek. We use Minty's trick for the term $\operatorname{div}\tau(\mathrm{D}v,c,\theta)$ and compactness for the convective term. The important fact is that we have chosen the subsequence so that, by (3.22d), for some $\dot v \in L^{p/(p-1)}(I; W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)^*)$, we have at our disposal

$$\frac{\partial v_\epsilon}{\partial t} + (-1)^k \epsilon \Delta^k v_\epsilon \to \dot v \quad \text{weakly in } \ L^{p/(p-1)}(I; W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)^*). \tag{3.61}$$

For any w smooth with a compact support in Q, it holds that

$$\begin{aligned} \langle \dot v, w\rangle &= \lim_{\epsilon\to 0}\left\langle \frac{\partial v_\epsilon}{\partial t} + (-1)^k\epsilon\Delta^k v_\epsilon, w\right\rangle = \lim_{\epsilon\to 0}\int_Q \frac{\partial v_\epsilon}{\partial t} w + \epsilon\nabla^k v_\epsilon \,\vdots\, \nabla^k w \,\mathrm{d}x\,\mathrm{d}t \\ &= \lim_{\epsilon\to 0}\int_Q -v_\epsilon\frac{\partial w}{\partial t} + \epsilon\nabla^k v_\epsilon \,\vdots\, \nabla^k w \,\mathrm{d}x\,\mathrm{d}t = \int_Q -v\frac{\partial w}{\partial t}\,\mathrm{d}x\,\mathrm{d}t \end{aligned} \tag{3.62}$$

because $v_\epsilon \to v$ weakly in $L^p(I; W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))$ thanks to (3.22b) and also because $\|\epsilon\nabla^k v_\epsilon\|_{L^2(Q;\mathbb{R}^{3^{k+1}})} = \epsilon\mathcal{O}(1/\sqrt{\epsilon}) = \mathcal{O}(\sqrt{\epsilon}) \to 0$ due to (3.25) so that

$$\lim_{\epsilon\to 0}\int_Q \epsilon\nabla^k v_\epsilon \,\vdots\, \nabla^k w \,\mathrm{d}x\,\mathrm{d}t = 0. \tag{3.63}$$

This shows that $\dot v$ is the distributional derivative of v; let us denote it naturally as $\frac{\partial v}{\partial t}$. In particular, we have shown that

$$\frac{\partial v}{\partial t} = \dot v \in L^{p/(p-1)}(I; W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)^*). \tag{3.64}$$

Furthermore, for $w \in L^2(I; W^{k,2}_0(\Omega;\mathbb{R}^3)) \cap L^p(I; W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))$, by monotonicity of $\tau(\cdot, c_\epsilon(t,x), \theta_\epsilon(t,x))$, it holds that

$$\begin{aligned} 0 &\le \int_Q (\tau(\mathrm{D}v_\epsilon, c_\epsilon, \theta_\epsilon) - \tau(\mathrm{D}w, c_\epsilon, \theta_\epsilon)) : \mathrm{D}(v_\epsilon - w)\,\mathrm{d}x\,\mathrm{d}t \\ &= \int_Q z\cdot c_\epsilon\nabla\phi_\epsilon\cdot(v_\epsilon - w) - \frac{\partial v_\epsilon}{\partial t}\cdot(v_\epsilon - w) + (v_\epsilon\otimes v_\epsilon) : \nabla(v_\epsilon - w) \\ &\qquad - \epsilon\nabla^k v_\epsilon \,\vdots\, \nabla^k(v_\epsilon - w) - \tau(\mathrm{D}w, c_\epsilon, \theta_\epsilon) : \mathrm{D}(v_\epsilon - w)\,\mathrm{d}x\,\mathrm{d}t \\ &\le \int_Q z\cdot c_\epsilon\nabla\phi_\epsilon\cdot(v_\epsilon - w) - \frac{\partial v_\epsilon}{\partial t}\cdot(v_\epsilon - w) + (v_\epsilon\otimes v_\epsilon) : \nabla(v_\epsilon - w) \\ &\qquad + \epsilon\nabla^k v_\epsilon \,\vdots\, \nabla^k w - \tau(\mathrm{D}w, c_\epsilon, \theta_\epsilon) : \mathrm{D}(v_\epsilon - w)\,\mathrm{d}x\,\mathrm{d}t. \end{aligned} \tag{3.65}$$

Then we can bound from above the limit superior. The important fact is that $\frac{\partial v}{\partial t} \in L^{p/(p-1)}(I; W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)^*)$ is in duality to v due to (3.64) and the estimate (3.22b). First, let us realize that (3.61) implies

$$\begin{aligned} v_\epsilon(T) = v_{0\epsilon} + \int_0^T \frac{\partial v_\epsilon}{\partial t}\,\mathrm{d}t = v_{0\epsilon} + \int_0^T \left(\frac{\partial v_\epsilon}{\partial t} + (-1)^k\epsilon\Delta^k v_\epsilon\right)\mathrm{d}t \\ -(-1)^k\int_0^T \epsilon\Delta^k v_\epsilon\,\mathrm{d}t \to v_0 + \int_0^T \frac{\partial v}{\partial t}\,\mathrm{d}t = v(T) \end{aligned} \tag{3.66}$$

weakly in $W^{k,2}(\Omega;\mathbb{R}^3)^*$. Due to the estimate (3.22b), $v_\epsilon(T)$ also converges weakly in $L^2(\Omega;\mathbb{R}^3)$; hence we can conclude that even

$$v_\epsilon(T) \to v(T) \quad \text{weakly in } \ L^2(\Omega;\mathbb{R}^3). \tag{3.67}$$

Thus we can use the usual bound

$$\liminf_{\epsilon\to 0}\int_Q \frac{\partial v_\epsilon}{\partial t}v_\epsilon \,\mathrm{d}x\,\mathrm{d}t = \liminf_{\epsilon\to 0}\frac{1}{2}\big\|v_\epsilon(T)\big\|^2_{L^2(\Omega;\mathbb{R}^3)} - \frac{1}{2}\big\|v_0\big\|^2_{L^2(\Omega;\mathbb{R}^3)} \\ \geq \frac{1}{2}\big\|v(T)\big\|^2_{L^2(\Omega;\mathbb{R}^3)} - \frac{1}{2}\big\|v_0\big\|^2_{L^2(\Omega;\mathbb{R}^3)} = \int_0^T \Big\langle \frac{\partial v}{\partial t}, v\Big\rangle \,\mathrm{d}t. \tag{3.68}$$

We also exploit (3.63) and the strong convergence $\tau(\mathrm{D}w, c_\epsilon, \theta_\epsilon) \to \tau(\mathrm{D}w, c, \theta)$ in $L^p(Q;\mathbb{R}^{3\times 3})$. Thus, from (3.65), we eventually get

$$0 \geq \int_0^T \bigg(\Big\langle \frac{\partial v}{\partial t}, v-w\Big\rangle - \int_\Omega z\cdot c\nabla\phi\cdot(v-w) \\ -(v\otimes v):\nabla(v-w) + \tau(\mathrm{D}w,c,\theta):\mathrm{D}(v-w)\,\mathrm{d}x\bigg)\,\mathrm{d}t. \tag{3.69}$$

Now we can extend this inequality for all $w \in L^p(I;W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))$ by continuity. Then, substituting $w = v + \delta\tilde{w}$, canceling $\delta > 0$, passing $\delta \to 0$, and choosing $\tilde{w}$ arbitrary, we prove that v satisfies

$$\int_0^T \bigg(\Big\langle \frac{\partial v}{\partial t}, \tilde{w}\Big\rangle + \int_\Omega (v\otimes v):\nabla\tilde{w} + \tau(\mathrm{D}v,c,\theta):\mathrm{D}\tilde{w} - z\cdot c\nabla\phi\cdot\tilde{w}\,\mathrm{d}x\bigg)\,\mathrm{d}t = 0 \tag{3.70}$$

for any $\tilde{w} \in L^p(I;W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))$. Hence v is a weak solution to (2.1a).

Let us prove the most essential and most difficult fact, namely, the strong convergence of $\mathrm{D}v_\epsilon$ to $\mathrm{D}v$ in $L^p(Q;\mathbb{R}^{3\times 3})$. We will use $\int_0^T \langle \frac{\partial(v_\epsilon - v)}{\partial t}, v_\epsilon - v\rangle \,\mathrm{d}t \geq -\frac{1}{2}\|v_{0\epsilon} - v_0\|^2_{L^2(\Omega;\mathbb{R}^3)}$; here again it is important that $\frac{\partial v}{\partial t}$ belongs to $L^{p/(p-1)}(I;W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3)^*)$ and is thus in duality to v due to (3.64) and the estimate (3.22b), and also that $\frac{\partial v_\epsilon}{\partial t}$ lives in $L^2(Q;\mathbb{R}^3)$ due to (3.22c) so it is certainly in duality with $v_\epsilon - v$. By uniform monotonicity (3.9b) of $\tau(\cdot, c_\epsilon(t,x), \theta_\epsilon(t,x))$, we get

$$\eta_2\big\|\mathrm{D}v_\epsilon - \mathrm{D}v\big\|^p_{L^p(Q;\mathbb{R}^{3\times 3}))} \leq \int_0^T\int_\Omega \big(\tau(\mathrm{D}v_\epsilon, c_\epsilon, \theta_\epsilon) - \tau(\mathrm{D}v, c_\epsilon, \theta_\epsilon)\big):\mathrm{D}(v_\epsilon - v)\,\mathrm{d}x\,\mathrm{d}t \\ \leq \int_0^T \bigg(\Big\langle \frac{\partial(v_\epsilon - v)}{\partial t}, v_\epsilon - v\Big\rangle \\ + \int_\Omega \big(\tau(\mathrm{D}v_\epsilon, c_\epsilon, \theta_\epsilon) - \tau(\mathrm{D}v, c_\epsilon, \theta_\epsilon)\big):\mathrm{D}(v_\epsilon - v)\,\mathrm{d}x\bigg)\,\mathrm{d}t + \frac{1}{2}\big\|v_{0\epsilon} - v_0\big\|^2_{L^2(\Omega;\mathbb{R}^3)} \\ = \int_0^T \bigg(\Big\langle \frac{\partial(v_\epsilon - v)}{\partial t}, v_\epsilon - v\Big\rangle + \int_\Omega \big(\tau(\mathrm{D}v_\epsilon, c_\epsilon, \theta_\epsilon) - \tau(\mathrm{D}v, c, \theta)\big):\mathrm{D}(v_\epsilon - v)\,\mathrm{d}x \\ + \int_\Omega \big(\tau(\mathrm{D}v, c, \theta) - \tau(\mathrm{D}v, c_\epsilon, \theta_\epsilon)\big):\mathrm{D}(v_\epsilon - v)\,\mathrm{d}x\bigg)\,\mathrm{d}t + \frac{1}{2}\big\|v_{0\epsilon} - v_0\big\|^2_{L^2(\Omega;\mathbb{R}^3)} \\ =: I_\epsilon^{(1)} + I_\epsilon^{(2)} + I_\epsilon^{(3)} + I_\epsilon^{(4)}. \tag{3.71}$$

Using (3.70) with $\tilde{w} := v_\epsilon - v$, the integrals $I_\epsilon^{(1)}$ and $I_\epsilon^{(2)}$ can be estimated in its sum as follows:

$$I_\epsilon^{(1)} + I_\epsilon^{(2)} = \int_0^T\int_\Omega \bigg(\frac{\partial v_\epsilon}{\partial t}\cdot v_\epsilon + \tau(\mathrm{D}v_\epsilon, c_\epsilon, \theta_\epsilon):\mathrm{D}v_\epsilon\bigg)\,\mathrm{d}x\,\mathrm{d}t \\ - \int_0^T\int_\Omega \bigg(\frac{\partial v_\epsilon}{\partial t}\cdot v + \tau(\mathrm{D}v_\epsilon, c_\epsilon, \theta_\epsilon):\mathrm{D}v\bigg)\,\mathrm{d}x\,\mathrm{d}t \tag{3.72}$$

$$
\begin{aligned}
&-\int_0^T\Big(\Big\langle\frac{\partial v}{\partial t},v_\epsilon\Big\rangle+\int_\Omega\tau(\mathrm{D}v,c,\theta):\mathrm{D}v_\epsilon\;\mathrm{d}x\Big)\,\mathrm{d}t\\
&+\int_0^T\Big(\Big\langle\frac{\partial v}{\partial t},v\Big\rangle+\int_\Omega\tau(\mathrm{D}v,c,\theta):\mathrm{D}v\;\mathrm{d}x\Big)\,\mathrm{d}t\\
=&\int_Q\Big(-\epsilon|\nabla^k v_\epsilon|^2-z\cdot c_\epsilon\nabla\phi_\epsilon\cdot v_\epsilon+(v_\epsilon\otimes v_\epsilon):\mathrm{D}v_\epsilon\\
&+\epsilon\nabla^k v_\epsilon\,\vdots\,\nabla^k\tilde v+z\cdot c_\epsilon\nabla\phi_\epsilon\cdot\tilde v-(v_\epsilon\otimes v_\epsilon):\mathrm{D}\tilde v\\
&+z\cdot c\nabla\phi\cdot v_\epsilon-(v\otimes v):\mathrm{D}v_\epsilon\\
&-z\cdot c\nabla\phi\cdot v+(v\otimes v):\mathrm{D}v\\
&-\frac{\partial v_\epsilon}{\partial t}\cdot(v-\tilde v)-\tau(\mathrm{D}v_\epsilon,c_\epsilon,\theta_\epsilon):\mathrm{D}(v-\tilde v)\Big)\,\mathrm{d}x\,\mathrm{d}t\\
\le&\int_Q\Big(-z\cdot c_\epsilon\nabla\phi_\epsilon\cdot(v_\epsilon-\tilde v)+(v_\epsilon\otimes v_\epsilon):\mathrm{D}(v_\epsilon-\tilde v)\\
&+\epsilon\nabla^k v_\epsilon\,\vdots\,\nabla^k\tilde v+z\cdot c\nabla\phi\cdot(v_\epsilon-v)-(v\otimes v):\mathrm{D}(v_\epsilon-v)\\
&-\frac{\partial v_\epsilon}{\partial t}\cdot(v-\tilde v)-\tau(\mathrm{D}v_\epsilon,c_\epsilon,\theta_\epsilon):\mathrm{D}(v-\tilde v)\Big)\,\mathrm{d}x\,\mathrm{d}t
\end{aligned}
$$

for any $\tilde v\in L^2(I;W_0^{k,2}(\Omega;\mathbb{R}^3))$. Now we can pass to the limit with $\epsilon\to 0$. The important trick is based on (3.61) with (3.64) and on integration-by-parts in time and on (3.67), which allows for

$$
\begin{aligned}
(3.73)\quad &\lim_{\epsilon\to 0}\int_Q\Big(\frac{\partial v_\epsilon}{\partial t}\cdot(v-\tilde v)+\epsilon\nabla^k v_\epsilon\,\vdots\,\nabla^k\tilde v\Big)\,\mathrm{d}x\,\mathrm{d}t\\
&=\lim_{\epsilon\to 0}\int_0^T\Big(\Big\langle(-1)^{k+1}\epsilon\Delta^k v_\epsilon-\frac{\partial v_\epsilon}{\partial t},\tilde v\Big\rangle-\Big\langle\frac{\partial v}{\partial t},v_\epsilon\Big\rangle\Big)\,\mathrm{d}t\\
&\quad+\int_\Omega\big(v_\epsilon(T)\cdot v(T)-v_{0\epsilon}\cdot v_0\big)\,\mathrm{d}x\\
&=-\int_0^T\Big(\Big\langle\frac{\partial v}{\partial t},\tilde v\Big\rangle+\Big\langle\frac{\partial v}{\partial t},v\Big\rangle\Big)\,\mathrm{d}t+\big\|v(T)\big\|^2_{L^2(\Omega;\mathbb{R}^3)}-\big\|v_0\big\|^2_{L^2(\Omega;\mathbb{R}^3)}\\
&=\int_0^T\Big\langle\frac{\partial v}{\partial t},v-\tilde v\Big\rangle\,\mathrm{d}t.
\end{aligned}
$$

The limit passage in the convective term $(v_\epsilon\otimes v_\epsilon):\mathrm{D}(v_\epsilon-\tilde v)\to(v\otimes v):\mathrm{D}(v-\tilde v)$ in $L^1(Q)$ is standard, using the strong convergence $v_\epsilon\to v$ in $L^{5p/3-\xi}(Q;\mathbb{R}^3)$ which can be proved by interpolating $L^\infty(I;L^2(\Omega))$ and $L^p(I;W^{1-\delta,p}(\Omega))$ by the Gagliardo–Nirenberg inequality (cf. [11, sect. I.3], and by using the Aubin–Lions theorem to have strong convergence in $L^p(I;W^{1-\delta,p}(\Omega))$ for any $\delta>0$; here the restriction $p>11/5$ is originated. As for the next-to-last term in (3.71), we have $\lim_{\epsilon\to 0}I_\epsilon^{(3)}=0$ because $\tau(\mathrm{D}v,c_\epsilon,\theta_\epsilon)\to\tau(\mathrm{D}v,c,\theta)$ strongly in $L^p(Q;\mathbb{R}^{3\times 3})$. By our assumption $v_{0\epsilon}\to v_0$ in $L^2(\Omega;\mathbb{R}^3)$, we also have $\lim_{\epsilon\to 0}I_\epsilon^{(4)}=0$. Now we can pass to the limit superior in

(3.71) with $\epsilon \to 0$ to obtain

$$
\begin{aligned}
(3.74)\qquad &\limsup_{\epsilon\to 0} \eta_2 \big\|\mathrm{D}v_\epsilon - \mathrm{D}v\big\|^p_{L^p(Q;\mathbb{R}^{3\times 3})} \\
&\le \int_0^T \bigg(\Big\langle \frac{\partial v}{\partial t}, v - \tilde v \Big\rangle + \int_\Omega (v\otimes v) : \mathrm{D}(v{-}\tilde v) - z{\cdot}c\nabla\phi{\cdot}(v{-}\tilde v)\,\mathrm{d}x \bigg)\,\mathrm{d}t \\
&\quad + C \limsup_{\epsilon\to 0} \Big(1 + \big\|v_\epsilon\big\|^{p-1}_{L^p(I;W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))}\Big) \big\|v - \tilde v\big\|_{L^p(I;W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))}
\end{aligned}
$$

with C being the constant from (3.9c). Using the estimates (3.22b) and (3.22d) and passing with $\tilde v$ to v in the norm topology of $L^p(I;W^{1,p}_{0,\mathrm{DIV}}(\Omega;\mathbb{R}^3))$, we can see that $\limsup_{\epsilon\to 0}\|\mathrm{D}v_\epsilon - \mathrm{D}v\|_{L^p(Q;\mathbb{R}^{3\times 3})} \le 0$; i.e., $\mathrm{D}v_\epsilon \to \mathrm{D}v$ strongly.

Having this strong convergence, we can pass to the limit in the term $\tau(\mathrm{D}v_\epsilon, c_\epsilon, \theta_\epsilon) : \mathrm{D}v_\epsilon \to \tau(\mathrm{D}v, c, \theta) : \mathrm{D}v$ in $L^1(Q)$ in the right-hand side of the heat equation (3.58d). For the limit passage in the convective term in (3.58d) it suffices to prove, in the weak formulation, that $v_\epsilon\theta_\epsilon \to v\theta$ weakly in $L^1(Q)$, which is simple due to the weak convergence $v_\epsilon \to v$ in $L^{5p/3}(Q;\mathbb{R}^3)$ based on (3.22b) with (3.44) and, by the Aubin–Lions theorem with the interpolation based on (3.22j) and (3.22k), the strong convergence $\theta_\epsilon \to \theta$ in $L^{5/3-\delta}(Q)$; thus we get $v_\epsilon\theta_\epsilon \to v\theta$ weakly even in $L^{55/48-\delta}(Q;\mathbb{R}^3)$ (see also (3.43)).

Limit passage in the other terms is routine; e.g., $\mathfrak{D}(c_\epsilon,\theta_\epsilon)\nabla c_\epsilon \cdot \nabla\phi_\epsilon \to \mathfrak{D}(c,\theta)\nabla c \cdot \nabla\phi$ because $(c_\epsilon,\theta_\epsilon) \to (c,\theta)$ strongly in $L^2(Q;\mathbb{R}^L) \times L^{5/3-\delta}(Q)$, $\nabla c_\epsilon \to \nabla c$ weakly in $L^2(Q;\mathbb{R}^{L\times 3})$ and $\nabla\phi_\epsilon \to \nabla\phi$ strongly in $L^\infty(I;L^r(\Omega;\mathbb{R}^3))$ for any $r < +\infty$, and, similarly, $(\mathfrak{m}(c_\epsilon,\theta_\epsilon)\otimes\nabla\phi_\epsilon){:}(z\otimes\nabla\phi_\epsilon) \to (\mathfrak{m}(c,\theta)\otimes\nabla\phi){:}(z\otimes\nabla\phi)$ in $L^1(Q)$. The limit passage in the term $c_\mathrm{v}\frac{\partial\theta_\epsilon}{\partial t}$ is made easy after integration-by-parts based on the estimates (3.22j) and (3.22k); i.e., $\int_Q c_\mathrm{v}\theta_\epsilon\frac{\partial w}{\partial t}\,\mathrm{d}x\,\mathrm{d}t + \int_\Omega c_\mathrm{v}\theta_{0\epsilon}w(0,\cdot)\,\mathrm{d}x$ indeed converges to $\int_Q c_\mathrm{v}\theta\frac{\partial w}{\partial t}\,\mathrm{d}x\,\mathrm{d}t + \int_\Omega c_\mathrm{v}\theta_0 w(0,\cdot)\,\mathrm{d}x$ for any test function $w \in C^1(Q)$ as used in (3.5).

Eventually, the constraints $\sum_{\ell=1}^L c_{\epsilon,\ell} = 1$ and $c_{\epsilon,\ell} \ge 0$ for $\ell = 1,\ldots,L$, which have been proved valid for the approximate solution (see (3.58f)), are inherited by the limit, too. □

Remark 3.11 (weak solutions). For $p \ge 3$, the estimate (3.22g) involves $r = 2$, and then c is the conventional weak solution to the Nernst–Planck equation (2.1c). The weak solution to the whole system (2.1) needs regularity of v. This was proved in [33] for a very narrow interval of p's (of length only about 0.0528), namely $\frac{9}{4} \le p < \frac{1+\sqrt{13}}{2}$, by using deep regularity results from [20] holding, however, only for a stress tensor τ independent of compositon and temperature and having a potential. Let us recall that regularity for the Navier–Stokes equation is generally recognized as an extremely difficult problem which is, at this writing, open and, in particular for $p = 2$ and τ linear, assigned to a $1 million Clay Mathematics Institute award.

4. Discussion of the model and particular cases. The general idea for determining the phenomenological fluxes j_ℓ is a drift/diffusion model like that in Roosbroeck's model of semiconductors [32]; for comparison of semiconductors and electrolytes; see, e.g., [37, p. 20]. In the simplest linear case, the phenomenological fluxes in Roosbroeck's model appear as

$$(4.1)\qquad j_\ell := m z_\ell c_\ell \nabla\phi + d\nabla c_\ell = m c_\ell \nabla\mu_\ell, \qquad \text{where} \quad \mu_\ell := \rho \ln c_\ell + z_\ell\phi,$$

where $m > 0$ is a mobility and $d > 0$ a diffusivity coefficient, and where μ_ℓ is the *electrochemical potential* of the ℓth constituent involving the ratio $\rho = d/m$; any influence of the temperature and its gradient on j_ℓ (in particular Soret's cross-effect) is neglected. In bipolar semiconductors, we have $L = 2$ and $z_1 = -z_2$, but here for multicomponent electrolytes we admit $L > 2$ in general. The form (4.1), however, does not satisfy (3.7d) except in the very trivial case $z_\ell m = 0$. More generally, mobilities in concrete mixtures may vary considerably for various components, especially if the size of molecules of particular constituents varies considerably from constituent to constituent [14], and then it is standardly considered that

$$j_\ell := \sum_{k=1}^{L} \mathbb{M}_{k\ell}(c)\nabla\mu_k, \tag{4.2}$$

where $\mu_\ell := \rho \ln c_\ell + z_\ell\phi$ is again from (4.1) now with $\rho = R\theta_{\rm R}$ with $\theta_{\rm R}$ a reference temperature and R a universal gas constant; let us mention that we consider z_ℓ to involve Faraday's constant.

To satisfy the zero-sum condition for the fluxes, i.e., (3.7c), (3.7d) with $\beta = 0$, the matrix $[\mathbb{M}_{k\ell}(c)]$ should satisfy

$$\forall k = 1, \ldots, L,\ \forall c \in G_1^+ : \qquad \sum_{\ell=1}^{L} \mathbb{M}_{k\ell}(c) = 0 \tag{4.3}$$

because then obviously

$$\sum_{\ell=1}^{L} j_\ell = \sum_{\ell=1}^{L}\sum_{k=1}^{L} \mathbb{M}_{k\ell}(c)\nabla\mu_k = \sum_{k=1}^{L}\ \underbrace{\sum_{\ell=1}^{L} \mathbb{M}_{k\ell}(c)}_{=0}\ \nabla\mu_k = 0. \tag{4.4}$$

Moreover, by the celebrated (Nobel prize winning) Onsager's principle [26, 27], the matrix $[\mathbb{M}_{k\ell}(c)]$ should be symmetric.

Example 4.1 (symmetric models). The zero-sum condition (4.3) for $[\mathbb{M}_{k\ell}(c)]$ has actually been adopted, e.g., in [10, 17, 28] or, in a bit different context of multicomponent alloys, in [15, 16, 23], where essentially the following matrix has been considered:

$$\mathbb{M}_{k\ell}(c) := m_\ell c_\ell \left(\delta_{k\ell} - \frac{m_k c_k}{\sum_{l=1}^{L} m_l c_l} \right) \tag{4.5}$$

with m_ℓ being "actual" mobilities of particular constituents (assumed to be) known from experiments. Such $[\mathbb{M}_{k\ell}]$ is symmetric and satisfies (4.3) because obviously $\sum_{\ell=1}^{L} \mathbb{M}_{k\ell}(c) = m_k c_k - (\sum_{\ell=1}^{L} m_\ell c_\ell) m_k c_k / (\sum_{l=1}^{L} m_l c_l) = 0$. Moreover, (4.5) also makes j_ℓ proportional to c_ℓ, which is a natural property.

Substituting (4.5) into (4.2) gives

$$j_\ell = m_\ell\big(\rho\nabla c_\ell + c_\ell z_\ell \nabla\phi\big) - \frac{m_\ell c_\ell}{\sum_{l=1}^{L} m_l c_l} \left(\sum_{k=1}^{L} m_k \big(\rho\nabla c_k + c_k z_k \nabla\phi\big) \right). \tag{4.6}$$

Comparing it with (3.8), we can see that our diffusion matrix $\mathfrak{D} = [\mathfrak{D}_{k\ell}]_{\ell,k=1}^{L}$ is now

$$\mathfrak{D}_{k\ell} = \mathfrak{D}_{k\ell}(c) = \rho m_\ell \left(\delta_{k\ell} - \frac{m_k c_\ell}{\sum_{l=1}^{L} m_l c_l} \right) \tag{4.7}$$

and our condition (3.10a) is indeed satisfied and also that (3.7c) is satisfied with $\beta = 0$ because $\sum_{\ell=1}^L \mathfrak{D}_{k\ell} = \rho m_k - \rho m_k(\sum_{\ell=1}^L c_\ell m_\ell)/(\sum_{l=1}^L c_l m_l)$=0. Therefore we can see that the coercivity and the monotonicity assumptions (3.7b) would be satisfied if and only if the matrix $[\mathfrak{D}_{k\ell}]$ given by (4.7) is positive definite uniformly with respect to $c \in G_1^+$. In fact, the positive definiteness of $[\mathfrak{D}_{k\ell}]$ in (3.7b) suffices to verify for the symmetric part of $[\mathfrak{D}_{k\ell}]$ only, and it suffices to hold on the manifold $\sum \nabla c_\ell = 0$ and in particular for $[\mathfrak{D}_{k\ell}] + \frac{\beta}{L}(\mathbf{1}\otimes\mathbf{1})$ with some $\beta \geq 0$ so that (3.7c) then holds with this β; see also [17, Chap. 7]. As for the effective mobilities $\mathfrak{m}_\ell$, comparing (4.6) with (3.8) yields

$$\mathfrak{m}_\ell = \mathfrak{m}_\ell(c) = m_\ell z_\ell c_\ell - m_\ell c_\ell \frac{\sum_{k=1}^L m_k c_k z_k}{\sum_{l=1}^L m_l c_l}, \tag{4.8}$$

and we can see that our conditions (3.7d) and (3.10b) are indeed satisfied.

Remark 4.2. The form (4.6) was suggested in [36, Remark 4.4], namely, if $m_\ell \rho \delta_{k\ell}$ and $m_\ell \delta_{k\ell}$ are taken, respectively, for quantities $d_{k\ell}$ and $m_{k\ell}$ in [36].

Remark 4.3 (special case: equal mobilities). As in (4.1), the very special situation with equal mobilities $m := m_1 = \cdots = m_L$, the formula (4.6) gives

$$j_\ell = m z_\ell c_\ell \nabla\phi + m\rho\nabla c_\ell - m c_\ell \left(\rho \sum_{k=1}^L \nabla c_k + \sum_{k=1}^L z_k c_k \nabla\phi\right); \tag{4.9}$$

again (3.7) with $\beta = 0$ holds. In fact, one can even omit $\sum_{k=1}^L \nabla c_k$ in (4.9) because it expectedly vanishes if $\sum_{k=1}^L c_k = 1$; then (3.7c) will be satisfied with $\beta = 1$ and such a j_ℓ is exactly what has been considered in [33, 34, 35, 36]; i.e.,

$$j_\ell = m c_\ell (z_\ell - q)\nabla\phi + m\rho\nabla c_\ell, \qquad \text{where again } q = \sum_{\ell=1} c_\ell z_\ell. \tag{4.10}$$

Example 4.4 (nonsymmetric models). Some other models neglect cross-effects and treat one selected constitutent, say L, in a nonsymmetric way by the formula

$$j_\ell = m_\ell c_\ell \nabla(\mu_\ell - \mu_L) \text{ for } \ell = 1, \ldots, L{-}1 \text{ and } j_L = -\sum_{\ell=1}^{L-1} j_\ell; \tag{4.11}$$

see, e.g., [18, formula (2.26)]. In this case, the symmetric matrix $[\mathbb{M}_{k\ell}(c)]$ satisfying (4.3) is given by

$$\mathbb{M}_{k\ell}(c) := \begin{cases} m_\ell c_\ell \delta_{k\ell} & \text{for } k < L,\ \ell < L, \\ -m_\ell c_\ell & \text{for } k = L,\ \ell < L, \\ -m_k c_k & \text{for } k < L,\ \ell = L, \\ \sum_{l=1}^{L-1} m_l c_l & \text{for } k = L,\ \ell = L. \end{cases} \tag{4.12}$$

Substituting μ_ℓ from (4.1) into (4.11) yields

$$j_\ell = \begin{cases} m_\ell\big(\rho\nabla c_\ell - \rho\frac{c_\ell}{c_L}\nabla c_L + (z_\ell - z_L)c_\ell\nabla\phi\big) & \text{for } \ell < L, \\ \sum_{k=1}^{L-1} m_k\big(\rho\frac{c_k}{c_L}\nabla c_L - \rho\nabla c_k + (z_L - z_k)c_\ell\nabla\phi\big) & \text{for } \ell = L. \end{cases} \tag{4.13}$$

This is sometimes used in electrochemistry either for hydrogen ions as the L-component or, in the case of very diluted water solutions, for water as the L-component.

Remark 4.5 (thermodynamics of the model). It is with a certain internal consistency and beauty that the thermodynamics of the model based on μ_ℓ from (4.1) can be derived from a single thermodynamical potential, namely, the specific *free energy* in the form

$$\psi = \psi(c, \theta, \phi, E) = \rho \sum_{\ell=1}^{L} c_\ell \big(\ln c_\ell - 1\big) - c_\mathrm{v}\,\theta \ln \frac{\theta}{\theta_\mathrm{R}} + \sum_{\ell=1}^{L} c_\ell z_\ell \phi - \frac{\varepsilon}{2} E^2, \tag{4.14}$$

where E plays the role of the *intensity of electric field*. The partial derivatives of ψ define, respectively, the *electrochemical potential* μ_ℓ, the *entropy* s, the *total electric charge* q, and the *electric induction* D, namely,

$$\mu_i = \frac{\partial \psi}{\partial c_i}, \qquad s = -\frac{\partial \psi}{\partial \theta}, \qquad q = \frac{\partial \psi}{\partial \phi}, \qquad D = -\frac{\partial \psi}{\partial E}. \tag{4.15}$$

This indeed yields expected relations, namely,

$$\mu_i = \rho \ln c_i + z_i \phi, \tag{4.16a}$$

$$s = c_\mathrm{v} \left(\ln \frac{\theta}{\theta_\mathrm{R}} + 1 \right), \tag{4.16b}$$

$$q = \sum_{\ell=1}^{L} c_\ell z_\ell, \tag{4.16c}$$

$$D = \varepsilon E. \tag{4.16d}$$

Furthermore, the *internal energy* u is then defined through Gibbs' relation as

$$u = \psi + \theta s = c_\mathrm{v} \theta + \rho \sum_{\ell=1}^{L} c_\ell \big(\ln c_\ell - 1\big) + \sum_{\ell=1}^{L} c_\ell z_\ell \phi - \frac{\varepsilon}{2} E^2. \tag{4.17}$$

It is interesting that $\frac{\varepsilon}{2}E^2$ has a negative sign in (4.17); let us remark that this term $-\frac{\varepsilon}{2}E^2$ can indeed be found in the literature (e.g., in [12, p. 342]).

The energetics of the model (2.1) considered, for simplicity, with $\alpha = 0$ in (2.8), has been derived in [36], resulting in

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \Big(\underbrace{\frac{\varrho}{2}|v|^2}_{\substack{\text{kinetic}\\ \text{energy}}} + \underbrace{\frac{\varepsilon}{2}|\nabla\phi|^2}_{\substack{\text{electrostatic}\\ \text{energy}}} + \underbrace{c_\mathrm{v}\theta}_{\substack{\text{heat part}\\ \text{of internal}\\ \text{energy } u}} \Big) \,\mathrm{d}x = \int_\Omega \underbrace{h(c,\theta)}_{\substack{\text{heat production by}\\ \text{chemical reactions}}} \,\mathrm{d}x. \tag{4.18}$$

Under the constitutive relation $E = -\nabla\phi$, we have

$$\begin{aligned} \int_\Omega q\phi - \frac{\varepsilon}{2}E^2 \,\mathrm{d}x &= -\int_\Omega (\varepsilon\Delta\phi)\phi + \frac{\varepsilon}{2}|\nabla\phi|^2 \,\mathrm{d}x \\ &= \int_\Omega \varepsilon\nabla\phi\cdot\nabla\phi - \frac{\varepsilon}{2}|\nabla\phi|^2 \,\mathrm{d}x = \int_\Omega \frac{\varepsilon}{2}|\nabla\phi|^2 \,\mathrm{d}x \end{aligned} \tag{4.19}$$

and we can rewrite the above energy balance (4.18) in terms of u and μ_ℓ as

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \left(\frac{\varrho}{2}|v|^2 + u - \sum_{\ell=1}^{L} c_\ell \mu_\ell + q\phi \right) \mathrm{d}x = \int_\Omega h(c,\theta) \,\mathrm{d}x. \tag{4.20}$$

Alternatively, (4.20) can equally be written in terms of u and c_ℓ as

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega\Big(\frac{\varrho}{2}|v|^2+u-\rho\sum_{\ell=1}^L c_\ell \ln c_\ell\Big)\,\mathrm{d}x=\int_\Omega h(c,\theta)\,\mathrm{d}x. \tag{4.21}$$

In addition, ϕ solving (2.1e) is just the critical point of the overall free energy $\Psi:\phi\mapsto\int_\Omega\psi(c,\theta,\phi,-\nabla\phi)\,\mathrm{d}x$; interestingly, as $\Psi(c,\theta,\cdot,\cdot)$ given by (4.14) is concave, this critical point is the global maximum.

The *Clausius–Duhem inequality* (under our zero-flux boundary conditions (2.8), i.e., in the isolated system) reads as

$$\begin{aligned}0&\le\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega s\,\mathrm{d}x=\int_\Omega\frac{c_{\rm v}}{\theta}\frac{\partial\theta}{\partial t}\,\mathrm{d}x\\&=\int_\Omega\frac{1}{\theta}\Big(\mathrm{div}(\kappa\nabla\theta-c_{\rm v}v\theta)+\tau(\mathrm{D}v,c,\theta):\mathrm{D}v+\sum_{\ell=1}^L f_\ell\cdot j_\ell+h\Big)\,\mathrm{d}x\\&=\int_\Omega\Big(\mathrm{div}\Big(\frac{\kappa\nabla\theta}{\theta}\Big)-c_{\rm v}v\nabla\ln\theta+\kappa\frac{|\nabla\theta|^2}{\theta^2}+\frac{\tau(\mathrm{D}v,c,\theta){:}\mathrm{D}v}{\theta}+\sum_{\ell=1}^L\frac{f_\ell{\cdot}j_\ell}{\theta}+\frac{h}{\theta}\Big)\,\mathrm{d}x,\end{aligned}$$

where $f_\ell=-z_\ell c_\ell\nabla\phi$ is the Lorenz force acting on the ℓ-constituent. The first and second terms in the last integral vanish in a thermally isolated system, and the third and fourth terms are always nonnegative (if $\theta>0$), while the nonnegativity of $\int_\Omega(\sum_{\ell=1}^L f_\ell{\cdot}j_\ell+h)/\theta\,\mathrm{d}x$ is a condition on j_ℓ and h.

Remark 4.6 (thermodynamics in the special case of equal mobilities). As already observed in [33, 36], the special case (4.10) gives the heat sources $\sum_{\ell=1}^L f_\ell\cdot j_\ell$ in the form

$$\sum_{\ell=1}^L f_\ell\cdot j_\ell=m\rho\nabla q{\cdot}\nabla\phi+\sum_{\ell=1}^L mc_\ell z_\ell^2|\nabla\phi|^2-mq^2|\nabla\phi|^2. \tag{4.22}$$

The meaning of these terms is the following: The first term $m\rho\nabla q{\cdot}\nabla\phi$ is the power (per unit volume) of the electric current arising by the diffusion flux, which can create local cooling effects; this is related to the *Peltier effect* mentioned already in Remark 2.1. This cooling effect may seemingly violate the entropy production law, but, at least in equilibrium situations (i.e., here spatially isothermal cases when $\theta(t,\cdot)$ is constant), the overall entropy production due to this term on Ω is nonnegative: indeed, by using Green's formula, one gets

$$\int_\Omega\nabla q{\cdot}\nabla\phi\,\mathrm{d}x=-\int_\Omega\varepsilon\nabla(\Delta\phi){\cdot}\nabla\phi\mathrm{d}x=\int_\Omega\varepsilon|\Delta\phi|^2\mathrm{d}x\ge0. \tag{4.23}$$

In the anisothermal case, we would get the nonnegative entropy production if the coefficient m in (4.10) were proportional to the absolute temperature θ, as it is really considered, e.g., in the kinetic theory of gases and known as Einstein's law. Such dependence would, however, make derivation of the a priori estimates (3.22e) difficult because $\inf\theta>0$ would have to be proved. The second term $\sum_{\ell=1}^L mc_\ell z_\ell^2|\nabla\phi|^2$ in (4.22) is the power of hypothetical *Joule's heat* produced by the electric currents j_ℓ in ideally diluted water solutions. The third term $-mq^2|\nabla\phi|^2=-mf_{\rm R}^2$ reduces it and represents the rate of cooling by the force which balances the volume-additivity constraint. Furthermore, the total actual Joule's heat is always nonnegative because the second term in (4.22) always dominates the third one thanks to the algebraic inequality

$$\sum_{\ell=1}^{L} c_\ell z_\ell^2 \geq \left(\sum_{\ell=1}^{L} c_\ell z_\ell\right)^2 \tag{4.24}$$

if (2.2) holds; cf. [36, Remark 2.2].

Acknowledgments. The author is deeply indebted to Petr Kaplický, Josef Málek, František Maršík, Milan Pokorný, and Ivan Samohýl for inspiring and very helpful discussions. Also many thanks go to the anonymous referees for suggesting many improvements and detecting various discrepancies in the original version of this article.

REFERENCES

[1] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions.* I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[2] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions.* II, Comm. Pure Appl. Math., 17 (1964), pp. 35–92.

[3] L. Andrej, I. Dvořák, and F. Maršík, *Biotermodynamika,* Akademia, Prague, 1982 (revised ed., F. Maršík and I. Dvořák, Akademia, Prague, 1998).

[4] J.-P. Aubin, *Un théorème de compacité,* C. R. Acad. Sci. Paris, 256 (1963), pp. 5042–5044.

[5] R. Balescu, *Equilibrium and Nonequilibrium Statistical Mechanics,* Wiley, New York, 1975.

[6] P. Baras and M. Pierre, *Problèmes paraboliques semilinéaires avec données mesures*, Appl. Anal., 18 (1984), pp. 111–149.

[7] L. Boccardo, A. Dall'aglio, T. Gallouët, and L. Orsina, *Nonlinear parabolic equations with measure data*, J. Funct. Anal., 147 (1997), pp. 237–258.

[8] L. Boccardo and T. Gallouët, *Nonlinear elliptic and parabolic equations involving measure data*, J. Funct. Anal., 87 (1989), pp. 149–169.

[9] L. Consiglieri, *Weak solutions for a class of non-Newtonian fluids with energy transfer*, J. Math. Fluid Mech., 2 (2000), pp. 267–293.

[10] P. D. de Gennes, *Dynamics of fluctuations and spinodal decomposition in polymer blends*, J. Chem. Phys., 72 (1980), pp. 4756–4763.

[11] E. DiBenedetto, *Degenerate Parabolic Equations*, Springer, New York, 1993.

[12] S. R. de Groot and P. Mazur, *Non-equilibrium Thermodynamics*, North–Holland, Amsterdam, 1962.

[13] C. Eckart, *The thermodynamics of irreversible processes.* II. *Fluid mixtures*, Phys. Rev., 58 (1940), pp. 269–275.

[14] M. Eisen, *Mathematical Models in Cell Biology and Cancer Chemotheraphy,* Springer, Berlin, 1979.

[15] H. Garcke, B. Nestler, and B. Stinner, *A diffuse interface model for alloys with multiple components and phases*, SIAM J. Appl. Math., 64 (2004), pp. 775–799.

[16] H. Garcke, B. Nestler, and B. Stinner,, *Multicomponent alloy solidification: Phase-field modeling and simulations,* Phys. Rev. E, 71 (2005), 041609.

[17] V. Giovangigli, *Multicomponent Flow Modeling,* Birkhäuser Boston, Boston, 1999.

[18] J. Kim and J. Lowengrub, *Phase field model and simulation of three-phase flows*, Interfaces Free Bound., 7 (2005), pp. 435–466.

[19] J. L. Lions, *Quelques Méthodes de Résolution des Problémes aux Limites non linéaires*, Dunod, Paris, 1969.

[20] J. Málek, J. Nečas, and M. Růžička, *On weak solutions to a class of non-Newtonian incompressible fluids in bounded three-dimensional domains: The case* $p \geq 2$, Adv. Differential Equations, 6 (2001), pp. 257–302.

[21] J. Málek, J. Nečas, M. Rokyta, and M. Růžička, *Weak and Measure-Valued Solutions to Evolutionary PDEs*, Chapman & Hall, London, 1996.

[22] N. Mills, *Incompressible mixtures of Newtonian fluids*, Internat. J. Engrg. Sci., 4 (1966), pp. 97–112.

[23] A. Miranville and G. Schimperna, *Generalized Cahn–Hilliard equations for multicomponent alloy*, Adv. Math. Sci. Appl., to appear.
[24] I. Müller and T. Ruggeri, *Rational Extended Thermodynamics*, 2nd ed., Springer, New York, 1998.
[25] J. Nečas and M. Šilhavý, *Multipolar viscous fluids*, Quart. Appl. Math., 49 (1991), pp. 247–265.
[26] L. Onsager, *Reciprocal relations in irreversible processes*, Phys. Rev. II, 37 (1931), pp. 405–426.
[27] L. Onsager, *Reciprocal relations in irreversible processes*, Phys. Rev. II, 38 (1931), pp. 2265–2279.
[28] F. Otto and W. E, *Thermodynamically driven incompressible mixtures*, J. Chem. Phys., 107 (1997), pp. 10177–10184.
[29] I. Prigogine, *Étude Thermodynamique des Phénomenès Irréversibles*, Desoer, Liege, Belgium, 1947.
[30] K. R. Rajagopal and L. Tao, *Mechanics of Mixtures*, World Scientific, River Edge, NJ, 1995.
[31] K. R. Rajagopal, A. S. Wineman, and M. Gandhi, *On boundary conditions for a certain class of problems in mixture theory*, Internat. J. Engrg. Sci., 24 (1986), pp. 1453–1463.
[32] W. van Roosbroeck, *Theory of flow of electrons and holes in Germanium and other semiconductors*, Bell System Tech. J., 29 (1950), pp. 560–607.
[33] T. Roubíček, *Incompressible ionized fluid mixtures: A non-Newtonian approach*, IASME Trans. (WSEAS), 2 (2005), pp. 1190–1197.
[34] T. Roubíček, *Incompressible fluid mixtures of ionized constituents*, in Proceedings of Trends in Applications of Mathematics to Mechanics, Y. Wang and K. Hutter, eds., Shaker, Aachen, Germany, 2005, pp. 429–440.
[35] T. Roubíček, *Nonlinear Partial Differential Equations with Applications*, Birkhäuser-Verlag, Basel, 2005.
[36] T. Roubíček, *Incompressible ionized fluid mixtures*, Contin. Mech. Thermodyn., 17 (2006), pp. 496–509.
[37] I. Rubinstein, *Electro-Diffusion of Ions,* SIAM, Philadelphia, 1990.
[38] I. Samohýl, *Thermodynamics of Irreversible Processes in Fluid Mixtures,* Teubner, Leipzig, Germany, 1987.
[39] I. Samohýl, *Thermodynamics of mixtures of reacting and non-reacting fluids with heat conduction, diffusion and viscosity*, Internat. J. Non-Linear Mech., 32 (1997), pp. 241–257.
[40] I. Samohýl, *Application of Truesdell's model of mixture to ionic liquid mixture*, Comp. Math. Appl., 53 (2007), pp. 182–197.
[41] I. Samohýl and M. Šilhavý, *Mixture invariance and its applications*, Arch. Rational Mech. Anal., 109 (1990), pp. 299–321.
[42] C. Truesdell and R. Toupin, *The Classical Field Theories*, Handbuch der Physik III/1, Springer, Berlin, 1960.

SIAM J. MATH. ANAL.
Vol. 39, No. 3, pp. 891–915

KINETIC FORMULATION FOR A PARABOLIC CONSERVATION LAW. APPLICATION TO HOMOGENIZATION*

ANNE-LAURE DALIBARD†

Abstract. We derive a kinetic formulation for the parabolic scalar conservation law $\partial_t u + \mathrm{div}_y A(y,u) - \Delta_y u = 0$. This allows us to define a weaker notion of solutions in L^1, which is enough to recover the L^1 contraction principle. We also apply this kinetic formulation to a homogenization problem studied in a previous paper; namely, we prove that the kinetic solution u^ε of $\partial_t u^\varepsilon + \mathrm{div}_x A(x/\varepsilon, u^\varepsilon) - \varepsilon \Delta_x u^\varepsilon = 0$ behaves in L^1_{loc} as $v(x/\varepsilon, \bar{u}(t,x))$, where v is the solution of a cell problem and $\bar{u}$ the solution of the homogenized problem.

Key words. scalar conservation law, kinetic formulation, homogenization

AMS subject classifications. 35K55, 35B27

DOI. 10.1137/060662770

1. Introduction.

This paper is devoted to the study of the solution $u \in \mathcal{C}([0,\infty), L^1(Y)) \cap L^2_{\mathrm{loc}}([0,\infty), H^1_{\mathrm{per}}(Y)) \cap L^\infty_{\mathrm{loc}}([0,\infty) \times Y)$ of the equation

$$\begin{cases} \partial_t u(t,y) + \mathrm{div}_y A(y, u(t,y)) - \Delta_y u(t,y) = 0, & t > 0,\ y \in Y, \\ u(t=0,y) = u_0(y), \end{cases} \tag{1.1}$$

where $Y = [0,1]^N$ is the N-dimensional torus; $A = A(y,v) \in \mathbb{R}^N$, $y \in Y$, $v \in \mathbb{R}$, is a given N-dimensional flux, periodic in the space variable y.

In [5], a kinetic formulation was derived for such heterogeneous conservation laws (in fact, that work was achieved for hyperbolic laws, but it can be generalized to parabolic laws with no difficulty), based on the previous papers of Lions, Perthame, and Tadmor concerning hyperbolic homogeneous conservation laws (see [14], [13], [18], [16], and the general presentation in [17]). However, this formulation is not entirely satisfactory: indeed, it is based on the comparison between the solution $u(t,y)$ of the conservation law and the constants via the function $\mathbf{1}_{v<u(t,y)}$, where v is an additional fluctuation variable. But the constants, which happen to be stationary solutions of homogeneous conservation laws, no longer play a special role in the context of heterogeneous conservation laws. Hence, our goal in this article is to derive a kinetic formulation based on the study of the stationary solutions of (1.1). Let us mention a related work of Audusse and Perthame [2], which defines a notion of entropy solution which is not based on Kruzkhov's inequalities, but rather on the comparison with special stationary solutions, and which is sufficient to derive the L^1 contraction principle.

Let us make precise some notation which will be used later on: if $\mathcal{C}^\infty_{\mathrm{per}}(Y)$ denotes the space of Y-periodic functions in $\mathcal{C}^\infty(\mathbb{R}^N)$, then

$$W^{k,p}_{\mathrm{per}}(Y) := \overline{\mathcal{C}^\infty_{\mathrm{per}}(Y)}^{W^{k,p}(Y)},$$
$$W^{1,\infty}_{\mathrm{per,loc}}(Y \times \mathbb{R}) := \{u = u(y,v) \in W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^{N+1}), u \text{ is } Y\text{-periodic in } y\},$$

*Received by the editors June 12, 2006; accepted for publication (in revised form) February 27, 2007; published electronically September 19, 2007.
http://www.siam.org/journals/sima/39-3/66277.html

†CEREMADE-UMR 7534, Université Paris-Dauphine, Place du maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France (dalibard@ceremade.dauphine.fr).

$$\mathcal{D}_{\rm per}([0,\infty)\times Y\times\mathbb{R}) := \{u = u(t,y,v)\in \mathcal{C}^\infty([0,\infty)\times\mathbb{R}^{N+1}),$$
$$u \text{ is periodic in } y \text{ and } \exists R>0,\ u(t,y,v)=0 \text{ if } t+|v|\geq R\},$$
$$\langle v\rangle := \frac{1}{|Y|}\int_Y v(y)\,dy \quad \forall v\in L^1(Y).$$

First, let us recall a few results on the stationary solutions of (1.1), which were studied in [3].

Proposition 1.1. *Let $A = A(y,v)\in W^{1,\infty}_{per,loc}(Y\times\mathbb{R})^N$.*

Let $a_i(y,v) := \partial_v A_i(y,v)$, $1\leq i\leq N$, $b(y,v) := \mathrm{div}_y A(y,v) \in L^\infty_{loc}(\mathbb{R}^{N+1})$. Assume that there exist real numbers $C_0>0$, $m\in[0,\infty)$, $n\in[0,\frac{N+2}{N-2})$ when $N\geq 3$, such that for all $(y,p)\in Y\times\mathbb{R}$

$$|a_i(y,p)| \leq C_0\left(1+|p|^m\right) \quad \forall 1\leq i\leq N, \tag{1.2}$$
$$|b(y,p)| \leq C_0\left(1+|p|^n\right). \tag{1.3}$$

Assume as well that the couple (m,n) satisfies at least one of the following conditions:

$$m=0 \tag{1.4}$$

or

$$0\leq n<1 \tag{1.5}$$

or

$$n<\min\left(\frac{N+2}{N},2\right) \quad \text{and} \quad \exists p_0\in\mathbb{R}\ \forall y\in Y \quad b(y,p_0)=0. \tag{1.6}$$

Then for all $p\in\mathbb{R}$, there exists a unique solution $v(\cdot,p)\in H^1_{per}(Y)$ of the equation

$$-\Delta_y v(y,p) + \mathrm{div}_y A(y,v(y,p)) = 0, \quad \langle v(\cdot,p)\rangle = p. \tag{1.7}$$

For all $p\in\mathbb{R}$, $v(\cdot,p)$ belongs to $W^{2,q}_{per}(Y)$ for all $1<q<+\infty$ and satisfies the following a priori estimate: for all $R>0$, there exists a constant $C_R>0$ depending only on N, Y, C_0, m, n, q, p_0, and R, such that

$$||v(\cdot,p)||_{W^{2,q}(Y)} \leq C_R \quad \forall p\in\mathbb{R},\ |p|\leq R. \tag{1.8}$$

Moreover, for all $p\in\mathbb{R}$, $\partial_p v(\cdot,p)\in H^1_{per}(Y)$ and is a solution of

$$-\Delta_y\frac{\partial v}{\partial p} + \mathrm{div}_y\left[a(y,v(y,p))\frac{\partial v}{\partial p}\right] = 0, \quad \left\langle\frac{\partial v}{\partial p}\right\rangle = 1. \tag{1.9}$$

And for all $R>0$, there exists $\alpha>0$ depending only on N, Y, C_0, m, n, q, p_0, and R, such that for all $(y,p)\in Y\times(-R,R)$,

$$\frac{\partial v}{\partial p}(y,p)\geq\alpha>0.$$

Equation (1.7) is also called a "cell problem" on account of its significance in homogenization problems.

Following the idea of Audusse and Perthame (see [2]), we now give a notion of entropy solution for (1.1) based on the comparison with stationary solutions.

Definition 1.2. *Assume the hypotheses of Proposition* 1.1 *are satisfied.*

Let $u \in \mathcal{C}([0,\infty), L^1(Y)) \cap L^2_{loc}([0,\infty), H^1_{per}(Y)) \cap L^\infty_{loc}([0,\infty)\times Y)$ *be a solution of* (1.1). *We say that* u *is an entropy solution of* (1.1) *if* u *satisfies the inequality*

$$\partial_t (u(t,y)-v(y,p))_+ + \mathrm{div}_y \left[\mathbf{1}_{u>v(y,p)} (A(y,u)-A(y,v(y,p)))\right] - \Delta_y (u(t,y)-v(y,p))_+ \le 0 \tag{1.10}$$

for all $p \in \mathbb{R}$ *and in the sense of distributions on* $[0,\infty)\times Y$.

Notice that this notion of entropy solution is different (at least in its formulation) from the one of Kruzkhov, since the latter is based on the comparison with constants. However, inequality (1.10) was known by Kruzkhov, since it can be considered as a particular case of the comparison principle (notice that $v(y,p)$ is a stationary solution of (1.1)). It will be proved in the second section, under suitable regularity assumptions on the flux function A, that all solutions of (1.1) are entropy solutions in the sense of Definition 1.2.

Let us mention here an important application of inequality (1.10) and of the kinetic formulation which follows from (1.10): we give in this paper another proof for a homogenization result proved in [3], which we recall here for the reader's convenience.

Proposition 1.3. *Assume that* $A \in W^{1,\infty}_{per,loc}(\mathbb{R}^{N+1})^N$ *satisfies the assumptions of Proposition* 1.1, *and that* $\partial_{y_j} a_i \in L^1_{loc}(\mathbb{R}^{N+1})$, $\partial_v a_i \in L^1_{loc}(\mathbb{R}^{N+1})$ *for* $1 \le i \le N+1$, $1 \le j \le N$.

For $\varepsilon > 0$, *let* $v^\varepsilon \in L^\infty_{loc}([0,\infty)\times\mathbb{R}^N) \cap \mathcal{C}([0,\infty), L^1_{loc}(\mathbb{R}^N)) \cap L^2_{loc}([0,\infty), H^1_{loc}(\mathbb{R}^N))$ *be a solution of the parabolic scalar conservation law:*

$$\frac{\partial v^\varepsilon}{\partial t}(t,x) + \sum_{i=1}^N \frac{\partial}{\partial x_i} A_i\left(\frac{x}{\varepsilon}, v^\varepsilon(t,x)\right) - \varepsilon \Delta_x v^\varepsilon = 0, \quad t \ge 0,\ x \in \mathbb{R}^N, \tag{1.11}$$

$$v^\varepsilon(t=0) = v_0\left(x, \frac{x}{\varepsilon}\right). \tag{1.12}$$

Let $p \in \mathbb{R}$, *and let* $v = v(y,p)$ *be the unique solution in* $H^1_{per}(Y)$ *of the cell problem* (1.7). *Define*

$$\bar{A}_i(p) := \frac{1}{|Y|}\int_Y A_i(y, v(y,p))\, dy. \tag{1.13}$$

Assume also that v_0 *is "well-prepared," i.e., satisfies*

$$v_0(x,y) = v(y, \bar{v}_0(x)) \tag{1.14}$$

for some $\bar{v}_0 \in L^1 \cap L^\infty(\mathbb{R}^N)$.

Then as ε *goes to* 0,

$$v^\varepsilon(t,x) - v\left(\frac{x}{\varepsilon}, \bar{v}(t,x)\right) \to 0 \quad \text{in } L^2_{loc}([0,\infty)\times\mathbb{R}^N),$$

where $\bar{v} = \bar{v}(t,x) \in \mathcal{C}([0,\infty), L^1(\mathbb{R}^N)) \cap L^\infty([0,\infty)\times\mathbb{R}^N)$ *is the unique entropy solution of the hyperbolic scalar conservation law*

$$\begin{cases} \dfrac{\partial \bar{v}}{\partial t} + \displaystyle\sum_{i=1}^N \frac{\partial \bar{A}_i(\bar{v}(t,x))}{\partial x_i} = 0, \\ \bar{v}(t=0,x) = \bar{v}_0(x) \in L^1 \cap L^\infty(\mathbb{R}^N). \end{cases} \tag{1.15}$$

Actually, the result proved in section 3 is more general than Proposition 1.3 but is much more complicated to state at this stage. In particular, we work in an L^1 rather than L^∞ setting, which appears to us to be entirely new for this kind of equation; this point will be developed a little further in Remark 3.2. We emphasize that inequality (1.10) was already used in [3], but we believe that the proof given here provides better insight into the homogenization process.

Let us mention related results of W. E (see [6], [7]) and W. E and Serre (see [8]), who use two-scale Young measures instead of the kinetic formulation in a hyperbolic context. In fact, the proof of [3] is close to those of these articles, although the viscous term in (1.11) is absent from the problems studied by W. E in [6] and W. E and Serre in [8]. Indeed, the scaling in our problem is chosen so that the viscosity has the same order of magnitude than the size of the oscillations in the flux function, and thus the viscosity has an effect at a microscopic level only. Notice that the (macroscopic) homogenized problem (1.15) is hyperbolic; this justifies the use of hyperbolic tools, such as Young measures or a kinetic formulation, in the study of (1.11).

We also wish to point out that the expression of the homogenized flux in the case studied by W. E and Serre in [8] when $N = 1$ is the same as that in (1.13). However, the corrector v appearing in the expression is not the same in both cases: indeed, in the hyperbolic case studied by W. E and Serre, v is a solution of

$$\partial_y A(y, v(y,p)) = 0.$$

In particular, v is not unique in general, although the homogenized flux is. We refer the interested reader to [8] and [12] for details; the latter uses an equivalent formulation using Hamilton–Jacobi equations.

The organization of this article is as follows: first we derive a kinetic formulation for (1.1). As usual, this allows us to define a weaker notion of solutions of the parabolic conservation law (1.1), called *kinetic solutions*. We also derive formally the L^1 contraction principle for kinetic solutions of (1.1). Then we use this formulation to give another proof of Proposition 1.3 in section 3. Eventually, in section 4 we give a rigorous proof for the derivation of the L^1 contraction principle announced in section 2.

2. Kinetic formulation. This section is devoted to the derivation of a kinetic formulation for (1.1). Throughout the section, we assume that the hypotheses of Proposition 1.1 are satisfied, that is, $A \in W^{1,\infty}_{\text{per,loc}}(Y \times \mathbb{R})$, and A satisfies either (1.4), (1.5), or (1.6). Additionally, we assume that

$$a(y,\cdot) \in \mathcal{C}(\mathbb{R})^N \quad \text{for a.e. } y \in Y. \tag{2.1}$$

Under such hypotheses, the following result is easily deduced from Proposition 1.1.

LEMMA 2.1. *For a.e. $y \in Y$, $p \mapsto v(y,p)$ is a $\mathcal{C}^1$ diffeomorphism from $\mathbb{R}$ to $(\alpha_-(y), \alpha_+(y))$, where $\alpha_\pm(y) = \lim_{p\to\pm\infty} v(y,p)$.*

Its reciprocal application is denoted by $w(y,\cdot)$

$$w(y) : (\alpha_-(y), \alpha_+(y)) \to \mathbb{R}.$$

REMARK 2.1. *Notice that $+\infty$ (resp., $-\infty$) is an admissible value for α_+ (resp., α_-). In fact, it can be checked that*

$$\langle \alpha_\pm \rangle = \pm\infty,$$

and there are cases when

$$\alpha_\pm(y) = \pm\infty \quad \forall y \in Y.$$

Indeed, for all $y \in Y$, the family $(v(y,p) - v(y,0))_{p>0}$ is increasing in p and nonnegative. Moreover,

$$\langle v(\cdot,p) - v(\cdot,0)\rangle = p \quad \forall p \in \mathbb{R}.$$

Hence according to Lebesgue's monotone convergence theorem, $\langle \alpha_+ - v(\cdot,0)\rangle = +\infty$, and thus $\langle \alpha_+\rangle = +\infty$. If we assume additionally that $m = 0$ in hypothesis (1.2) *(i.e., we assume that* (1.4) *is satisfied), then it is proved in* [3, Lemma 6] *that*

$$\lim_{p\to+\infty} \inf_{y\in Y} v(y,p) = +\infty.$$

In that case, $\alpha_+(y) = +\infty$ for all $y \in Y$.

We begin our study of (1.1) with the following lemma.

LEMMA 2.2. *Let $u \in \mathcal{C}([0,\infty); L^1(Y)) \cap L^2_{loc}(0,\infty; H^1_{per}(Y)) \cap L^\infty_{loc}([0,\infty)\times Y)$ be an arbitrary solution of* (1.1). *Assume that the flux $A \in W^{1,\infty}_{per,loc}(Y\times\mathbb{R})$ satisfies* (2.1) *and the hypotheses of Proposition* 1.1.

Then the function u satisfies the following equality in the sense of distributions on $[0,\infty)\times Y\times\mathbb{R}_p$:

(2.2)
$$\partial_t (u - v(y,p))_+ + \mathrm{div}_y \left[\mathbf{1}_{u>v(y,p)}(A(y,u) - A(y,v(y,p)))\right] - \Delta_y (u - v(y,p))_+ = -m,$$

where

$$m(t,y,p) = \frac{1}{\frac{\partial v}{\partial p}(y,p)} \left|\nabla_y (u(t,y) - v(y,p))\right|^2 \delta(p = w(y,u(t,y)))$$

is a nonnegative measure on $(0,\infty)\times Y\times\mathbb{R}$.

Consequently, u is an entropy solution of (1.1) *in the sense of Definition* 1.2.

We postpone the proof of Lemma 2.2 to the end of section 2. Let us stress that (2.2) is to be understood in the sense of distributions in $[0,\infty)\times Y\times\mathbb{R}$. Such an equality would indeed be meaningless were it considered for $p \in \mathbb{R}$ fixed.

Let us now write down the kinetic formulation for (1.1). Let u be an entropy solution of (1.1); differentiating (2.2) with respect to p leads to

(2.3)
$$\frac{\partial}{\partial t}\left(\frac{\partial v(y,p)}{\partial p} f^+\right) + \frac{\partial}{\partial y_i}\left(\frac{\partial v(y,p)}{\partial p} a_i(y,v(y,p)) f^+\right) - \Delta_y \left(\frac{\partial v(y,p)}{\partial p} f^+\right) = \frac{\partial m(t,y,p)}{\partial p},$$

where $f^+(t,y,p) = \mathbf{1}_{u(t,y)>v(y,p)}$.

The same kind of equation holds for $f^- = \mathbf{1}_{u(t,y)<v(y,p)} = 1 - f^+$ (recall (1.9)):

(2.4)
$$\frac{\partial}{\partial t}\left(\frac{\partial v(y,p)}{\partial p} f^-\right) + \frac{\partial}{\partial y_i}\left(\frac{\partial v(y,p)}{\partial p} a_i(y,v(y,p)) f^-\right) - \Delta_y \left(\frac{\partial v(y,p)}{\partial p} f^-\right) = -\frac{\partial m(t,y,p)}{\partial p}.$$

This leads to a notion of kinetic solution.

DEFINITION 2.3. *Assume that the flux A satisfies the hypotheses of Proposition* 1.1 *and* (2.1). *Let $u = u(t,y) \in \mathcal{C}([0,\infty); L^1(Y)) \cap L^2_{loc}(0,\infty; H^1_{per}(Y))$ such that*

$$\alpha_-(y) < u(t,y) < \alpha_+(y) \quad \textit{for a.e. } (t,y) \in [0,\infty)\times Y.$$

We say that u is a kinetic solution of (1.1) *if $f^+ = \mathbf{1}_{u(t,y)>v(y,p)}$ satisfies* (2.3) *in the sense of distributions with the initial data $f^+(t=0,y,p) = \mathbf{1}_{u_0(y)>v(y,p)}$, and if there exists a function $\mu \in L^\infty(\mathbb{R})$ such that $\mu(p) \to 0$ as $|p| \to \infty$, and*

$$\int_0^\infty \int_Y m(t,y,p)\,dy\,dt \le \mu(p) \quad \textit{in } \mathcal{D}'(\mathbb{R}). \tag{2.5}$$

Precisely, u is a kinetic solution of (1.1) *if* (2.5) *holds and if for every test function* $\psi = \psi(t,y,p) \in \mathcal{D}_{per}([0,\infty) \times Y \times \mathbb{R})$, *we have*

$$\int_0^\infty \int_{Y\times\mathbb{R}} f^+(t,y,p) \frac{\partial v(y,p)}{\partial p} \{\partial_t \psi + a_i(y, v(y,p))\partial_{y_i}\psi + \Delta_y \psi\}\, dt\, dy\, dp \tag{2.6}$$
$$= \int_0^\infty \int_{Y\times\mathbb{R}} m(t,y,p)\partial_p \psi(t,y,p)\, dtdydp - \int_{Y\times\mathbb{R}} \mathbf{1}_{u_0(y)>v(y,p)} \frac{\partial v(y,p)}{\partial p} \psi(0,y,p) dydp.$$

Notice that without any loss of generality, we can choose a function μ in (2.5) which is nonincreasing on $(0,\infty)$ and nondecreasing on $(-\infty, 0)$.

It is easily checked that the notions of entropy and kinetic solutions are equivalent as long as u is bounded in some kind of L^∞ norm.

Proposition 2.4. *Assume that A satisfies* (2.1) *and the hypotheses of Proposition* 1.1. *Let $u = u(t,y) \in \mathcal{C}([0,\infty); L^1(Y)) \cap L^2_{loc}(0,\infty; H^1_{per}(Y))$. Assume that there exist real numbers $\beta_1, \beta_2 \in \mathbb{R}$ such that*

$$v(y,\beta_1) \le u(t,y) \le v(y,\beta_2) \quad \text{for a.e. } (t,y) \in (0,\infty) \times Y. \tag{2.7}$$

Then u is an entropy solution of (1.1) *if and only if u is a kinetic solution.*

We are then able to prove the L^1 contraction principle thanks to the kinetic formulation; we wish to emphasize that when u satisfies (2.7), this result is not new by any means, and has been known since the articles of Kruzkhov [10], [11]. However, we present here a different proof (see section 4), using merely regularizations by convolution following [16], [17]. Moreover, we prove that the L^1 contraction principle holds for a larger class of solutions.

Theorem 2.5. *Assume the hypotheses of Proposition* 1.1 *are satisfied, with $a \in W^{1,1}_{per,loc}(Y \times \mathbb{R})^N$, and*

$$\partial_v a \in L^\infty_{loc}(Y\times\mathbb{R})^N, \tag{2.8}$$

(2.9)
$$\forall R>0,\ \exists\, \alpha, C>0,\ \forall (y,y') \in Y^2,\ \forall v \in (-R,R) \quad |a(y,v) - a(y',v)| \le C|y-y'|^\alpha.$$

Let u_1, u_2 be two kinetic solutions of (1.1). *Then*

$$||(u_1(t) - u_2(t))_+||_{L^1(Y)} \le ||(u_1(t=0) - u_2(t=0))_+||_{L^1(Y)}. \tag{2.10}$$

Moreover, if for all $T>0$

$$\int_0^T \int_Y \int_{\mathbb{R}} \frac{\partial v(y,p)}{\partial p} |a(y, v(y,p))|\, \mathbf{1}_{u_2(t,y)<v(y,p)<u_1(t,y)}\, dt\, dy\, dp < +\infty, \tag{2.11}$$

then the following inequality holds in the sense of distributions on $[0,\infty) \times Y$:

$$\frac{\partial}{\partial t}(u_1-u_2)_+ + \frac{\partial}{\partial y_i}\left[\mathbf{1}_{u_1>u_2} \left(A_i(y,u_1) - A_i(y,u_2)\right)\right] - \Delta_y (u_1-u_2)_+ \le 0. \tag{2.12}$$

Remark 2.2. *Hypothesis* (2.11) *is necessary in order to retrieve inequality* (2.12). *However, if the sole purpose is to derive the L^1 contraction inequality* (2.10), *hypothesis* (2.11) *is no longer required. Hypothesis* (2.11) *implies that the function*

$$(t,y) \mapsto \mathbf{1}_{u_1>u_2} \left[A(y,u_1(t,y)) - A(y,u_2(t,y))\right]$$

belongs to $L^1((0,T)\times Y)^N$ for all $T>0$. Notice that such an integrability property is not obvious in general, since we no longer assume that $u\in L^\infty_{loc}$, and thus $A(\cdot,u)$ does not belong to L^∞_{loc} either.

Let us explain formally how inequality (2.12) is derived: let u_1, u_2 be two kinetic solutions of (1.1). We set $f_1=\mathbf{1}_{u_1(t,y)>v(y,p)}$, $f_2=\mathbf{1}_{u_2(t,y)<v(y,p)}$,

$$m_i=|\nabla_y u_i(t,y)-\nabla_y v(y,p)|^2\frac{1}{\frac{\partial v(y,p)}{\partial p}}\delta(p=w(y,u_i(t,y))),\quad i=1,2.$$

Then

$$\frac{\partial}{\partial t}\left(\frac{\partial v}{\partial p}f_1\right)+\frac{\partial}{\partial y_i}\left(\frac{\partial v}{\partial p}a_i(y,v(y,p))f_1\right)-\Delta_y\left(\frac{\partial v}{\partial p}f_1\right)=\frac{\partial m_1}{\partial p},\tag{2.13}$$

$$\frac{\partial}{\partial t}\left(\frac{\partial v}{\partial p}f_2\right)+\frac{\partial}{\partial y_i}\left(\frac{\partial v}{\partial p}a_i(y,v(y,p))f_2\right)-\Delta_y\left(\frac{\partial v}{\partial p}f_2\right)=-\frac{\partial m_2}{\partial p}.\tag{2.14}$$

Multiply (2.13) by f_2, and (2.14) by f_1; recalling (1.9), we add the two equations thus obtained and are led to

$$\begin{aligned}\frac{\partial}{\partial t}\left(\frac{\partial v}{\partial p}f_1f_2\right)+\frac{\partial}{\partial y_i}\left(\frac{\partial v}{\partial p}a_i(y,v(y,p))f_1f_2\right)-\Delta_y\left(\frac{\partial v}{\partial p}f_1f_2\right)&\\ =\frac{\partial m_1}{\partial p}f_2-\frac{\partial m_2}{\partial p}f_1-2\frac{\partial v}{\partial p}\nabla_y f_1\cdot\nabla_y f_2.&\end{aligned}\tag{2.15}$$

Set $\varphi_i(t,y)=w(y,u_i(t,y))$ $(i=1,2)$, i.e., $v(y,\varphi_i(t,y))=u_i(t,y)$. Then

$$\nabla_y\varphi_i(t,y)=\frac{1}{\frac{\partial v}{\partial p}(y,\varphi_i(t,y))}\left[\nabla_y u_i(t,y)-\nabla_y v(y,\varphi_i(t,y))\right].$$

Notice that

$$\begin{aligned}f_1&=\mathbf{1}_{u_1(t,y)>v(y,p)}=\mathbf{1}_{\varphi_1(t,y)>p},\\ f_2&=\mathbf{1}_{u_2(t,y)<v(y,p)}=\mathbf{1}_{\varphi_2(t,y)<p},\end{aligned}$$

and thus, setting $\eta_1=1$ and $\eta_2=-1$,

$$\begin{aligned}\frac{\partial f_i}{\partial p}&=-\eta_i\,\delta(p=\varphi_i(t,y)),\\ \nabla_y f_i&=\eta_i\nabla_y\varphi_i(t,y)\delta(p=\varphi_1(t,y)).\end{aligned}$$

We refer to the proof of Lemma 2.2, at the end of the present section, for a derivation of the above equalities in the sense of distributions on $[0,\infty)\times Y\times\mathbb{R}$.

On the other hand, for any function $G\in W^{1,\infty}_{\rm loc}(\mathbb{R})$,

$$\begin{aligned}\int_{\mathbb{R}}G'(v(y,p))f_1f_2\frac{\partial v(y,p)}{\partial p}\,dp&=\int_{\mathbb{R}}G'(v(y,p))\mathbf{1}_{u_2(t,y)<v(y,p)<u_1(t,y)}\frac{\partial v(y,p)}{\partial p}\,dp\\ &=\mathbf{1}_{u_2(t,y)<u_1(t,y)}\left[G(u_1(t,y))-G(u_2(t,y))\right].\end{aligned}$$

Hence, integrating (2.15) with respect to p on $\mathbb{R}$ yields

$$\begin{aligned}
&\frac{\partial}{\partial t}(u_1-u_2)_+ + \frac{\partial}{\partial y_i}\mathbf{1}_{u_2(t,y)<u_1(t,y)}\left[A_i(y,u_1(t,y))-A_i(y,u_2(t,y))\right]-\Delta_y(u_1-u_2)_+\\
&=\int_{\mathbb{R}} -m_1\partial_p f_2 + m_2\partial_p f_1 - 2\frac{\partial v}{\partial p}\nabla_y f_1\cdot\nabla_y f_2\,dp\\
&=-\int_{\mathbb{R}}|\nabla_y u_1(t,y)-\nabla_y v(y,\varphi_1)|^2\frac{1}{\frac{\partial v(y,p)}{\partial p}}\delta(p=\varphi_1)\delta(p=\varphi_2)dp\\
&\quad-\int_{\mathbb{R}}|\nabla_y u_2(t,y)-\nabla_y v(y,\varphi_2)|^2\frac{1}{\frac{\partial v(y,p)}{\partial p}}\delta(p=\varphi_2)\delta(p=\varphi_1)dp\\
&\quad+2\int_{\mathbb{R}}\frac{\partial v}{\partial p}(y,p)\,\nabla_y\varphi_1(t,y)\cdot\nabla_y\varphi_2(t,y)\delta(p=\varphi_1)\delta(p=\varphi_2)dp\\
&=-\int_{\mathbb{R}}\frac{1}{\frac{\partial v(y,p)}{\partial p}}\delta(p=\varphi_1)\delta(p=\varphi_2)|\nabla_y(u_1-u_2)(t,y)-\nabla_y v(y,\varphi_1)+\nabla_y v(y,\varphi_2)|^2\,dp\\
&\le 0,
\end{aligned}$$

which is exactly the L^1 contraction principle between u_1 and u_2.

However, the calculations above are entirely formal, since the product of Dirac masses is not a well-defined object, and f_1, f_2 do not have enough regularity to perform nonlinear calculations. Thus, regularizations are necessary in order to justify the contraction principle, which is proved in section 4.

Proof of Lemma 2.2. Notice first that since $u(t,y)$ and $v(y,p)$ are both solutions of (1.1), we always have

$$\partial_t\left[u(t,y)-v(y,p)\right]+\mathrm{div}_y\left[A(y,u)-A(y,v(y,p))\right]-\Delta_y\left[u(t,y)-v(y,p)\right]=0.$$

Thanks to the regularizing parabolic (resp., elliptic) term, the regularity of u (resp., v) is sufficient for us to use the chain rule, and thus

$$\begin{aligned}
\mathbf{1}_{u(t,y)>v(y,p)}\partial_t\left[u(t,y)-v(y,p)\right]&=\partial_t\left[u(t,y)-v(y,p)\right]_+,\\
\mathbf{1}_{u(t,y)>v(y,p)}\mathrm{div}_y\left[A(y,u)-A(y,v(y,p))\right]&=\mathrm{div}_y\left[\mathbf{1}_{u(t,y)>v(y,p)}\left(A(y,u)-A(y,v(y,p))\right)\right],\\
\mathbf{1}_{u>v(y,p)}\Delta_y\left[u-v(y,p)\right]&=\Delta_y\left[u-v(y,p)\right]_+-\nabla_y\mathbf{1}_{u>v(y,p)}\cdot\nabla_y\left[u-v(y,p)\right].
\end{aligned}$$

Similar calculations can be found, for instance, in [10], [11] and are in fact at the heart of Kruzkhov's method for proving the L^1 contraction principle.

The major difficulty comes from the term $\nabla_y\mathbf{1}_{u(t,y)>v(y,p)}$. Notice that $\mathbf{1}_{u(t,y)>v(y,p)} = \mathbf{1}_{w(y,u(t,y))>p}$. When $p\in\mathbb{R}$ is considered as a fixed parameter, we have

$$\nabla_y\mathbf{1}_{u>v(y,p)}=\nu\otimes\mathcal{H}^{n-1}_{\partial\omega},$$

where $\omega := \{y\in Y;\ w(y,u(t,y))>p\}$, $\mathcal{H}^{n-1}_{\partial\omega}$ is the $(n-1)$-dimensional Hausdorff measure along $\{w(y,u(t,y))=p\}$, and ν is the unit normal vector field oriented from $\{w(y,u(t,y))<p\}$ to $\{w(y,u(t,y))>p\}$. In general, no simplification occurs. However, when deriving a kinetic formulation for (1.1), we are interested only in the computation of $\nabla_y\mathbf{1}_{u>v(y,p)}$ in the sense of distributions on $[0,\infty)\times Y\times\mathbb{R}_p$ (see, for instance, [14], [13], and [17, section 3.2]). In that case, we can give another expression for the gradient of $\mathbf{1}_{u>v(y,p)}$, namely,

$$\begin{aligned}\nabla_y \mathbf{1}_{u(t,y)>v(y,p)} &= \nabla_y \mathbf{1}_{w(y,u(t,y))>p}\\ &= \nabla_y \left(w(y,u(t,y))\right)\delta(p=w(y,u(t,y)))\\ &= \frac{1}{\frac{\partial v}{\partial p}(y,p)}\nabla_y(u(t,y)-v(y,p))\delta(p=w(y,u(t,y))).\end{aligned}$$

Notice that the above expression, although meaningless if considered for $p\in\mathbb{R}$ fixed, is well-defined in the sense of distributions on $[0,\infty)\times Y\times\mathbb{R}_p$.

Thus (2.2) is proved. Consequently, all solutions of (1.1) satisfy inequality (1.10) in the sense of distributions on $[0,\infty)\times Y\times\mathbb{R}$. And it is then easily checked that if a solution u of (1.1) satisfies (1.10) in the sense of distributions in t,y,p, then u satisfies (1.10) for all p in the sense of distributions in t, y. $\square$

3. An application to homogenization. We provide here a proof for Proposition 1.3. The kinetic formulation derived above allows a better understanding of the homogenization process, and the proof is much more elegant than the original one in [3], which used two-scale Young measures.

We will work in the context of kinetic solutions of (1.11): Let $\varepsilon>0$, and let $u^\varepsilon\in L^\infty_{\text{loc}}([0,\infty);L^1_{\text{loc}}(\mathbb{R}^N))\cap L^2_{\text{loc}}(0,+\infty;H^1_{\text{loc}}(\mathbb{R}^N))$. We assume that

$$f^\varepsilon(t,x,p):=\mathbf{1}_{v\left(\frac{x}{\varepsilon},p\right)<u^\varepsilon(t,x)}$$

is a solution in the sense of distributions of

$$\begin{aligned}&\partial_t\left(v_p\left(\frac{x}{\varepsilon},p\right)f^\varepsilon\right)+\partial_{x_i}\left[a_i\left(\frac{x}{\varepsilon},v\left(\frac{x}{\varepsilon},p\right)\right)v_p\left(\frac{x}{\varepsilon},p\right)f^\varepsilon\right]-\varepsilon\Delta_x\left(v_p\left(\frac{x}{\varepsilon},p\right)f^\varepsilon\right)=\partial_p m^\varepsilon,\\ &f^\varepsilon(t=0)=\mathbf{1}_{v\left(\frac{x}{\varepsilon},p\right)<u_0\left(x,\frac{x}{\varepsilon}\right)},\end{aligned}\tag{3.1}$$

where

$$m^\varepsilon(t,x,p):=\varepsilon\left|\nabla_x u^\varepsilon(t,x)-\nabla_y v\left(\frac{x}{\varepsilon},p\right)\right|^2\frac{1}{v_p\left(\frac{x}{\varepsilon},p\right)}\delta\left(p=w\left(\frac{x}{\varepsilon},u^\varepsilon(t,x)\right)\right).$$

We assume that the hypotheses of Proposition 1.1 are satisfied, together with (2.1), so that $w(y,p)$ is well-defined (see Lemma 2.1). We have used the notation $v_p(y,p)=\partial_p v(y,p)$.

The hypotheses on f^ε are the following:

(H1) $u_0(x,y)=v(y,\bar u_0(x))$ for some $\bar u_0\in L^1(\mathbb{R}^N)$.

(H2) $u_0-v(y,0)\in L^1(\mathbb{R}^N,\mathcal{C}_{\text{per}}(Y))$; this means that

$$\int_{\mathbb{R}^N}\sup_{y\in Y}|v(y,\bar u_0(x))-v(y,0)|\,dx<+\infty,$$

which is slightly stronger than requiring $\bar u_0\in L^1$.

(H3) $f^\varepsilon(t,x,p)\to 0$ (resp., $1-f^\varepsilon\to 0$) as $p\to+\infty$ (resp., as $p\to-\infty$) for a.e. $(t,x)\in[0,\infty)\times\mathbb{R}^N$ and for all $\varepsilon>0$. Equivalently,

$$\alpha_-\left(\frac{x}{\varepsilon}\right)<u^\varepsilon(t,x)<\alpha_+\left(\frac{x}{\varepsilon}\right)\quad\text{for a.e. }(t,x)\in(0,\infty)\times\mathbb{R}^N,$$

where α_- and α_+ were defined in Lemma 2.1.

(H4) For all $\varepsilon > 0$, there exists a function $\mu_\varepsilon \in L^\infty(\mathbb{R})$ such that $\mu_\varepsilon(p) \to 0$ as $|p| \to \infty$ and

$$\int_0^\infty \int_{\mathbb{R}^N} m^\varepsilon(t,x,p)\,dt\,dx \le \mu_\varepsilon(p) \quad \forall p \in \mathbb{R}.$$

(H5) For all $\varepsilon > 0$, the function

$$(t,x,p) \mapsto \frac{\partial v}{\partial p}\left(\frac{x}{\varepsilon}, p\right)\left[\mathbf{1}_{p>0} f^\varepsilon(t,x,p) + \mathbf{1}_{p<0}(1 - f^\varepsilon(t,x,p))\right]$$

belongs to $L^\infty_{\mathrm{loc}}([0,\infty), L^1(\mathbb{R}^{N+1}))$. Equivalently, the function

$$(t,x) \mapsto u^\varepsilon(t,x) - v\left(\frac{x}{\varepsilon}, 0\right)$$

belongs to $L^\infty_{\mathrm{loc}}([0,\infty), L^1(\mathbb{R}^N))$.

A function $u^\varepsilon \in L^\infty_{\mathrm{loc}}([0,\infty); L^1_{\mathrm{loc}}(\mathbb{R}^N)) \cap L^2_{\mathrm{loc}}(0,+\infty; H^1_{\mathrm{loc}}(\mathbb{R}^N))$ such that f^ε is a solution of (3.1) and such that (H3)–(H5) are satisfied is called a *kinetic solution* of the parabolic scalar conservation law (1.11). Notice that we do not assume that (1.11) is satisfied in the sense of distributions.

Let us now state the result we prove in this section.

THEOREM 3.1. *Assume that A satisfies the hypotheses of Proposition* 1.1 *and* (2.1)*. Let $u^\varepsilon \in L^\infty_{loc}([0,\infty); L^1_{loc}(\mathbb{R}^N)) \cap L^2_{loc}(0,+\infty; H^1_{loc}(\mathbb{R}^N))$ be a kinetic solution of* (1.11) *such that hypotheses* (H1)–(H5) *are satisfied. Then*

$$u^\varepsilon(t,x) - v\left(\frac{x}{\varepsilon}, \bar{u}(t,x)\right) \to 0$$

in $L^1_{loc}([0,\infty) \times \mathbb{R}^N)$, where $\bar{u} \in L^\infty([0,\infty), L^1(\mathbb{R}^N))$ is the kinetic solution of (1.15) *with initial data $\bar{u}_0$.*

REMARK 3.1. *When hypothesis* (H1) *on the microscopic profile of the initial data is not satisfied, it is proved in the L^∞ case in* [4] *that there is an initial layer of typical size ε, during which the solution adapts itself to the profile dictated by the microscopic structure. The proof of this result relies on the parabolic structure of the equation, which cannot be used here since the kinetic formulation is essentially a hyperbolic tool.*

REMARK 3.2. *It can be checked that* (H2)–(H5) *are always satisfied when $\bar{u}_0 \in L^\infty \cap L^1(\mathbb{R}^N)$ and $u^\varepsilon \in L^\infty_{loc}$ is an entropy solution. However, we wish to stress that hypothesis* (H3) *does not imply that $u^\varepsilon \in L^\infty_{\mathrm{loc}}([0,\infty) \times \mathbb{R}^N)$ in general. For instance, in the case when hypothesis* (1.4) *is satisfied, we have $\alpha_\pm = \pm\infty$, as explained in Remark* 2.1*. Hence in that case, hypothesis* (H3) *is always satisfied, and the only bound required on u^ε is* (H5)*, which is an L^1 bound. Consequently, we refer to* (H2)–(H5) *as an "L^1 setting," in contrast with the "L^∞ setting" of entropy solutions.*

At last, let us mention that the function μ_ε in hypothesis (H4) *can in fact be derived from* (3.1) *(see Lemma* 3.2 *below) if it is known that* (H4) *holds for some function μ_ε; nonetheless,* (H4) *cannot be avoided and is necessary for Lemma* 3.2 *to hold.*

We will prove the convergence in several steps; first, we introduce the two-scale weak limit $f(t,x,y,p)$ of f^ε. Then the key point in the analysis is to show that $f(t,x,y,p) = \mathbf{1}_{p<\bar{u}(t,x)}$, where $\bar{u}$ is the solution of the homogenized problem. Hence, we first prove that f does not depend on y. Then we derive the macroscopic equation solved by f and prove that $f(t=0) = \mathbf{1}_{p<\bar{u}_0(x)}$; this entails that $f = \mathbf{1}_{p<\bar{u}}$, and $\bar{u}$ can be identified thanks to the equation solved by f. Eventually, we prove the strong convergence in L^1_{loc}.

We begin with a few preliminary bounds on m^ε and f^ε, of which we give only a rough idea of the proof (see, for instance, [17, Proposition 4.1.7 and Lemma 3.1.7] for the derivation of similar inequalities).

LEMMA 3.2. *Assume that* (H1)–(H5) *are satisfied.*

- *There exists a constant $C > 0$ such that for all $\varepsilon > 0$, for a.e. $t > 0$,*

$$\int_{\mathbb{R}^{N+1}} v_p\left(\frac{x}{\varepsilon}, p\right)\left(\mathbf{1}_{p>0} f^\varepsilon(t,x,p) + \mathbf{1}_{p<0}(1-f^\varepsilon)(t,x,p)\right)\, dx\, dp \le C.$$

- *There exists a constant $C > 0$ such that for all $p_0 > 0$, $\varepsilon > 0$,*

$$\int_0^\infty \int_{\mathbb{R}^N} m^\varepsilon(t,x,p_0)\, dx\, dt \le \int_{\mathbb{R}^N} \left(v\left(\frac{x}{\varepsilon}, \bar{u}_0(x)\right) - v\left(\frac{x}{\varepsilon}, p_0\right)\right)_+ dx \le C.$$

The same kind of bound holds for $p_0 < 0$.
Thus $m^\varepsilon((0,+\infty)\times\mathbb{R}^N\times(-R,R))$ is bounded for all $R > 0$ uniformly in ε.

- *For all $t \ge 0$, for all $p_0 > 0$, and for all $\varepsilon > 0$,*

$$\int_{\mathbb{R}^N} \left(u^\varepsilon(t,x) - v\left(\frac{x}{\varepsilon}, p_0\right)\right)_+ dx \le \int_{\mathbb{R}^N} \left(v\left(\frac{x}{\varepsilon}, \bar{u}_0(x)\right) - v\left(\frac{x}{\varepsilon}, p_0\right)\right)_+ dx. \tag{3.2}$$

We deduce from the second bound in the lemma that we can take in (H4)

$$\begin{aligned}\mu_\varepsilon(p) := \mathbf{1}_{p>0} &\int_{\mathbb{R}^N} \left(v\left(\frac{x}{\varepsilon}, \bar{u}_0(x)\right) - v\left(\frac{x}{\varepsilon}, p\right)\right)_+ dx \\ &+ \mathbf{1}_{p<0} \int_{\mathbb{R}^N} \left(v\left(\frac{x}{\varepsilon}, \bar{u}_0(x)\right) - v\left(\frac{x}{\varepsilon}, p\right)\right)_- dx.\end{aligned}$$

Then μ_ε is bounded in L^∞, uniformly in ε. Moreover, it will be proved in the very last step of the proof that for all p, $\mu_\varepsilon(p)$ converges as $\varepsilon \to 0$ toward $\mu_0(p)$ for some function $\mu_0 \in L^\infty(\mathbb{R})$ vanishing at infinity.

Proof. Thanks to the integrability assumptions (H4)–(H5) on f^ε and m^ε, we prove that for any test function $S' \in \mathcal{D}(\mathbb{R})$, for all $t > 0$, we have

$$\begin{aligned}\int_{\mathbb{R}^{N+1}} S'(p) f^\varepsilon(t,x,p) v_p\left(\frac{x}{\varepsilon}, p\right) dx\, dp - \int_{\mathbb{R}^{N+1}} S'(p) f^\varepsilon(t=0,x,p) v_p\left(\frac{x}{\varepsilon}, p\right) dx\, dp \\ = -\int_0^t \int_{\mathbb{R}^{N+1}} m^\varepsilon(t,x,p) S''(p)\, dt\, dx\, dp.\end{aligned}$$

Then, using the fact that μ_ε vanishes at infinity, we prove that the above equality holds for more general functions S. In particular, the choice $S'(p) = \mathbf{1}_{p>0}$ (and thus $S''(p) = \delta(p=0)$) yields the first bound on f^ε, and the choice $S'(p) = \mathbf{1}_{p>p_0}$ for some $p_0 > 0$ gives the one on m^ε. Moreover

$$\begin{aligned}\int_{\mathbb{R}^{N+1}} \mathbf{1}_{p>p_0} f^\varepsilon(t,x,p) v_p\left(\frac{x}{\varepsilon}, p\right) dx\, dp &= \int_{\mathbb{R}^{N+1}} \mathbf{1}_{v\left(\frac{x}{\varepsilon}, p_0\right) < v\left(\frac{x}{\varepsilon}, p\right) < u^\varepsilon(t,x)} v_p\left(\frac{x}{\varepsilon}, p\right) dx\, dp \\ &= \int_{\mathbb{R}^N} \left[u^\varepsilon(t,x) - v\left(\frac{x}{\varepsilon}, p_0\right)\right]_+ dx,\end{aligned}$$

and thus the choice $S'(p) = \mathbf{1}_{p>p_0}$ also yields the bound on u^ε. ∎

We now use the concept of two-scale convergence, defined by Allaire in [1] following an idea of Nguetseng (see [15]), in order to find a two-scale limit for f^ε.

PROPOSITION 3.3. *Let $\{v^\varepsilon\}_{\varepsilon>0}$ be a bounded sequence of $L^2(\Omega)$, where Ω is an open set of $\mathbb{R}^N$. Then as $\varepsilon \to 0$, there exists a subsequence, still denoted by ε, and $v^0 \in L^2(\Omega \times Y)$, such that*

$$\int_\Omega \psi\left(x, \frac{x}{\varepsilon}\right) v^\varepsilon(x)\, dx \to \int_{\Omega\times Y} \psi(x,y) v^0(x,y)\, dx\, dy$$

for all $\psi \in \mathcal{C}_{per}(Y, L^2(\Omega))$.

It is then said that the sequence $\{v^\varepsilon\}_{\varepsilon>0}$ "two-scale" converges to v_0.

This concept is easily generalized to functions in L^∞ (the proof goes along the same lines as the one given in [1]), which allows us to prove the following.

LEMMA 3.4. *There exists a function $f(t,x,y,p) \in L^\infty((0,\infty)\times\mathbb{R}^N \times Y \times \mathbb{R})$ and a subsequence, still denoted by ε, such that f^ε two-scale converges to f.*

It is easily checked that $0 \le f \le 1$ almost everywhere. Since v_p, f, and $1-f$ are nonnegative, Lemma 3.2 entails that there exists a constant C such that

$$\int_{\mathbb{R}^N\times Y\times\mathbb{R}} \{\mathbf{1}_{p>0} f(t,x,y,p) + \mathbf{1}_{p<0}(1-f(t,x,y,p))\}\, v_p(y,p) dx dy dp \le C \quad \text{for a.e. } t>0.$$

The goal is now to identify the equations solved by f in order to prove that f is an indicator function. Hence, we now focus on the microscopic (i.e., in y) and macroscopic (i.e., in t,x) equations solved by f.

First step. Microscopic profile. Multiplying (3.1) by a test function of the form $\varepsilon\varphi(t,x,x/\varepsilon,p)$, with $\varphi \in \mathcal{D}_{\rm per}((0,\infty)\times\mathbb{R}^N\times Y\times\mathbb{R})$, and passing to the limit as $\varepsilon \to 0$ leads to the equation

$$-\Delta_y\left(\frac{\partial v}{\partial p} f\right) + \mathrm{div}_y\left(a(y,v(y,p))\frac{\partial v}{\partial p} f\right) = 0 \tag{3.3}$$

in the sense of distributions on $(0,\infty)\times\mathbb{R}^N\times Y\times\mathbb{R}$. Let us point out that $a(y,v(y,p))$ is an "admissible" test function in the sense of Allaire (see [1]) thanks to the continuity assumption (2.1).

Then we regularize (3.3) in the variables t,x,y,p thanks to a convolution kernel, and pass to the limit as the parameter of the regularization vanishes. We easily deduce that (3.3) in fact holds almost everywhere in t,x,p in the variational sense in y.

Notice that the constant function equal to 1 on Y, denoted by $\bar{1}$, is a positive solution of the dual problem

$$-\Delta_y \bar{1} - a(y,v(y,p))\cdot\nabla_y \bar{1} = 0.$$

Consequently, by the Krein–Rutman theorem, we infer that any solution g of the equation

$$-\Delta_y g + \mathrm{div}_y\left(a(y,v(y,p))g\right) = 0$$

can be written $g(y) = c\frac{\partial v(y,p)}{\partial p}$, where c is a constant in y (recall (1.9)).

Thus $f(t,x,y,p)$ does not depend on y, and $f = f(t,x,p)$.

Second step. Evolution equation. Now, we multiply (3.1) by a test function of the form $\varphi(t,x,p)$, with $\varphi(t,x,p) = 0$ when $|p| \ge R$, $R>0$ arbitrary; thanks to

Lemma 3.2, $m^\varepsilon((0,\infty)\times\mathbb{R}^N\times(-R,R))$ is bounded uniformly in ε, and thus up to the extraction of a subsequence, there exists a measure $\bar m_R$ such that

$$m^\varepsilon \rightharpoonup \bar m_R \quad \text{in } w-M^1((0,\infty)\times\mathbb{R}^N\times(-R,R)).$$

We define, for any $p\in\mathbb{R}$,

$$\bar a(p)=\frac{1}{|Y|}\int_Y a(y,v(y,p))\frac{\partial v}{\partial p}\,dy;$$

recall also that

$$\frac{1}{|Y|}\int_Y \frac{\partial v}{\partial p}\,dy=1.$$

Then f satisfies, in the sense of distributions on $(0,\infty)\times\mathbb{R}^N\times(-R,R)$,

$$\partial_t f+\operatorname{div}_x(\bar a(p)f)=\frac{\partial \bar m_R}{\partial p}. \tag{3.4}$$

We deduce that for any $0<R<R'$, $\bar m_R=\bar m_{R'}$ on $(0,\infty)\times\mathbb{R}^N\times(-R,R)$. Consequently, the measure $\bar m$, defined by $\bar m=\bar m_R$ on $(0,\infty)\times\mathbb{R}^N\times(-R,R)$, is well-defined. Hence (3.4) holds in $(0,\infty)\times\mathbb{R}^{N+1}$ with $\bar m_R$ replaced by $\bar m$, and $\bar m\in\mathcal{C}(\mathbb{R}_p,w-M^1([0,\infty)\times\mathbb{R}^N_x))$. Moreover the measure $\bar m$ inherits the following property from the bounds on m^ε: for almost every $p\in\mathbb{R}$,

$$\int_0^\infty\int_{\mathbb{R}^N\times Y}\bar m(t,x,y,p)\,dt\,dx\,dy\le\mu_0(p), \tag{3.5}$$

and μ_0 belongs to L^∞ and vanishes at infinity, as we shall prove in the fourth step.

Equation (3.4) looks very much like the kinetic formulation for a homogeneous and hyperbolic scalar conservation law (see, for instance, [14], [13], and [17, Chapter 3]). However, we have to work out a few points before coming to a conclusion.

Third step. Identification of f as an indicator function. First, the function which occurs in the kinetic formulation is the function $\chi:\mathbb{R}^2\to\{1,-1,0\}$ defined by

$$\chi(v,u):=\begin{cases}1 & \text{if } 0<v<u,\\ -1 & \text{if } u<v<0,\\ 0 & \text{otherwise.}\end{cases}$$

Here, if $u^\varepsilon(t,x)-v\left(x/\varepsilon,\bar u(t,x)\right)$ converges strongly to 0, as we intend to prove, then $f=\mathbf{1}_{v(y,p)<v(y,\bar u(t,x))}=\mathbf{1}_{p<\bar u(t,x)}$; hence, a good candidate for a function $\chi(v,\bar u(t,x))$ seems to be

$$g(t,x,p)=\mathbf{1}_{p>0}f-\mathbf{1}_{p<0}(1-f)=f-\mathbf{1}_{p<0}.$$

The function g satisfies the same equation as f, and

$$\operatorname{sgn}(p)g=\mathbf{1}_{p>0}f+\mathbf{1}_{p<0}(1-f)=|g|\in[0,1].$$

Moreover,

$$\frac{\partial g}{\partial p}=\delta(p=0)+\partial_p f. \tag{3.6}$$

Recall that

$$\partial_p f^\varepsilon(t,x,p) = -\delta\left(p - w\left(\frac{x}{\varepsilon}, u^\varepsilon(t,x)\right)\right);$$

Hence $-\partial_p f^\varepsilon(t,x,p)$ is a nonnegative measure, uniformly bounded in ε on compact sets of $(0,\infty)\times\mathbb{R}^{N+1}$. Since $\partial_p f^\varepsilon$ weakly converges to $\partial_p f$, we deduce that $\partial_p f$ is a nonpositive locally finite measure.

It remains to check that

$$g(t=0,x,p) = \chi(p,\bar{u}_0(x)); \tag{3.7}$$

this equality is in fact not obvious: if $f^\varepsilon(t=0,x,p) = f_0\left(x, x/\varepsilon, p\right)$, then it is false in general that $f(t=0,x,y,p) = f_0(x,y,p)$. Indeed, there might be initial layers of typical size ε. These are not taken into account when passing to the two-scale limit because the test functions do not select the microscopic information in time. In order to see the possible initial layers, we should have taken test functions of the kind

$$\psi\left(t, \frac{t}{\varepsilon}, x, \frac{x}{\varepsilon}, p\right).$$

Here, it is unnecessary to consider test functions which have microscopic oscillations in time because the initial data is well prepared. Hence, there is no initial layer in this case. In other words, the u^ε are uniformly continuous in time at time $t=0$ (with values in L^1_{loc}). In terms of the kinetic formulation, this result follows directly from the fact that

$$f^\varepsilon(t=0,x,p) = \mathbf{1}_{v\left(\frac{x}{\varepsilon},p\right) < v\left(\frac{x}{\varepsilon},\bar{u}_0(x)\right)} = \mathbf{1}_{p<\bar{u}_0(x)}.$$

Hence $f^\varepsilon(t=0)$ does not depend on ε. Consequently, multiplying (3.1) by a test function $\varphi(t,x,p)\in\mathcal{D}([0,\infty)\times\mathbb{R}^{N+1})$ yields

$$\begin{aligned}&\int_0^\infty\!\!\int_{\mathbb{R}^{N+1}} f^\varepsilon(t,x,p)\frac{\partial v}{\partial p}\left(\frac{x}{\varepsilon},p\right)\left\{\partial_t\varphi + a_i\left(\frac{x}{\varepsilon}, v\left(\frac{x}{\varepsilon},p\right)\right)\partial_{x_i}\varphi + \varepsilon\Delta_x\varphi\right\}\,dt\,dx\,dp\\ &= \int_0^\infty\!\!\int_{\mathbb{R}^{N+1}} m^\varepsilon(t,x,p)\partial_p\varphi(t,x,p)dtdxdp - \int_{\mathbb{R}^{N+1}}\frac{\partial v}{\partial p}\left(\frac{x}{\varepsilon},p\right)\mathbf{1}_{p<\bar{u}_0(x)}\varphi(t=0,x,p)dxdp.\end{aligned}$$

Passing to the limit as $\varepsilon\to 0$ entails that

$$f(t=0,x,p) = \mathbf{1}_{p<\bar{u}_0(x)},$$

and thus

$$g(t=0,x,p) = \chi(p,\bar{u}_0(x)).$$

Gathering (3.4), (3.5), (3.6), (3.7), we infer that g is a generalized kinetic solution (see Definition 4.1.2 in [17]) of the scalar conservation law

$$\frac{\partial u}{\partial t} + \frac{\partial \bar{A}_i(u)}{\partial x_i} = 0,$$

where

$$\bar{A}_i'(p) = \bar{a}_i(p).$$

Now, we can apply Theorem 4.3.1 in [17]: there exists $\bar{u}(t,x) \in L^\infty([0,\infty); L^1(\mathbb{R}^N))$ such that $g(t,x,p) = \chi(p, \bar{u}(t,x))$ almost everywhere, and $\bar{u}$ is a kinetic solution of the above scalar conservation law.

And since

$$\mathbf{1}_{p>0} f - \mathbf{1}_{p<0}(1-f) = \chi(p, \bar{u}(t,x)),$$

we deduce that

$$f(t,x,p) = \mathbf{1}_{p<\bar{u}(t,x)}$$

almost everywhere.

Fourth step. Strong convergence. Let us now prove that this result entails that

$$u^\varepsilon(t,x) - v\left(\frac{x}{\varepsilon}, \bar{u}(t,x)\right) \to 0$$

in L^1_{loc}.

(i) *Convergence of* $u^\varepsilon \wedge v(x/\varepsilon, p_0)$ *for all* $p_0 > 0$: Take an arbitrary cut-off function $\varphi = \varphi(t,x)$ with compact support in $[0,\infty) \times \mathbb{R}^N$, $p_0 > 0$ and set

$$\psi(t,x,y,p) := \mathbf{1}_{\bar{u}(t,x)<p<p_0} \frac{\partial v}{\partial p}(y,p)\, \varphi(t,x).$$

Since $f^\varepsilon(t,x,p) = \mathbf{1}_{v\left(\frac{x}{\varepsilon},p\right)<u^\varepsilon(t,x)}$ two-scale converges to $f = \mathbf{1}_{p<\bar{u}(t,x)}$, we deduce that as $\varepsilon \to 0$

$$\int_0^\infty \int_{\mathbb{R}^{N+1}} \psi\left(t,x,\frac{x}{\varepsilon},p\right) \mathbf{1}_{v\left(\frac{x}{\varepsilon},p\right)<u^\varepsilon(t,x)}\, dp\, dx\, dt \to 0.$$

And the left-hand side can be transformed as follows:

$$\begin{aligned}
&\int_0^\infty \int_{\mathbb{R}^{N+1}} \psi\left(t,x,\frac{x}{\varepsilon},p\right) \mathbf{1}_{v\left(\frac{x}{\varepsilon},p\right)<u^\varepsilon(t,x)}\, dp\, dx\, dt \\
&= \int_0^\infty \int_{\mathbb{R}^{N+1}} \mathbf{1}_{v\left(\frac{x}{\varepsilon},\bar{u}(t,x)\right)<v\left(\frac{x}{\varepsilon},p\right)<u^\varepsilon(t,x)\wedge v\left(\frac{x}{\varepsilon},p_0\right)}\, \varphi(t,x) \frac{\partial v}{\partial p}\left(\frac{x}{\varepsilon},p\right) dp\, dx\, dt \\
&= \int_0^\infty \int_{\mathbb{R}^{N+1}} \mathbf{1}_{v\left(\frac{x}{\varepsilon},\bar{u}(t,x)\right)<v<u^\varepsilon(t,x)\wedge v\left(\frac{x}{\varepsilon},p_0\right)}\, \varphi(t,x) dv\, dx\, dt \\
&= \int_0^\infty \int_{\mathbb{R}^N} \varphi(t,x) \left[u^\varepsilon(t,x) \wedge v\left(\frac{x}{\varepsilon},p_0\right) - v\left(\frac{x}{\varepsilon},\bar{u}(t,x)\right)\right]_+ dx\, dt.
\end{aligned}$$

Take any compact set $K \subset [0,\infty) \times \mathbb{R}^N$ and choose a test function $\varphi \in \mathcal{D}([0,\infty) \times \mathbb{R}^N)$ such that $0 \le \varphi \le 1$, and $\varphi \equiv 1$ on K. Then for all $\varepsilon > 0$,

$$\begin{aligned}
&\left\| \left[u^\varepsilon(t,x) \wedge v\left(\frac{x}{\varepsilon},p_0\right) - v\left(\frac{x}{\varepsilon},\bar{u}(t,x)\right)\right]_+ \right\|_{L^1(K)} \\
&\le \int_0^\infty \int_{\mathbb{R}^N} \varphi(t,x) \left[u^\varepsilon(t,x) \wedge v\left(\frac{x}{\varepsilon},p_0\right) - v\left(\frac{x}{\varepsilon},\bar{u}(t,x)\right)\right]_+ dx\, dt.
\end{aligned}$$

In the inequality above, we have used the fact that $u_+ = \max(u,0)$ is always nonnegative. Thus we deduce that for all $p_0 > 0$

$$\left\| \left[u^\varepsilon(t,x) \wedge v\left(\frac{x}{\varepsilon},p_0\right) - v\left(\frac{x}{\varepsilon},\bar{u}(t,x)\right)\right]_+ \right\|_{L^1_{\text{loc}}([0,\infty)\times\mathbb{R}^N)} \underset{\varepsilon\to 0}{\longrightarrow} 0.$$

The same kind of result holds for $p_0 < 0$.

(ii) *Convergence of* u^ε: Let $T > 0$, $R > 0$, and set $Q := (0,T) \times B_R$. For $p_0 > 0$ arbitrary, we have

$$\begin{aligned}
&\left\| \left[u^\varepsilon(t,x) - v\left(\frac{x}{\varepsilon}, \bar{u}(t,x)\right)\right]_+ \right\|_{L^1(Q)} \\
\le &\left\| \left[u^\varepsilon \wedge v\left(\frac{x}{\varepsilon}, p_0\right) - v\left(\frac{x}{\varepsilon}, \bar{u}(t,x)\right)\right]_+ \right\|_{L^1(Q)} + \left\| \left[u^\varepsilon - u^\varepsilon \wedge v\left(\frac{x}{\varepsilon}, p_0\right)\right]_+ \right\|_{L^1(Q)} \\
\le &\left\| \left[u^\varepsilon \wedge v\left(\frac{x}{\varepsilon}, p_0\right) - v\left(\frac{x}{\varepsilon}, \bar{u}(t,x)\right)\right]_+ \right\|_{L^1(Q)} + \left\| \left[u^\varepsilon - v\left(\frac{x}{\varepsilon}, p_0\right)\right]_+ \right\|_{L^1(Q)} \\
\le &\left\| \left[u^\varepsilon \wedge v\left(\frac{x}{\varepsilon}, p_0\right) - v\left(\frac{x}{\varepsilon}, \bar{u}(t,x)\right)\right]_+ \right\|_{L^1(Q)} + T\int_{\mathbb{R}^N} \left[v\left(\frac{x}{\varepsilon}, \bar{u}_0(x)\right) - v\left(\frac{x}{\varepsilon}, p_0\right)\right]_+ dx
\end{aligned}$$

thanks to inequality (3.2).

According to (H2), we have $[v(y,\bar{u}_0) - v(y,p_0)]_+ \in L^1(\mathbb{R}^N; \mathcal{C}_{\text{per}}(Y))$; thus, using a result of Allaire (see [1]), we deduce

$$\int_{\mathbb{R}^N} \left[v\left(\frac{x}{\varepsilon}, \bar{u}_0(x)\right) - v\left(\frac{x}{\varepsilon}, p_0\right)\right]_+ dx \to \int_{\mathbb{R}^N \times Y} \left[v\left(y, \bar{u}_0(x)\right) - v\left(y, p_0\right)\right]_+ dx\, dy$$

as $\varepsilon \to 0$ for all $p_0 > 0$. Since $||\left(v(y,p) - v(y,p')\right)_+||_{L^1(Y)} = (p-p')_+$ for all $p, p' \in \mathbb{R}$, we derive the bound

$$\begin{aligned}
&\int_0^T \int_{B_R} \left[u^\varepsilon(t,x) - v\left(\frac{x}{\varepsilon}, \bar{u}(t,x)\right)\right]_+ dx\, dt \\
\le &\int_0^T \int_{B_R} \left[u^\varepsilon(t,x) \wedge v\left(\frac{x}{\varepsilon}, p_0\right) - v\left(\frac{x}{\varepsilon}, \bar{u}(t,x)\right)\right]_+ dx\, dt \\
&+ T \left| \int_{\mathbb{R}^N} \left[v\left(\frac{x}{\varepsilon}, \bar{u}_0(x)\right) - v\left(\frac{x}{\varepsilon}, p_0\right)\right]_+ dx - ||(\bar{u}_0 - p_0)_+||_{L^1(\mathbb{R}^N)} \right| \\
&+ T||(\bar{u}_0 - p_0)_+||_{L^1(\mathbb{R}^N)}.
\end{aligned}$$

In the above inequality, take p_0 large enough so that $||(\bar{u}_0 - p_0)_+||_{L^1(\mathbb{R}^N)}$ is small enough, and then for this p_0, take $\varepsilon > 0$ small enough so that the two other terms are small (notice that the first one vanishes thanks to the first step). We deduce that

$$\left[u^\varepsilon(t,x) - v\left(\frac{x}{\varepsilon}, \bar{u}(t,x)\right)\right]_+ \to 0$$

in $L^1_{\text{loc}}([0,\infty) \times \mathbb{R}^N)$, and Theorem 3.1 is proved.

Moreover, we have proved that for all $p > 0$,

$$\begin{aligned}
\lim_{\varepsilon \to 0} \mu_\varepsilon(p) &= \lim_{\varepsilon \to 0} \int_{\mathbb{R}^N} \left[v\left(\frac{x}{\varepsilon}, \bar{u}_0(x)\right) - v\left(\frac{x}{\varepsilon}, p\right)\right]_+ dx \\
&= \int_{\mathbb{R}^N \times Y} \left[v\left(y, \bar{u}_0(x)\right) - v\left(y, p\right)\right]_+ dx\, dy \\
&= ||(\bar{u}_0 - p)_+||_{L^1(\mathbb{R}^N)} =: \mu_0(p).
\end{aligned}$$

Thus μ_0 vanishes at infinity, and the result stated after Lemma 3.2 holds.

4. Rigorous proof of the L^1 contraction principle. This section is devoted to the proof of inequality (2.12) under assumption (2.11) and the hypotheses of Theorem 2.5. The main ideas behind the proof were exposed in the formal calculations of section 2; however, regularizations are necessary in order to justify nonlinear manipulations of the type

$$f_1\partial_t f_2 + f_2\partial_t f_1 = \partial_t(f_1 f_2),$$

as well as the reduction of the right-hand side.

As in [16], [17, Chapter 4], we will merely regularize the equation by convolution; let $\varepsilon > 0$ be a small parameter, $\zeta_1 \in \mathcal{D}(\mathbb{R})$, $\zeta_2 \in \mathcal{D}(\mathbb{R}^N)$, $\zeta_3 \in \mathcal{D}(\mathbb{R})$, such that

$$\zeta_i \geq 0 \quad (i = 1, 2, 3),$$
$$\operatorname{supp} \zeta_1 \subset [-1, 0], \quad \operatorname{supp} \zeta_2 \subset B_1, \quad \operatorname{supp} \zeta_3 \subset [-1, 1], \quad \zeta_1(0) = 0,$$
$$\int_{\mathbb{R}} \zeta_1 = \int_{\mathbb{R}^N} \zeta_2 = \int_{\mathbb{R}} \zeta_3 = 1.$$

We set, for $\varepsilon > 0$, $(t, x, p) \in \mathbb{R}^{N+2}$,

$$\phi_\varepsilon(t, x, p) := \frac{1}{\varepsilon^{N+2}}\zeta_1\left(\frac{t}{\varepsilon}\right)\zeta_2\left(\frac{x}{\varepsilon}\right)\zeta_3\left(\frac{p}{\varepsilon}\right),$$

and for $(t, x, p) \in [0, \infty) \times \mathbb{R}^N \times \mathbb{R}$,

$$f_i^\varepsilon(t, x, p) = \int_{\mathbb{R}^{N+2}} f_i(s, z, q)\phi_\varepsilon(t - s, x - z, p - q)\, ds\, dz\, dq,$$
$$m_i^\varepsilon(t, x, p) = \int_{\mathbb{R}^{N+2}} m_i(s, z, q)\phi_\varepsilon(t - s, x - z, p - q)\, ds\, dz\, dq.$$

(Notice that the convolution in the space variable x is meant in the whole of $\mathbb{R}^N$: f_i is thus considered as a function defined on $[0, \infty) \times \mathbb{R}^N \times \mathbb{R}$, periodic with period Y in its second variable. The function f_i^ε is of course Y-periodic as well.)

We begin with the derivation of the equation solved by f^ε.

LEMMA 4.1. *Set* $\tilde{a}_i(y, p) = a_i(y, v(y, p))\frac{\partial v(y,p)}{\partial p}$ *for* $1 \leq i \leq N$, $y \in Y$, $p \in \mathbb{R}$.

Then for $\varepsilon < 1/2$, f_j^ε $(j = 1, 2)$ *is a classical solution of*

$$\frac{\partial}{\partial t}\left(\frac{\partial v}{\partial p} f_j^\varepsilon\right) + \frac{\partial}{\partial y_i}\left(\tilde{a}_i(y, p) f_j^\varepsilon\right) - \Delta_y\left(\frac{\partial v}{\partial p} f_j^\varepsilon\right) = \eta_j\frac{\partial m_j^\varepsilon}{\partial p} + r_j^\varepsilon, \tag{4.1}$$

where $\eta_1 = 1$, $\eta_2 = -1$, *and the error term* r_j^ε *is equal to*

$$\begin{aligned} r_j^\varepsilon(t, y, p) = {} & \frac{\partial}{\partial t}\left[\frac{\partial v}{\partial p}(y, p) f_j^\varepsilon(t, y, p) - \left(\frac{\partial v}{\partial p} f_j\right) * \phi_\varepsilon(t, y, p)\right] \\ & + \frac{\partial}{\partial y_i}\left[\tilde{a}_i(y, p) f_j^\varepsilon(t, y, p) - (\tilde{a}_i f_j) * \phi_\varepsilon(t, y, p)\right] \\ & - \Delta_y\left[\frac{\partial v}{\partial p}(y, p) f_j^\varepsilon(t, y, p) - \left(\frac{\partial v}{\partial p} f_j\right) * \phi_\varepsilon(t, y, p)\right]. \end{aligned}$$

Moreover, for all $0 < \varepsilon < 1/2$, *for all* $p \in \mathbb{R}$,

$$\int_0^\infty \int_Y m_i^\varepsilon(t, y, p)\, dt\, dy \leq \max(\mu_i(p + 1), \mu_i(p - 1)),$$

where the functions μ_i *were introduced in hypothesis* (2.5) *in Definition* 2.3.

We postpone the proof of Lemma 4.1 until the end of the section.

Now, since f_j^ε is smooth we can multiply (4.1) written for f_1^ε (resp., f_2^ε) by f_2^ε (resp., f_1^ε) and add the two equations thus obtained. Following the calculations in section 2 leads to

$$\begin{aligned}\frac{\partial}{\partial t}\left(\frac{\partial v}{\partial p}f_1^\varepsilon f_2^\varepsilon\right) + \frac{\partial}{\partial y_i}\left(\tilde{a}_i(y,p)f_1^\varepsilon f_2^\varepsilon\right) - \Delta_y\left(\frac{\partial v}{\partial p}f_1^\varepsilon f_2^\varepsilon\right)\\ = \frac{\partial m_1^\varepsilon}{\partial p}f_2^\varepsilon - \frac{\partial m_2^\varepsilon}{\partial p}f_1^\varepsilon - 2\frac{\partial v}{\partial p}\nabla_y f_1^\varepsilon \cdot \nabla_y f_2^\varepsilon + r_1^\varepsilon f_2^\varepsilon + r_2^\varepsilon f_1^\varepsilon.\end{aligned}$$

Let $R>0$ arbitrary, and let $K_R \in \mathcal{D}(\mathbb{R})$ be a cut-off function such that

$$\begin{gathered}0 \le K_R(p) \le 1, \quad |K_R'(p)| \le 2 \quad \forall p \in \mathbb{R},\\ K_R(p) = 1 \quad \forall p \in [-R,R],\\ K_R(p) = 0 \quad \forall p \in (-\infty,-R-1] \cup [R+1,+\infty).\end{gathered}$$

Classically, the following convergence results hold for any test function $\theta = \theta(t,y) \in \mathcal{D}_{\text{per}}([0,\infty)\times Y)$ (recall (2.11)):

$$\lim_{R\to\infty}\lim_{\varepsilon\to 0}\int_0^\infty\int_{Y\times\mathbb{R}}\frac{\partial v}{\partial p}(y,p)f_1^\varepsilon f_2^\varepsilon \theta(t,y)K_R(p)\,dt\,dy\,dp = \int_Y (u_1-u_2)_+\theta(t,y)\,dt\,dy,$$

$$\begin{aligned}\lim_{R\to\infty}\lim_{\varepsilon\to 0}\int_0^\infty\int_{Y\times\mathbb{R}}\tilde{a}_i(y,p)f_1^\varepsilon f_2^\varepsilon \partial_{y_i}\theta(t,y)K_R(p)\,dt\,dy\,dp\\ = \int_0^\infty\int_Y \mathbf{1}_{u_1>u_2}\left[A_i(y,u_1) - A_i(y,u_2)\right]\partial_{y_i}\theta(t,y)\,dt\,dy.\end{aligned}$$

(If one is interested in deriving (2.10), without assumption (2.11), instead of (2.12), one should merely take $\theta \in \mathcal{D}([0,\infty))$, independent of y, at this stage; the left-hand side in the second equality above is zero in that case. The rest of the proof remains unchanged.)

On the other hand, it is easily proved that the first order terms in r_j^ε go to 0 in $L^1_{\text{loc}}((0,\infty)\times\mathbb{R}^{N+1})$ as $\varepsilon \to 0$ thanks to the assumption $a \in W^{1,1}_{\text{loc}}$. Hence, we now focus on the convergence of

$$\frac{\partial m_1^\varepsilon}{\partial p}f_2^\varepsilon - \frac{\partial m_2^\varepsilon}{\partial p}f_1^\varepsilon - 2\frac{\partial v}{\partial p}\nabla_y f_1^\varepsilon \cdot \nabla_y f_2^\varepsilon$$

and the second order terms in r_j^ε, that is,

$$\begin{aligned}-\Delta_y\left[\frac{\partial v}{\partial p}(y,p)f_1^\varepsilon(t,y,p) - \left(\frac{\partial v}{\partial p}f_1\right) * \phi_\varepsilon(t,y,p)\right]f_2^\varepsilon,\\ -\Delta_y\left[\frac{\partial v}{\partial p}(y,p)f_2^\varepsilon(t,y,p) - \left(\frac{\partial v}{\partial p}f_2\right) * \phi_\varepsilon(t,y,p)\right]f_1^\varepsilon.\end{aligned}$$

In the following, we set

$$\varphi_i(t,y) = w(y,u_i(t,y)) \quad \text{(i.e., } v(y,\varphi_i(t,y)) = u_i(t,y)\text{)}, \tag{4.2}$$

$$\gamma_i(t,y) = \frac{1}{\frac{\partial v}{\partial p}(y,\varphi_i(t,y))}\left[\nabla_y u_i(t,y) - (\nabla_y v)(y,\varphi_i(t,y))\right] = \nabla_y \varphi_i(t,y). \tag{4.3}$$

We recall that

$$m_i(t,y,p) = |\nabla_y \varphi_i(t,y)|^2 \frac{\partial v}{\partial p}(y,\varphi_i(t,y))\; \delta(p=\varphi_i(t,y)) \tag{4.4}$$

$$= |\gamma_i|^2(t,y)\frac{\partial v}{\partial p}(y,\varphi_i(t,y))\; \delta(p=\varphi_i(t,y)), \tag{4.5}$$

$$\nabla_y f_i(t,y,p) = \eta_i \nabla_y \varphi_i(t,y)\delta(p=\varphi_i(t,y)) \tag{4.6}$$

$$= \eta_i \gamma_i(t,y)\delta(p=\varphi_i(t,y)), \tag{4.7}$$

$$\partial_p f_i = -\eta_i \delta(p=\varphi_i(t,y)) \tag{4.8}$$

for $i=1,2$, where $\eta_1 = 1$ and $\eta_2=-1$.

First, for any test function $\theta = \theta(t,y) \in \mathcal{D}_{\text{per}}([0,+\infty)\times Y)$ such that $\theta \geq 0$, for $\varepsilon < 1$, $R>1$, we claim that

$$\begin{aligned}
&\int_0^\infty \int_{Y\times\mathbb{R}} \left[\frac{\partial m_1^\varepsilon}{\partial p} f_2^\varepsilon - \frac{\partial m_2^\varepsilon}{\partial p} f_1^\varepsilon - 2\frac{\partial v}{\partial p}\nabla_y f_1^\varepsilon \cdot \nabla_y f_2^\varepsilon\right]\theta(t,y)K_R(p)\,dt\,dy\,dp\\
&\leq \int_0^\infty \int_{\mathbb{R}^{2N+2}} \int_{Y\times\mathbb{R}} \phi^\varepsilon(t-s_1,y-y_1,p-\varphi_1)\phi^\varepsilon(t-s_2,y-y_2,p-\varphi_2)\theta(t,y)K_R(p)\\
&\quad\times 2\left[\gamma_1\cdot\gamma_2\left(\frac{\partial v}{\partial p}(y,p) - \sqrt{\frac{\partial v}{\partial p}(y_1,\varphi_1)\frac{\partial v}{\partial p}(y_2,\varphi_2)}\right)\right]\,dy\,dp\,dy_1\,dy_2\,ds_1\,ds_2\,dt\\
&\quad + 2||\theta||_\infty\left[\mu_1(R-1)+\mu_1(-R+1)+\mu_2(R-1)+\mu_2(-R+1)\right].
\end{aligned} \tag{4.9}$$

In the integral of the right-hand side above, γ_i, φ_i are evaluated at (s_i,y_i) $(i=1,2)$.

The derivation of this inequality is rather technical but straightforward if one follows the formal calculations of section 2. Let us focus on the first term of the left-hand side, namely,

$$\begin{aligned}
I_\varepsilon &:= \int_0^\infty \int_{Y\times\mathbb{R}} \frac{\partial m_1^\varepsilon}{\partial p} f_2^\varepsilon \theta(t,y)K_R(p)\,dt\,dy\,dp\\
&= -\int_0^\infty \int_{Y\times\mathbb{R}} m_1^\varepsilon\;\partial_p f_2^\varepsilon\;\theta(t,y)K_R(p)\,dt\,dy\,dp - \int_0^\infty\int_{Y\times\mathbb{R}} m_1^\varepsilon f_2^\varepsilon\;\theta(t,y)\;K_R'(p)\,dt\,dy\,dp\\
&=: -(I_{\varepsilon,1}+I_{\varepsilon,2}).
\end{aligned}$$

Remembering (4.5) and (4.8), we have

$$m_1^\varepsilon(t,y,p) = \int_{\mathbb{R}^{N+1}} |\gamma_1(s_1,y_1)|^2 \frac{\partial v}{\partial p}(y_1,\varphi_1(s_1,y_1))\phi^\varepsilon(t-s_1,y-y_1,p-\varphi_1(s_1,y_1))\,ds_1\,dy_1,$$

$$\partial_p f_2^\varepsilon = \int_{\mathbb{R}^{N+1}} \phi^\varepsilon(t-s_2,y-y_2,p-\varphi_2(s_2,y_2))\,ds_2\,dy_2,$$

and thus

$$\begin{aligned}
I_{\varepsilon,1} &= \int_0^\infty \int_{\mathbb{R}^{2N+2}} \int_{Y\times\mathbb{R}} \phi^\varepsilon(t-s_1,y-y_1,p-\varphi_1)\phi^\varepsilon(t-s_2,y-y_2,p-\varphi_2)\theta(t,y)K_R(p)\\
&\qquad\times\left[\sqrt{\frac{\partial v}{\partial p}(y_1,\varphi_1(s_1,y_1))}\gamma_1(s_1,y_1)\right]^2\,dy\,dp\,dy_1\,dy_2\,ds_1\,ds_2\,dt.
\end{aligned}$$

On the other hand, according to Lemma 4.1 and the assumptions on K_R,

$$\begin{aligned}|I_{\varepsilon,2}| &\leq \int_0^\infty \int_{Y\times\mathbb{R}} m_1^\varepsilon \theta(t,y)|K_R'(p)|\,dt\,dy\,dp\\ &\leq 2||\theta||_\infty \left[\mu_1(R-1)+\mu_1(-R+1)\right].\end{aligned}$$

The two other terms are treated in a similar way; eventually, we use the inequality $-(|a|^2+|b|^2)\leq -2a\cdot b$ for all $a,b\in\mathbb{R}^N$ with $a=\sqrt{v_p(y_1,\varphi_1)}\,\gamma_1$, $b=\sqrt{v_p(y_2,\varphi_2)}\,\gamma_2$, and γ_i, φ_i are evaluated at $(s_i,y_i)\in[0,\infty)\times\mathbb{R}^N$. This concludes the derivation of (4.9).

Next, we investigate the second order terms in r_j^ε, i.e.,

$$\begin{aligned}\int_0^\infty \int_{\mathbb{R}^{N+1}} -\Bigg\{\Delta_y\left[\frac{\partial v}{\partial p}(y,p)f_1^\varepsilon(t,y,p)-\left(\frac{\partial v}{\partial p}f_1\right)*\phi_\varepsilon(t,y,p)\right]f_2^\varepsilon\\ +\Delta_y\left[\frac{\partial v}{\partial p}(y,p)f_2^\varepsilon(t,y,p)-\left(\frac{\partial v}{\partial p}f_2\right)*\phi_\varepsilon(t,y,p)\right]f_1^\varepsilon\Bigg\}\,\theta(t,y)K_R(p)\,dt\,dy\,dp.\end{aligned}$$

Integrating by parts, we obtain for instance

$$\begin{aligned}&\int_0^\infty \int_{Y\times\mathbb{R}} -\Delta_y\left[\frac{\partial v}{\partial p}f_1^\varepsilon-\left(\frac{\partial v}{\partial p}f_1\right)*\phi_\varepsilon\right]f_2^\varepsilon\ \theta(t,y)K_R(p)\,dt\,dy\,dp\\ =&\int_0^\infty \int_{Y\times\mathbb{R}}\left[\frac{\partial v}{\partial p}\nabla_y f_1^\varepsilon-\left(\frac{\partial v}{\partial p}\left(\nabla_y f_1\right)\right)*\phi_\varepsilon\right]\cdot\nabla_y f_2^\varepsilon\ \theta(t,y)K_R(p)\,dt\,dy\,dp\\ &+\int_0^\infty \int_{Y\times\mathbb{R}}\left[\nabla_y\frac{\partial v}{\partial p}f_1^\varepsilon-\left(\left(\nabla_y\frac{\partial v}{\partial p}\right)f_1\right)*\phi_\varepsilon\right]\cdot\nabla_y f_2^\varepsilon\ \theta(t,y)K_R(p)\,dt\,dy\,dp\\ &-\int_0^\infty \int_{Y\times\mathbb{R}}\left[\frac{\partial v}{\partial p}f_1^\varepsilon-\left(\frac{\partial v}{\partial p}f_1\right)*\phi_\varepsilon\right]\nabla_y f_2^\varepsilon\cdot\nabla\theta(t,y)K_R(p)\,dt\,dy\,dp\\ &-\int_0^\infty \int_{Y\times\mathbb{R}}\left[\frac{\partial v}{\partial p}f_1^\varepsilon-\left(\frac{\partial v}{\partial p}f_1\right)*\phi_\varepsilon\right]f_2^\varepsilon\Delta\theta(t,y)K_R(p)\,dt\,dy\,dp\\ =:&\ J_{\varepsilon,1}+J_{\varepsilon,2}+J_{\varepsilon,3}+J_{\varepsilon,4}.\end{aligned}$$

Notice that hypothesis (2.8) ensures that $\frac{\partial^2 v}{\partial p^2}$ exists and is Hölder continuous in y, with locally uniform bounds in p (see Theorem 8.24 in [9]), and hypothesis (2.9) entails that $\nabla_y\frac{\partial v}{\partial p}$ is Hölder continuous in y, with locally uniform bounds in p (see Theorem 8.32 in [9]).

Hence, $\frac{\partial v}{\partial p}$ belongs to $\mathcal{C}(\mathbb{R},\mathcal{C}^\alpha_{\text{per}}(Y))$ for some $\alpha\in(0,1)$. Moreover, thanks to classical H^1 bounds for elliptic equations, we deduce that $\nabla_y\frac{\partial v}{\partial p}$ belongs to $\mathcal{C}(\mathbb{R},L^2(Y))$. Together with identity (4.7), these regularity results easily entail that for all $R>0$,

$$\lim_{\varepsilon\to0} J_{\varepsilon,i}=0\quad\text{for } i=2,3,4.$$

In the following, we denote by $\omega_R:[0,\infty)\to[0,\infty)$ a function such that $\lim_{0^+}\omega_R=0$ and

$$|J_{\varepsilon,2}|+|J_{\varepsilon,3}|+|J_{\varepsilon,4}|\leq\omega_R(\varepsilon)\quad\forall R>0,\ \forall\varepsilon>0.$$

Without loss of generality, we can also assume that the first order terms in r_j^ε are bounded by $\omega_R(\varepsilon)$.

We now focus on the term $J_{\varepsilon,1}$; thanks to identity (4.7), we have

$$J_{\varepsilon,1} = -\int_0^\infty\int_{\mathbb{R}^{2N+2}}\int_{Y\times\mathbb{R}} \phi_\varepsilon(t-s_1, y-y_1, p-\varphi_1)\phi_\varepsilon(t-s_2, y-y_2, p-\varphi_2) \times \left[\frac{\partial v}{\partial p}(y,p) - \frac{\partial v}{\partial p}(y_1,\varphi_1)\right]\gamma_1\cdot\gamma_2\;\theta(t,y)\;K_R(p)\;dy\;dp\;dy_1\;dy_2\;ds_1\;ds_2\;dt,$$

and γ_i, φ_i are evaluated at t_i, s_i. Gathering all the terms, we infer

$$\begin{aligned}
&\int_0^\infty\int_{Y\times\mathbb{R}}\left[\frac{\partial m_1^\varepsilon}{\partial p}f_2^\varepsilon - \frac{\partial m_2^\varepsilon}{\partial p}f_1^\varepsilon - 2\frac{\partial v}{\partial p}\nabla_y f_1^\varepsilon\cdot\nabla_y f_2^\varepsilon + r_1^\varepsilon f_2^\varepsilon + r_2^\varepsilon f_1^\varepsilon\right]\theta K_R\\
&\le \int \phi_\varepsilon(t-s_1, y-y_1, p-\varphi_1)\phi_\varepsilon(t-s_2, y-y_2, p-\varphi_2)\theta(t,y)K_R(p)\gamma_1\cdot\gamma_2\\
&\quad\times\left(\frac{\partial v}{\partial p}(y_1,\varphi_1) + \frac{\partial v}{\partial p}(y_2,\varphi_2) - 2\sqrt{\frac{\partial v}{\partial p}(y_1,\varphi_1)\frac{\partial v}{\partial p}(y_2,\varphi_2)}\right)dy\;dp\;dy_1\;dy_2\;ds_1\;ds_2\;dt\\
&\quad + 2\|\theta\|_\infty\left[\mu_1(R-1)+\mu_1(-R+1)+\mu_2(R-1)+\mu_2(-R+1)\right] + C\omega_R(\varepsilon)\\
&\le \int_0^\infty\int_{\mathbb{R}^{2N+2}}\int_{Y\times\mathbb{R}}\phi_\varepsilon(t-s_1, y-y_1, p-\varphi_1)\phi_\varepsilon(t-s_2, y-y_2, p-\varphi_2)\theta(t,y)K_R(p)\\
&\quad\times\gamma_1\cdot\gamma_2\left[\sqrt{\frac{\partial v}{\partial p}(y_1,\varphi_1)} - \sqrt{\frac{\partial v}{\partial p}(y_2,\varphi_2)}\right]^2 dy\;dp\;dy_1\;dy_2\;ds_1\;ds_2\;dt\\
&\quad + 2\|\theta\|_\infty\left[\mu_1(R-1)+\mu_1(-R+1)+\mu_2(R-1)+\mu_2(-R+1)\right] + C\omega_R(\varepsilon).
\end{aligned}$$

The function $\frac{\partial v}{\partial p}$ belongs to $W^{1,\infty}_{\mathrm{per,loc}}(Y\times\mathbb{R})$ and is bounded away from zero on bounded subsets of $Y\times\mathbb{R}$. As a consequence, there exists a constant $C_R>0$, depending on R, such that for all $\varepsilon>0$,

$$\left|\sqrt{\frac{\partial v}{\partial p}(y_1,\varphi_1)} - \sqrt{\frac{\partial v}{\partial p}(y_2,\varphi_2)}\right| \le C_R\,\varepsilon$$

whenever $|y_1-y_2|\le 2\varepsilon$, $|\varphi_1-\varphi_2|\le 2\varepsilon$, and $|\varphi_1|,|\varphi_2|\le R$. We set

$$\phi(t,x,p) := \zeta_1(t)\,\zeta_2(x)\,\zeta_3(p).$$

Performing changes of variables in the integral on the right-hand side, we obtain

$$\begin{aligned}
&\int_0^\infty\int_{Y\times\mathbb{R}}\left[\frac{\partial m_1^\varepsilon}{\partial p}f_2^\varepsilon - \frac{\partial m_2^\varepsilon}{\partial p}f_1^\varepsilon - 2\frac{\partial v}{\partial p}\nabla_y f_1^\varepsilon\cdot\nabla_y f_2^\varepsilon + r_1^\varepsilon f_2^\varepsilon + r_2^\varepsilon f_1^\varepsilon\right]\theta K_R\\
&\le C_R\,\varepsilon^2\int_0^\infty\int_{\mathbb{R}^{2N+2}}\int_{Y\times\mathbb{R}}\phi_\varepsilon(t-s_1, y-y_1, p-\varphi_1)\phi_\varepsilon(t-s_2, y-y_2, p-\varphi_2)\\
&\quad\times\theta(t,y)K_R(p)\gamma_1\cdot\gamma_2\;dy\;dp\;dy_1\;dy_2\;ds_1\;ds_2\;dt\\
&\quad + 2\|\theta\|_\infty\left[\mu_1(R-1)+\mu_1(-R+1)+\mu_2(R-1)+\mu_2(-R+1)\right] + C\omega_R(\varepsilon)\\
&\le C_R\,\varepsilon\int_0^\infty\int_{\mathbb{R}^{2N+2}}\int_{Y\times\mathbb{R}}\theta(t,y)K_R(p+\varepsilon\varphi_1(t-\varepsilon s_1, y-\varepsilon y_1))\\
&\quad\times\phi(s_1,y_1,p)\phi\left(s_2,y_2,p+\frac{\varphi_1(t-\varepsilon s_1,y-\varepsilon y_1)-\varphi_2(t-\varepsilon s_2, y-\varepsilon y_2)}{\varepsilon}\right)\\
&\quad\times\gamma_1(t-\varepsilon s_1, y-\varepsilon y_1)\cdot\gamma_2(t-\varepsilon s_2, y-\varepsilon y_2)\;dy\;dp\;dy_1\;dy_2\;ds_1\;ds_2\;dt\\
&\quad + 2\|\theta\|_\infty\left[\mu_1(R-1)+\mu_1(-R+1)+\mu_2(R-1)+\mu_2(-R+1)\right] + C\omega_R(\varepsilon)\\
&\le C_R\,\varepsilon + 2\|\theta\|_\infty\left[\mu_1(R-1)+\mu_1(-R+1)+\mu_2(R-1)+\mu_2(-R+1)\right] + C\omega_R(\varepsilon),
\end{aligned}$$

so that eventually, for all test functions $\theta(t,y)\in\mathcal{D}_{\rm per}([0,\infty)\times\mathbb{R}^N)$ such that $\theta\geq 0$,

(4.10)
$$\limsup_{R\to\infty}\limsup_{\varepsilon\to 0}\int_0^\infty\int_{Y\times\mathbb{R}}\left[\frac{\partial m_1^\varepsilon}{\partial p}f_2^\varepsilon-\frac{\partial m_2^\varepsilon}{\partial p}f_1^\varepsilon-2\frac{\partial v}{\partial p}\nabla_y f_1^\varepsilon\cdot\nabla_y f_2^\varepsilon+r_1^\varepsilon f_2^\varepsilon+r_2^\varepsilon f_1^\varepsilon\right]$$
$$\times\,\theta(t,y)K_R(p)dt\,dy\,dp\leq 0.$$

Consequently, in the limit, we obtain for any test function $\theta(t,y)\in\mathcal{D}_{\rm per}([0,\infty)\times\mathbb{R}^N)$ such that $\theta\geq 0$

$$\int_0^\infty\int_Y (u_1-u_2)_+\partial_t\theta(t,y)+\mathbf{1}_{u_1>u_2}\left[A(y,u_1)-A(y,u_2)\right]\cdot\nabla_y\theta(t,y)\,dtdy$$
$$\geq\int_Y (u_1(t=0,y)-u_2(t=0,y))_+\,\theta(t=0,y)\,dy,$$

which means exactly that

$$\partial_t(u_1-u_2)_+ +\mathrm{div}_y\left[1_{u_1>u_2}\left(A(y,u_1)-A(y,u_2)\right)\right]\leq 0 \tag{4.11}$$

in the sense of distributions.

Integrating this last inequality on $(0,T)\times Y$ for any $T>0$ yields

$$||(u_1(t=T)-u_2(t=T))_+||_{L^1(Y)}\leq ||(u_1(t=0)-u_2(t=0))_+||_{L^1(Y)}. \tag{4.12}$$

Hence the derivation of (2.10) and (2.12) is complete; there remains only to prove Lemma 4.1. The argument goes along the same lines as Lemma 4.2.1 in [17].

Proof of Lemma 4.1. Notice that (4.1) is equivalent to

$$\frac{\partial}{\partial t}\left(\frac{\partial v}{\partial p}f_j\right)*\phi^\varepsilon+\frac{\partial}{\partial y_i}\left(\tilde{a}_i f_j\right)*\phi^\varepsilon-\Delta_y\left(\frac{\partial v}{\partial p}f_j\right)*\phi^\varepsilon=\eta_j\frac{\partial m_j^\varepsilon}{\partial p}. \tag{4.13}$$

Thus we focus on the derivation of (4.13) for f_1; let $(t,y,p)\in[0,\infty)\times Y\times\mathbb{R}$ be arbitrary. Following [17], one is tempted to consider the test function

$$(s,z,q)\mapsto\phi^\varepsilon(t-s,y-z,p-q)=\frac{1}{\varepsilon^{N+2}}\zeta_1\left(\frac{t-s}{\varepsilon}\right)\zeta_2\left(\frac{y-z}{\varepsilon}\right)\zeta_3\left(\frac{p-q}{\varepsilon}\right)$$

in Definition 2.3 of kinetic solutions. However, such a test function is not periodic in z, as required in Definition 2.3; but the support of $z\mapsto\zeta_2\left((y-z)/\varepsilon\right)$ is a subset of $\bar{B}(y,\varepsilon)$, the closed ball centered on y and of radius ε. Thus, for $0<\varepsilon<1/2$,

$$\mathrm{supp}\,\zeta_2\left((y-\cdot)/\varepsilon\right)\subset\bar{B}(y,\varepsilon)\subset\Pi_{i=1}^N\left(y_i-\frac{1}{2},y_i+\frac{1}{2}\right).$$

Hence for $\varepsilon<1/2$, we can extend $\zeta_2\left((y-\cdot)/\varepsilon\right)$ by periodicity on $\mathbb{R}^N$; the function thus obtained is denoted by $\tilde{\zeta}_{y,\varepsilon}$ and belongs to $\mathcal{C}^\infty_{\rm per}(\mathbb{R}^N)$.

Now, for fixed $(t, y, p) \in [0, \infty) \times Y \times \mathbb{R}$, we define the test function

$$\psi : (s, z, q) \mapsto \frac{1}{\varepsilon^{N+2}} \zeta_1 \left(\frac{t-s}{\varepsilon} \right) \tilde{\zeta}_{y,\varepsilon}(z) \zeta_3 \left(\frac{p-q}{\varepsilon} \right).$$

By construction, ψ belongs to $\mathcal{D}_{\rm per}([0, \infty) \times Y \times \mathbb{R})$. Thus ψ is an admissible test function, and according to Definition 2.3,

$$\int_0^\infty \int_{Y \times \mathbb{R}} f_1(s, z, q) \frac{\partial v(z, q)}{\partial q} \left\{ \partial_s \psi + a_i(z, v(z, q)) \partial_{z_i} \psi + \Delta_z \psi \right\} ds\, dz\, dq$$
$$= \int_0^\infty \int_{Y \times \mathbb{R}} m(s, z, q) \partial_q \psi(s, z, q)\ dsdzdq - \int_{Y \times \mathbb{R}} \mathbf{1}_{u_0(z) > v(z,q)} \psi(0, z, q) \frac{\partial v(z, q)}{\partial q} dzdq.$$

First, notice that since $\operatorname{supp} \zeta_1 \subset [-1, 0]$, we have $\psi(0, z, q) = 0$ for all z, q. Moreover, since f_1 and ψ are Y-periodic in their second variable, we have, for instance, setting $Y_y = \Pi_{i=1}^N (y_i - 1/2, y_i + 1/2) = y - e + Y$, where $e := (1/2, \ldots, 1/2) \in \mathbb{R}^N$,

$$\int_0^\infty \int_{Y \times \mathbb{R}} f_1 \frac{\partial v}{\partial q} \partial_s \psi = \int_0^\infty \int_{Y_y \times \mathbb{R}} f_1 \frac{\partial v}{\partial q} \partial_s \psi.$$

And when $z \in Y_y$, $\psi(s, z, q) = \phi^\varepsilon(t-s, y-z, p-q)$ by definition. Thus, using once again the assumption on the support of ζ_2,

$$\begin{aligned}
&\int_0^\infty \int_{Y \times \mathbb{R}} f_1(s, z, q) \frac{\partial v}{\partial q}(z, q) \partial_s \psi(s, z, q) ds\, dz\, dq \\
&= \int_0^\infty \int_{Y_y \times \mathbb{R}} f_1(s, z, q) \frac{\partial v}{\partial q}(z, q) \partial_s \phi^\varepsilon(t-s, y-z, p-q) ds\, dz\, dq \\
&= \int_0^\infty \int_{\mathbb{R}^N \times \mathbb{R}} f_1(s, z, q) \frac{\partial v}{\partial q}(z, q) \partial_s \phi^\varepsilon(t-s, y-z, p-q) ds\, dz\, dq \\
&= -\int_0^\infty \int_{\mathbb{R}^N \times \mathbb{R}} f_1(s, z, q) \frac{\partial v}{\partial q}(z, q) \partial_t \phi^\varepsilon(t-s, y-z, p-q) ds\, dz\, dq \\
&= -\partial_t \left[(f_1 v_p) * \phi^\varepsilon \right](t, y, p).
\end{aligned}$$

The other terms are treated in a similar way; we obtain

$$\int_0^\infty \int_{Y \times \mathbb{R}} f_1(s, z, q) \tilde{a}_i(z, q) \partial_{z_i} \psi(s, z, q)\ ds\, dz\, dq$$
$$= \int_0^\infty \int_{\mathbb{R}^N \times \mathbb{R}} f_1(s, z, q) \tilde{a}_i(z, q) \partial_{z_i} \phi^\varepsilon(t-s, y-z, p-q)\ ds\, dz\, dq = -\partial_{y_i} \left[f_1 \tilde{a}_i \right] * \phi^\varepsilon(t, y, p),$$

$$\int_0^\infty \int_{Y \times \mathbb{R}} f_1(s, z, q) v_q(z, q) \Delta_z \psi(s, z, q) ds\, dz\, dq = \Delta_y \left[(f_1 v_p) * \phi^\varepsilon \right](t, y, p),$$

$$\int_0^\infty \int_{Y \times \mathbb{R}} m(s, z, q) \partial_q \psi(s, z, q)\ ds\, dz\, dq = -\partial_p m_1^\varepsilon(t, y, p).$$

There remains to derive the bound on m_1^ε: by definition,

$$\begin{aligned}
&\int_0^\infty \int_Y m_1^\varepsilon(t,y,p)\,dt\,dy \\
&= \int_0^\infty \int_Y \int_{\mathbb{R}^{N+2}} m_1(s,z,q)\phi^\varepsilon(t-s,y-z,p-q)ds\,dz\,dq\,dt\,dy \\
&= \int_Y \int_{\mathbb{R}^{N+1}} \int_0^\infty \int_0^\infty m_1(s,z,q)\phi^\varepsilon(t-s,y-z,p-q)\,dt\,ds\,dz\,dq\,dy \\
&= \int_Y \int_{\mathbb{R}^{N+1}} \int_0^\infty \int_{-s}^\infty m_1(s,z,q)\phi^\varepsilon(u,y-z,p-q)\,du\,ds\;\,dz\,dq\,dy \\
&\leq \int_Y \int_{\mathbb{R}^{N+1}} \int_0^\infty \int_{\mathbb{R}} m_1(s,z,q)\phi^\varepsilon(u,y-z,p-q)\,du\,ds\;\,dz\,dq\,dy \\
&\leq \frac{1}{\varepsilon^{N+1}} \int_Y \int_{\mathbb{R}^{N+1}} \int_0^\infty m_1(s,z,q)\zeta_2\left(\frac{y-z}{\varepsilon}\right)\zeta_3\left(\frac{p-q}{\varepsilon}\right)\,ds\;\,dz\,dq\,dy.
\end{aligned}$$

Then, with the same notation as earlier,

$$\begin{aligned}
\int_Y \int_{\mathbb{R}^N} m_1(s,z,q)\zeta_2\left(\frac{y-z}{\varepsilon}\right)\,dz\,dy &= \int_Y dy\left(\int_{Y_y} dz\; m_1(s,z,q)\zeta_2\left(\frac{y-z}{\varepsilon}\right)\right) \\
&= \int_{Y\times Y} m_1(s,y+y'-e,q)\zeta_2\left(\frac{-y'+e}{\varepsilon}\right)dy\,dy' \\
&= \int_{Y\times Y} m_1(s,y,q)\zeta_2\left(\frac{-y'+e}{\varepsilon}\right)dy\,dy' \\
&= \left(\int_Y m_1(s,y,q)\,dy\right)\times\left(\int_{\mathbb{R}^N}\zeta_2\left(\frac{-y'}{\varepsilon}\right)\,dy'\right).
\end{aligned}$$

In the penultimate step, we have used the periodicity of m_1.

Thus

$$\begin{aligned}
\int_0^\infty \int_Y m_1^\varepsilon(t,y,p)\,dt\,dy &\leq \frac{1}{\varepsilon}\int_{\mathbb{R}}\int_{z\in Y}\int_0^\infty m_1(s,z,q)\zeta_3\left(\frac{p-q}{\varepsilon}\right)\,ds\;\,dz\,dq \\
&\leq \frac{1}{\varepsilon}\int_{\mathbb{R}} \mu_1(q)\zeta_3\left(\frac{p-q}{\varepsilon}\right)\,dq \\
&\leq \int_{-1}^1 \mu_1(p-\varepsilon q)\zeta_3(q)\,dq.
\end{aligned}$$

The monotonicity of μ_1 yields the desired result.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[2] E. Audusse and B. Perthame, *Uniqueness for scalar conservation laws with discontinuous flux via adapted entropies*, Proc. Roy. Soc. Edinburgh Sect. A, 135 (2005), pp. 253–266.

[3] A.-L. Dalibard, *Homogenization of a quasilinear parabolic equation with vanishing viscosity*, J. Math. Pures Appl., 86 (2006), pp. 133–154.

[4] A.-L. Dalibard, *Initial layer for the homogenization of a quasilinear parabolic equation with vanishing viscosity*, Arch. Rat. Mech. Anal., 185 (2007), pp. 515–543.

[5] A.-L. Dalibard, *Kinetic formulation for heterogeneous scalar conservation laws*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 23 (2006), pp. 475–498.

[6] W. E, *Homogenization of linear and nonlinear transport equations*, Comm. Pure Appl. Math., 45 (1992), pp. 301–326.

[7] W. E, *Homogenization of scalar conservation laws with oscillatory forcing terms*, SIAM J. Appl. Math., 52 (1992), pp. 959–972.

[8] W. E and D. Serre, *Correctors for the homogenization of conservation laws with oscillatory forcing terms*, Asymptot. Anal., 5 (1992), pp. 311–316.

[9] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 3rd ed., Springer, Berlin, 2001.

[10] S. N. Kruzkhov, *Generalized solutions of the Cauchy problem in the large for nonlinear equations of first order*, Soviet Math. Dokl., 10 (1969), pp. 785–788.

[11] S. N. Kruzkhov, *First order quasilinear equations in several independent variables*, Math. USSR Sb., 10 (1970), pp. 217–243.

[12] P.-L. Lions, G. Papanicolaou, and S. R. S. Varadhan, *Homogenization of Hamilton-Jacobi Equations*, unpublished manuscript, 1987.

[13] P.-L. Lions, B. Perthame, and E. Tadmor, *Formulation cinétique des lois de conservation scalaires multidimensionnelles*, C. R. Acad. Sci. Paris Sér. I Math., 312 (1991), pp. 97–102.

[14] P.-L. Lions, B. Perthame, and E. Tadmor, *A kinetic formulation of multidimensional scalar conservation laws and related equations*, J. Amer. Math. Soc., 7 (1994), pp. 169–191.

[15] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.

[16] B. Perthame, *Uniqueness and error estimates in first order quasilinear conservation laws via the kinetic entropy defect measure*, J. Math. Pures Appl., 77 (1998), pp. 1055–1064.

[17] B. Perthame, *Kinetic Formulation of Conservation Laws*, Oxford Lecture Ser. Math. Appl. 21, Oxford University Press, Oxford, UK, 2002.

[18] B. Perthame and E. Tadmor, *A kinetic equation with kinetic entropy functions for scalar conservation laws*, Comm. Math. Phys., 136 (1991), pp. 501–517.

SIAM J. MATH. ANAL.
Vol. 39, No. 3, pp. 916–931

BOUNDS ON ENERGY AND HELICITY DISSIPATION RATES OF APPROXIMATE DECONVOLUTION MODELS OF TURBULENCE*

WILLIAM LAYTON†

Abstract. We consider a family of high-accuracy, approximate deconvolution models of turbulence. For body force driven turbulence, we prove directly from the model's equations of motion the following bounds on the model's time-averaged energy dissipation rate $\langle \varepsilon_{ADM} \rangle$ and helicity dissipation rate $\langle \gamma_{ADM}(w) \rangle$:

$$\langle \varepsilon_{ADM} \rangle \leq 2\frac{U^3}{L} + \mathrm{Re}^{-1}\frac{U^3}{L}\left(1 + \left(\frac{\delta}{L}\right)^2\right) \quad \text{and}$$

$$|\langle \gamma_{ADM}(w) \rangle| \leq \frac{U^3}{L^2} + \sqrt{2}\,\mathrm{Re}^{-\frac{1}{2}}\left(1 + \frac{\delta^2}{L^2}\right)^{\frac{1}{2}}\frac{U^3}{L^2} + \mathrm{Re}^{-1}\left(1 + \frac{\delta^2}{L^2}\right)\frac{U^3}{L^2},$$

where U, L are global velocity and length scales, respectively, and δ is the large eddy simulation filter radius. We also give a partial result on the helicity dissipation rate of solutions of the Navier–Stokes equations.

Key words. energy dissipation rate, helicity, helicity dissipation rate, large eddy simulation, turbulence, deconvolution

AMS subject classifications. Primary, 76F65; Secondary, 76D05

DOI. 10.1137/06066223X

1. Introduction.

Turbulent flows consist of complex, interacting three-dimensional (3d) eddies of various sizes down to the Kolmogorov microscale, $\eta = O(\mathrm{Re}^{-3/4})$ in $3d$. A direct numerical simulation of the persistent eddies in a $3d$ flow thus requires roughly $O(\mathrm{Re}^{+9/4})$ mesh points in space per time step. Therefore, direct numerical simulation of turbulent flows is often not computationally economical or even feasible. On the other hand, the largest structures in the flow (containing most of the flow's energy) are responsible for much of the mixing and most of the flow's momentum transport. One promising approach to predicting a flow's large structures is called *large eddy simulation* (LES). LES seeks to model and predict the evolution of local, spatial averages over a user-selected length scale δ. The correct treatment by an LES model of fundamental flow quantities is critical to predicting correctly a flow's large structures.

The first fundamental integral invariant of the Euler equations is kinetic energy

$$E(u)(t) := \frac{1}{|\Omega|}\int_\Omega \frac{1}{2} u \cdot u dx.$$

Inviscid conservation of kinetic energy leads (by the reasoning of Richardson and Kolmogorov) to the kinetic energy cascade for turbulent flows which has important and universal features. If the LES model does not dissipate enough energy, the statistics of any energy cascade predicted by the model can be wrong, and there can be an accumulation of energy around the smallest resolved scales (i.e., wiggles in the model's

*Received by the editors June 6, 2006; accepted for publication (in revised form) April 30, 2007; published electronically September 19, 2007.
http://www.siam.org/journals/sima/39-3/66223.html
†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (wjl@pitt.edu, www.math.pitt.edu/~wjl). Partially supported by NSF grant DMS 0508260.

predicted velocity). The energy dissipation rates in various LES models are adjusted in various ways, such as using mixed models (i.e., adding eddy viscosity) and picking the parameters introduced (e.g., Lilly [L67]) to match the energy dissipation rate of homogeneous, isotropic turbulence. On the other hand, parameter-free models have great advantages, and understanding their important statistics, such as their energy dissipation rate, is critical to advancing their reliability.

The second fundamental integral invariant of the $3d$ Euler equations (discovered in 1961 by Moreau [M61]) is helicity or streamwise vorticity

$$H(u)(t) := \frac{1}{|\Omega|}\int_\Omega u\cdot(\nabla\times u)dx.$$

The helicity of a flow vanishes if and only if a flow has a reflectional symmetry and has the interpretation as the degree to which vortex lines are knotted and intertwined (Moffatt [M84], Moffatt and Tsoniber [MT92]). The interaction of helicity and energy is thought to play a key role in organizing flows. However, much less is known about helicity than energy, and its mathematical study is more difficult than that of energy because more derivatives are involved and neither H nor γ has one sign. (The best current result concerning helicity of weak solutions of the Navier–Stokes equations (NSE) appears to be Foias, Hoang, and Nicolenko [FHN06], in which it is proven that if the body force is potential, then $H(u)(T)\to 0$ as $T\to\infty$ for $\nu>0$.) By the same reasoning as energy, a similarity theory of coupled helicity and energy cascades with universal statistics has also been developed by Brissaud et al. [BFL73], Andre and Lesieur [AL77], Chen, Chen, and Eyink [CCE03], and Chen et al. [CCEH03], by Ditlevsen and Giuliani [DG01a], [DG01b] (and observed by Bourne and Orszag [BO97]) in turbulent flows and for the family of approximate deconvolution model (ADM) turbulence models in [LMNR06]. Helical modes where both signs exist are fundamental to the analysis of helicity cascades. In this theory the time-averaged helicity dissipation rate plays a key role analogous to that of the time-averaged energy dissipation rate. In the inertial range (after suitable averaging), the only quantities that distinguish one flow from another in these two cascades are their energy and helicity dissipation rates. Thus, a model prediction of energy and helicity dissipation rates is critical for evaluating a model's physical fidelity.

Herein we consider one family of high-accuracy, parameter-free models, ADMs, and bound the ADM time-averaged energy and helicity dissipation rates. The bounds derived mirror both dissipation rates of the underlying solution of the NSE (in the limit $\delta\to 0$) and the estimates of it derived in [LN07b], [LMNR06] by dimensional analysis.

To begin, consider the NSE in a periodic box in $\mathbb{R}^3$, $\Omega=(0,L_\Omega)^3$:

$$\begin{gathered} u_t+u\cdot\nabla u-\nu\triangle u+\nabla p=f(x) \quad \text{in } \Omega=(0,L_\Omega)^3,\ t>0,\\ \nabla\cdot u=0 \quad \text{in } (0,L_\Omega)^3, \end{gathered} \tag{1.1}$$

subject to periodic (with zero mean) conditions

$$\begin{gathered} u(x+L_\Omega e_j,t)=u(x,t), \quad j=1,2,3, \quad \text{and}\\ \int_\Omega \phi dx=0 \ \text{ for } \ \phi=u,\ u_0,\ f,\ p. \end{gathered} \tag{1.2}$$

We suppose throughout that the data $u_0(x), f(x)$ are smooth and satisfy

$$\nabla\cdot u_0=0 \quad \text{and } \nabla\cdot f=0.$$

Many averaging operators are used in LES; see, e.g., Sagaut [S01], John [J04], and Berselli, Iliescu, and Layton [BIL06]. Herein we consider a *differential filter* (Germano [Ger86]) associated with length scale $\delta > 0$ related to the Yoshida regularization (and sometimes called a *Helmholz filter* in the alpha-model literature, e.g., Cheskidov et al. [CHOT05]) defined as follows. Given $\phi(x)$, $\overline{\phi(x)}$ is the unique L-periodic solution of

$$A\overline{\phi} := -\delta^2 \triangle\overline{\phi} + \overline{\phi} = \phi \text{ in } \Omega.$$

Averaging the NSE (i.e., applying A^{-1} to (1.1)) gives the exact space-filtered NSE for $\overline{u}$

$$\begin{aligned} \overline{u}_t + \overline{u \cdot \nabla u} - \nu\triangle\overline{u} + \nabla\overline{p} = \overline{f}(x) \text{ and} \\ \nabla \cdot \overline{u} = 0. \end{aligned}$$

This is not closed since (noting that $\overline{u \cdot \nabla u} = \nabla \cdot (\overline{uu})$)

$$\overline{u\,u} \;\neq \overline{u}\,\overline{u}.$$

There are many closure models used in LES; see Sagaut [S01], John [J04], Lesieur, Metais, and Comte [LMC05], and Berselli, Iliescu, and Layton [BIL06] for surveys. Approximate deconvolution models, studied herein, are used, with success, in many simulations of turbulent flows, e.g., the works of Adams, Stolz, and Kleiser [AS01], [AS02], [SA99], [SAK01a], [SAK01b], [SAK02]. They are among the most accurate of turbulence models, and one of the few turbulence models for which a mathematical confirmation of their effectiveness is known [LL06b] and [DE06]. Briefly, an approximate deconvolution operator (constructed in section 3) denoted by D_N is an operator satisfying

$$\phi = D_N(\overline{\phi}) + O(\delta^{2N+2}) \quad \text{for smooth } \phi.$$

Since $D_N\overline{u}$ approximates u to accuracy $O(\delta^{2N+2})$ in the smooth flow regions, it is justified to consider the closure approximation:

$$\overline{uu} \simeq \overline{D_N\overline{u}D_N\overline{u}} + O(\delta^{2N+2}). \tag{1.3}$$

Using this closure approximation, the resulting family of ADMs is given by[1]

$$\begin{aligned} w_t + \nabla \cdot (\overline{D_N w\; D_N w}) - \nu\triangle w + \nabla q = \overline{f}(x), \\ \nabla \cdot w = 0, N = 0, 1, 2, \ldots. \end{aligned} \tag{1.4}$$

As a special case, $D_0\overline{u} = \overline{u} + O(\delta^2)$ gives the zeroth order ADM:

$$w_t + \nabla \cdot (\overline{w\; w}) - \nu\triangle w + \nabla q = \overline{f}(x) \text{ and } \nabla \cdot w = 0.$$

We consider two important flow statistics: the *time-averaged energy* and *helicity dissipation rates*. The energy dissipation rate is a fundamental statistic in experimental and theoretical studies of turbulence, e.g., Sreenivasan [S84], [S98], Bourne and Orszag [BO97], Pope [P00], Frisch [Frisch], and Lesieur [Les97]. In the theory of turbulent energy cascades, the time-averaged energy dissipation is the only quantity

[1]In practical computations with ADMs an additional time relaxation term $\chi\,(w - \overline{w})$ has often been added to (1.4). This term can be used as a numerical regularization in any model and is studied in [LN07a], [ELN07], [AS02], [P06], and [Gue04].

that distinguishes the inertial range of one flow from another. In the early 1990s Constantin and Doering [CD92] (see also Doering and Gibbon [DG95]) established a direct link between the phenomenology of energy dissipation and that predicted for shear flows directly from the NSE. This work builds on the work of Busse [B78] and Howard [H72] (and others) and has developed in many important directions, including Childress, Kerswell, and Gilbert [CKG01], Kerswell [K98], and Wang [W97] (shear flows) and Foias [F97], Doering and Foias [DF02], and Cheskidov, Doering, and Petrov [CDP07] (body force driven flows). Because of the greater difficulties of studying helicity directly from the NSE, this connection remains open for helicity dissipation rates; see section 5.1.

Let $\langle\cdot\rangle$ denote long time averaging (defined in section 2). K41 phenomenology (e.g., Frisch [Frisch], Muschinski [Mus97], and Pope [P00]) in [LN07b] suggests the scaling of the energy dissipation rate $\langle\varepsilon_{ADM}\rangle$

$$\langle\varepsilon_{ADM}\rangle \approx \frac{U^3}{L}\left(1+\frac{\delta^2}{L^2}\right).$$

In section 4, we prove directly from the equations of motion (1.4) that the energy dissipation rate of the model satisfies

$$\langle\varepsilon_{ADM}\rangle \leq 2\frac{U^3}{L} + \mathrm{Re}^{-1}\frac{U^3}{L}\left(1+\left(\frac{\delta}{L}\right)^2\right).$$

Here U, L denote natural velocity and length scales, respectively, associated with the largest scales of the model (1.4), defined precisely in section 2.1.

The time-averaged helicity dissipation rate is defined in an analogous way. For the NSE, it is known[2] that the helicity satisfies the balance equation

$$H(u)(T) + \int_0^T \gamma(u)(t)dt = H(u_0) + \int_0^T \frac{1}{|\Omega|}(\nabla\times f, u)dt. \tag{1.5}$$

The term $\gamma(u)$ is thus taken to be the helicity dissipation rate and is given by

$$\gamma(u) := \frac{\nu}{|\Omega|}(\nabla\times u, \nabla\times\nabla\times u). \tag{1.6}$$

The helicity dissipation rate is defined several ways in the literature on the phenomenology of helicity cascades due to possible coupling with energy dissipation rates. Herein we use the above definition for the NSE, which is natural from the point of view of the above helicity balance equation. For the model (1.4), we use the definition of the model helicity dissipation rate $\gamma_{ADM}(w)$ that comes from the model's helicity balance in the same way as above for the NSE. In section 5, we prove that the model's helicity dissipation rate $|\langle\gamma_{ADM}(w)\rangle|$ satisfies

$$|\langle\gamma_{ADM}(w)\rangle| \leq \frac{U^3}{L^2} + \sqrt{2}\,\mathrm{Re}^{-\frac{1}{2}}\left(1+\frac{\delta^2}{L^2}\right)^{\frac{1}{2}}\frac{U^3}{L^2} + \mathrm{Re}^{-1}\left(1+\frac{\delta^2}{L^2}\right)\frac{U^3}{L^2}.$$

This estimate of $|\langle\gamma_{ADM}(w)\rangle|$ is consistent, as $\delta\to 0$, with the dimensional analysis estimate of $\frac{U^3}{L^2}$ for the NSE.

[2] Derived formally by multiplication by the vorticity, integration over the flow domain and integration by parts.

The higher order models give much more accurate predictions of flow quantities. However, the mathematical structures of the zeroth order model and the entire family of ADMs are closely related. Proofs will be given in detail for the zeroth order model and the corresponding proofs (which involve only additional subscripts) for the Nth ADM sketched. Aside from the family of ADMs, we also give a partial result on the helicity dissipation rate for the NSE and delineate why the technique (just) fails to isolate the NSE's helicity dissipation rate. Thus, it seems likely that the techniques used herein can be used to prove parallel estimates of energy and helicity dissipation rates for other models which have more regular solutions than the NSE and known helicity balance equations, such as the alpha model.

2. Notation and preliminaries. The long time average of a function $\phi(t)$ is defined, following [DF02], by

$$\langle \phi \rangle := Lim_{T\to\infty} \frac{1}{T}\int_0^T \phi(t)dt,$$

where Lim denotes a generalized limit. (It can also be defined by a limit superior which lengthens several proofs.)

First consider the zeroth order model. With $|\Omega|$ the volume of the flow domain, the scale of the body force and the large scale velocity are defined, respectively, by

$$F := \left(\frac{1}{|\Omega|}\int_\Omega |f(x)|^2 dx\right)^{\frac{1}{2}} \text{ and}$$
$$U := \left\langle \frac{1}{|\Omega|}\int_\Omega |w(x,t)|^2 dx\right\rangle^{\frac{1}{2}}.$$

Let $||\cdot||, (\cdot,\cdot)$ denote the usual $L^2(\Omega)$ norm and inner product, respectively (other norms are explicitly indicated by a subscript). The global length scale associated with the power input of the large scales, i.e., with $f(x)$, is

$$L := \min\left\{ L_\Omega, \frac{F}{||\nabla f||_{L^\infty(\Omega)}}, \frac{F}{(\frac{1}{|\Omega|}||\nabla f||^2)^{\frac{1}{2}}}, \frac{F}{(\frac{1}{|\Omega|}||\nabla\times f||^2)^{\frac{1}{2}}}, \frac{F^{\frac{1}{2}}}{(\frac{1}{|\Omega|}||\triangle f||^2)^{\frac{1}{4}}}\right\}.$$

It is easy to check that L has units of length and satisfies the inequalities:

$$\begin{aligned} &||\nabla f||_{L^\infty} \le \frac{F}{L},\\ &\frac{1}{|\Omega|}\int_\Omega |\nabla f(x)|^2 dx \le \frac{F^2}{L^2}, \frac{1}{|\Omega|}\int_\Omega |\nabla\times f(x)|^2 dx \le \frac{F^2}{L^2}, \quad \text{and}\\ &\frac{1}{|\Omega|}\int_\Omega |\triangle f(x)|^2 dx \le \frac{F^2}{L^4}. \end{aligned} \tag{2.1}$$

The kinetic viscosity is denoted ν, and the associated global Reynolds number is $\mathrm{Re} := \frac{LU}{\nu}$.

The energy dissipation rate induced by the model depends on the precise form of the model's kinetic energy balance. Let w denote the solution of the zeroth order model. The appropriate definitions (see Proposition 3.1, Remark 4.1, (3.2)–(3.5), as well as [LL03], [LL06a], [LL06b]) for the zeroth order model are

$$\varepsilon_{ADM-0}(w)(t) = \frac{\nu}{L^3}\{||\nabla w(t)||^2 + \delta^2||\triangle w(t)||^2\} \text{ and}$$
$$\langle \varepsilon_{ADM-0}\rangle := \langle \varepsilon_{ADM-0}(w)(t)\rangle.$$

Before introducing the notation for the general case we must first define the van Cittert approximate deconvolution operators.

2.1. Approximate deconvolution operators. The filtering or convolution operator $u \to \overline{u}$ is a bounded map: $L^2(\Omega) \to L^2(\Omega)$. If (as in the case we study) it is smoothing, its inverse cannot be bounded due to small divisor problems. An approximate deconvolution operator D_N is an approximate inverse $\overline{u} \to D_N(\overline{u}) \approx u$, (e.g., [Geu97]) which

- is a bounded operator on $L^2(\Omega)$,
- approximates u in some useful (typically asymptotic) sense, and
- satisfies other conditions necessary for the application at hand.

The deconvolution operator we consider was studied by van Cittert in 1931, e.g., Bertero and Boccacci [BB98], and its use in LES pioneered by Adams, Kleiser, and Stolz [AS01], [SA99], [AS02], [SAK01a], [SAK01b], [SAK02]. The Nth van Cittert approximate deconvolution operator D_N is defined by N steps of Picard iteration [BB98] for the fixed point problem:

$$\text{Given } \overline{u}, \text{ solve } u = u + \{\overline{u} - A^{-1}u\} \text{ for } u.$$

ALGORITHM 2.1 (van Cittert approximate deconvolution operator). *$u_0 = \overline{u}$, for $n = 1, 2, \ldots, N-1$, perform $u_{n+1} = u_n + \{\overline{u} - A^{-1}u_n\}$. Define $D_N\overline{u} := u_N$.*

By eliminating the intermediate steps, the Nth deconvolution operator D_N is given explicitly by

$$D_N\phi := \sum_{n=0}^{N} (I - A^{-1})^n \phi. \tag{2.2}$$

For example, the approximate deconvolution operators corresponding to $N = 0, 1, 2$ are

$$D_0\overline{u} = \overline{u},$$
$$D_1\overline{u} = 2\overline{u} - \overline{\overline{u}},$$
$$D_2\overline{u} = 3\overline{u} - 3\overline{\overline{u}} + \overline{\overline{\overline{u}}}.$$

DEFINITION 2.1. *The deconvolution-weighted inner product and norm $(\cdot,\cdot)_N$ and $||\cdot||_N$ are, respectively,*

$$(u, v)_N := (u, D_N v)\,, \quad ||u||_N := (u, u)_N^{\frac{1}{2}}.$$

LEMMA 2.2. *Consider the approximate deconvolution operator*

$$D_N : L^2(\Omega) \to L^2(\Omega).$$

D_N is a bounded, self-adjoint, positive-definite operator and satisfies

$$||\phi||^2 \le ||\phi||_N^2 \le (N+1)||\phi||^2 \ \forall \phi \in L^2(\Omega).$$

Proof. D_N is a function of the bounded, self-adjoint operator A^{-1} and is thus bounded and self-adjoint. By the spectral mapping theorem we have

$$\lambda(D_N) = \sum_{n=0}^{N} \lambda(I - A^{-1})^n = \sum_{n=0}^{N} (1 - \lambda(A^{-1}))^n, \text{ and}$$
$$0 < \lambda(A^{-1}) \le 1 \text{ by the definition of operator } A.$$

Thus, $1 \le \lambda(D_N) \le N+1$. Since D_N is a self-adjoint operator, this proves positive definiteness and the above equivalence of norms. □%newpage

3. Kinetic energy balance of ADM turbulence models. To see the mathematical key to the estimates of energy and helicity dissipation rates, we first recall from [LL03], [DE06] (see also [LL06a], and [MM06] for the more difficult case of no-slip boundary conditions) the energy equality for the ADM (1.4).

PROPOSITION 3.1. *If w is a weak or strong solution*[3] *of* (1.4), *w satisfies*

$$\frac{1}{2}[||w(T)||_N^2 + \delta^2||\nabla w(T)||_N^2] + \int_0^T \nu||\nabla w(t)||_N^2 + \nu\delta^2||\triangle w(t)||_N^2 dt = \frac{1}{2}[||\overline{u}_0||_N^2 + \delta^2||\nabla \overline{u}_0||_N^2] + \int_0^T (f, w(t))_N \, dt.$$

Proof (*sketch*). Let (w, q) denote a periodic solution of the ADM (1.4). Multiplying (1.4) by $AD_N w$ and integrating over Ω gives

$$\int_\Omega w_t \cdot AD_N w + \nabla \cdot (\overline{D_N w \ D_N w}) \cdot AD_N w - \nu \triangle w \cdot AD_N w + \nabla q \cdot AD_N w d\mathbf{x} = \int_\Omega \overline{f} \cdot AD_N w d\mathbf{x}.$$

The nonlinear term vanishes exactly because

$$\int_\Omega \nabla \cdot (\overline{D_N w \ D_N w}) \cdot AD_N w d\mathbf{x} = \int_\Omega A^{-1}(\nabla \cdot (D_N w \ D_N w)) \cdot AD_N w d\mathbf{x} = \int_\Omega \nabla \cdot (D_N w \ D_N w) \cdot D_N w d\mathbf{x} = 0.$$

Integrating by parts the remaining terms gives

$$\frac{d}{dt}\frac{1}{2}\{||w(t)||_N^2 + \delta^2||\nabla w(t)||_N^2\} + \nu\{||\nabla w(t)||_N^2 + \delta^2||\triangle w(t)||_N^2\} = (f, w(t))_N \,. \tag{3.1}$$

The results follow by integrating this from 0 to t. □

From Proposition 3.1, the ADM kinetic energy $E_{ADM-N}(w)(t)$, energy dissipation rate $\varepsilon_{ADM-N}(w)(t)$, time-averaged dissipation rate $\langle\varepsilon_{ADM-N}\rangle$, and power input $P_{ADM-N}(w)(t)$ are clearly identified.

$$E_{ADM-N}(w)(t) := \frac{1}{2|\Omega|}\{||w(t)||_N^2 + \delta^2||\nabla w(t)||_N^2\}, \tag{3.2}$$

$$\varepsilon_{ADM-N}(w)(t) := \frac{\nu}{|\Omega|}\{||\nabla w(t)||_N^2 + \delta^2||\triangle w(t)||_N^2\}, \tag{3.3}$$

$$\langle\varepsilon_{ADM-N}\rangle := \langle\varepsilon_{ADM-N}(w)(t)\rangle, \tag{3.4}$$

$$P_{ADM-N}(w)(t) := \frac{1}{|\Omega|}(f, w(t))_N. \tag{3.5}$$

[3]Unlike the NSE case, e.g., [Gal95], it is known that weak=strong for the ADM and that both exist and are unique.

Let $||\cdot||_N$ denote the deconvolution-weighted $L^2(\Omega)$ norm (Definition 2.1). The deconvolution weighted scales of the body force and large scale velocity are defined by

$$F_N := \left(\frac{1}{|\Omega|}||f||_N^2\right)^{\frac{1}{2}}, \text{ and } U_N := \left\langle \frac{1}{|\Omega|}||w||_N^2 \right\rangle^{\frac{1}{2}}.$$

Note that these are related to F and U by

$$F \le F_N \le (N+1)^{\frac{1}{2}} F, U \le U_N \le (N+1)^{\frac{1}{2}} U, \text{ and also}$$
$$\langle \varepsilon_{ADM-0}(w)\rangle \le \langle \varepsilon_{ADM-N}(w)\rangle \le (N+1)^{\frac{1}{2}} \langle \varepsilon_{ADM-0}(w)\rangle.$$

For $N = 1, 2, 3, \dots$ the deconvolution-weighted global length scale associated with the power input to the large scales, i.e., with $f(x)$, is defined to be

$$L_N := \min \left\{ L_\Omega, \frac{F_N}{||D_N^{\frac{1}{2}} \nabla f||_{L^\infty(\Omega)}}, \frac{F_N}{(\frac{1}{|\Omega|}||\nabla f||_N^2)^{\frac{1}{2}}}, \frac{F_N}{(\frac{1}{|\Omega|}||\nabla \times f||_N^2)^{\frac{1}{2}}}, \frac{F_N^{\frac{1}{2}}}{(\frac{1}{|\Omega|}||\triangle f||_N^2)^{\frac{1}{4}}} \right\}.$$

It is easy to check that L_N has units of length and satisfies the deconvolution-weighted form of the inequalities (2.1) above.

LEMMA 3.2. *As $\delta \to 0$, for $N = 0, 1, 2, \dots$*

$$E_{ADM-N}(w)(t) \to E(w)(t) = \frac{1}{2|\Omega|}||w(t)||^2,$$
$$\varepsilon_{ADM-N}(w)(t) \to \varepsilon(w)(t) = \frac{\nu}{2|\Omega|}||\nabla w(t)||^2, \textit{ and}$$
$$P_{ADM-N}(w)(t) \to P(w)(t) = \frac{1}{|\Omega|}(f(t), w(t)).$$

Proof. As $\delta \to 0$ all of the δ^2 terms drop out in the definitions above, $D_N \to I$, and $||\phi||_N \to ||\phi||$. ☐

COROLLARY 3.3. *Let $f = f(x) \in L^2(\Omega)$ and w be a solution of the ADM turbulence model* (1.4)*; then*

$$\sup_{t\in(0,\infty)} E_{ADM-N}(w)(t) \le C(data) < \infty,$$
$$\frac{1}{T}\int_0^T \varepsilon_{ADM-N}(w)(t)dt \le C(data) < \infty.$$

Proof. We begin with (3.1) from the proof of Proposition 3.1. Using the Poincaré and Cauchy–Schwarz inequalities we have from (3.1)

$$\frac{d}{dt}E_{ADM-N}(t) + \alpha E_{ADM-N}(t) \le ||f||_N^2$$

for some $\alpha > 0$, which implies $E_{ADM-N}(w)(t)$ is uniformly bounded in time. For boundedness of the time-averaged dissipation rate, divide the energy estimate of the

ADM turbulence model energy equality from Proposition 3.1 by T :

$$\frac{1}{T}E_{ADM-N}(w)(T)+\frac{1}{T}\int_0^T \varepsilon_{ADM-N}(w)(t)dt$$
$$=\frac{1}{T}E_{ADM-N}(w)(0)+\frac{1}{T}\int_0^T (f,w(t))_N dt$$
$$\leq \frac{1}{T}E_{ADM-N}(w)(0)+||f||_N\left[\frac{1}{T}\int_0^T ||w(t)||_N^2 dt\right]^{\frac{1}{2}} \tag{3.6}$$
$$\leq C(data). \qquad \square \tag{3.7}$$

4. Bounds on energy dissipation rates. We prove the following estimate on the model's time-averaged energy dissipation rates. Recall that the body force $f(x)$ is a smooth, divergence-free function. The global velocity and length scales are given in section 3.

THEOREM 4.1. *Let the data $f(x)$ and $u_0(x)$ be smooth, divergence-free functions. For all cases $N=0,1,2,3,\ldots$*

$$\langle\varepsilon_{ADM-N}(w)\rangle \leq 2\frac{U_N^3}{L_N}+\mathrm{Re}^{-1}\frac{U_N^3}{L_N}\left(1+\frac{\delta^2}{L_N^2}\right).$$

The proof in the general case follows the zeroth order case by adding subscripts as appropriate. The mathematics driving the proof in both cases is the precise estimate of the model's energy balance. In the general case, the model's kinetic energy balance satisfies the analog of the zeroth order's energy balance with norms replaced by deconvolution-weighted norms. We shall give the proof in detail for the notationally clearest, $N=0$, case (indicating the notational modifications in the general case).

PROPOSITION 4.2. *Let the data $f(x)$ and $u_0(x)$ be smooth, divergence-free functions. For the case $N=0$,*

$$\langle\varepsilon_{ADM-0}(w)\rangle \leq 2\frac{U^3}{L}+\mathrm{Re}^{-1}\frac{U^3}{L}\left(1+\frac{\delta^2}{L^2}\right).$$

The proof of the proposition and theorem follow next. They combine the energy estimate for the ADM in Proposition 3.1 with the breakthrough arguments of Foias [F97] and Doering and Foias [DF02] from the NSE case.

The case of general $N=0,1,2,\ldots$. The first of two key bounds is obtained by time averaging the energy inequality of Proposition 3.1. Using Corollary 3.3 we have for $N=0,1,2,\ldots$

$$Lim_{T\to\infty}\frac{1}{T}\int_0^T \nu||\nabla w(t)||_N^2+\nu\delta^2||\triangle w(t)||_N^2 dt \leq Lim_{T\to\infty}\frac{1}{T}\int_0^T (f,w(t))_N\, dt.$$

The Cauchy–Schwarz inequality and Corollary 3.3 imply

$$\langle\ \varepsilon_{ADM-N}\rangle \leq F_N U_N, \text{ and thus } \langle\varepsilon_{ADM-0}\rangle \leq FU. \tag{4.1}$$

Time averaging the ADM turbulence model (1.4) gives for $N=0,1,2,\ldots$

$$\langle w\rangle_t+\nabla\cdot(\langle\overline{D_N w\ D_N w}\rangle)-\nu\triangle\langle w\rangle+\nabla\langle q\rangle=\langle\overline{f}\rangle, \tag{4.2}$$
$$\nabla\cdot\langle w\rangle=0.$$

The case $N = 0$. Set $N = 0$ and recall that $D_0 = I$. Take the inner product of the time-averaged model (4.2) with Af. Note that because $f = f(x)$, $\langle \overline{f} \rangle = f(x)$. Further $(\overline{f}, Af) = (A^{-1}f, Af) = ||f||^2$ and analogously for the nonlinear term. Since $\nabla \cdot f = 0$, the pressure term vanishes. This gives

$$\frac{1}{|\Omega|}||f||^2 = \frac{1}{|\Omega|}(Af, \langle w \rangle_t) - \frac{1}{|\Omega|}(\nabla f, \langle D_0 w \ D_0 w \rangle) + \frac{\nu}{|\Omega|}(A \nabla f, \nabla \langle w \rangle).$$

The time derivative term vanishes in the limit as $T \to \infty$ by the Cauchy–Schwarz inequality and Corollary 3.3. The last term on the right-hand side (RHS) is integrated by parts, giving:

$$\frac{1}{|\Omega|}||f||^2 = -\frac{1}{|\Omega|}(\nabla f, \langle D_0 w \ D_0 w \rangle) + \frac{\nu}{|\Omega|}\{(\nabla f, \nabla \langle w \rangle) + \delta^2 (\triangle f, \triangle \langle w \rangle)\}.$$

Thus, using the Cauchy–Schwarz inequality,

$$\frac{1}{|\Omega|}||f||^2 \leq -\frac{1}{|\Omega|}(\nabla f, \langle D_0 w \ D_0 w \rangle) + \varepsilon_{ADM-0}(f)^{\frac{1}{2}} \varepsilon_{ADM-0}(\langle w \rangle)^{\frac{1}{2}}. \tag{4.3}$$

Next, consider the nonlinear term on the above RHS. By the definitions of L, F, U we have (recalling $D_0 w = w$)

$$\frac{1}{|\Omega|}(\nabla f, \langle w \ w \rangle) \leq ||\nabla f||_{L^\infty} \left\langle \frac{1}{|\Omega|}||w||^2 \right\rangle \leq \frac{FU^2}{L}. \tag{4.4}$$

By the triangle inequality we have

$$||\nabla \langle w \rangle||^2 \leq \langle ||\nabla w||^2 \rangle, \text{ and } ||\triangle \langle w \rangle||^2 \leq \langle ||\triangle w||^2 \rangle. \tag{4.5}$$

(This step is not sharp.) This implies, by the definitions of F, L,

$$\begin{gathered} \varepsilon_{ADM-0}(\langle w \rangle)^{\frac{1}{2}} \leq \langle \varepsilon_{ADM-0}(w) \rangle^{\frac{1}{2}}, \\ \frac{\nu}{|\Omega|}||\nabla f||^2 \leq \nu \frac{F^2}{L^2}, \text{ and } \frac{\nu \delta^2}{|\Omega|}||\triangle f||^2 \leq \nu \delta^2 \frac{F^2}{L^4}. \end{gathered} \tag{4.6}$$

Using the bounds (4.4), (4.6), and (4.7) in (4.3) gives

$$F^2 \leq \frac{FU^2}{L} + \left(\frac{\nu F^2}{L^2} + \frac{\nu \delta^2 F^2}{L^4} \right)^{\frac{1}{2}} \langle \varepsilon_{ADM-0}(w) \rangle^{\frac{1}{2}}. \tag{4.7}$$

From the first basic estimate $\langle \varepsilon_{ADM-0}(w) \rangle \leq FU$. Inserting this in the RHS and cancelling the obvious terms gives

$$\langle \varepsilon_{ADM-0}(w) \rangle \leq FU \leq \frac{U^3}{L} + U \left(\frac{\nu}{L^2} + \frac{\nu \delta^2}{L^4} \right)^{\frac{1}{2}} \langle \varepsilon_{ADM-0}(w) \rangle^{\frac{1}{2}}. \tag{4.8}$$

Thus, by Young's inequality

$$\langle \varepsilon_{ADM-0}(w) \rangle \leq 2\frac{U^3}{L} + \frac{\nu U^2}{L^2} \left(1 + \frac{\delta^2}{L^2} \right),$$

and Proposition 4.2 is proven:

$$\langle \varepsilon_{ADM-0}(w) \rangle \leq 2\frac{U^3}{L} + \mathrm{Re}^{-1} \frac{U^3}{L} \left(1 + \frac{\delta^2}{L^2} \right).$$

The case of general $N = 0, 1, 2, \ldots$. For the general case in Theorem 4.1, the claimed estimates hold by the same proof as the $N = 0$ case (above) with norms and inner products replaced by their deconvolution-weighted versions (i.e., by adding subscripts N).

Remark 4.1. *Exactly as in the NSE case, Foias* [F97] *and Doering and Foias* [DF02]*, the estimate can be improved by more careful treatment of the quadratic equation. The result is the elimination of the multiplier* 2 *on the RHS and a slight modification of the second term.*

5. Bounds on helicity dissipation rates. This section considers bounds on the time-averaged helicity dissipation rate for both the NSE and the ADM turbulence models (1.4). We consider the NSE case first and derive a partial result. The expected result predicted by dimensional analysis in the NSE case is recovered if it is known that the helicity of a solution of the NSE is eventually bounded—a property that seems physically obvious but mathematically intractable. Because of the enhanced kinetic energy bound available for the ADM turbulence model, we are able to prove a bound on the ADM helicity, $\langle \gamma_{ADM-N}(w) \rangle$.

5.1. Helicity dissipation for the NSE. Because of the incomplete nature of the final result, we proceed formally. Dividing the helicity balance equation by T gives

$$\frac{1}{T}H(u)(T) + \frac{1}{T}\int_0^T \gamma(u)(t)dt = \frac{1}{T}H(u_0) + \frac{1}{T}\int_0^T \frac{1}{|\Omega|}(\nabla \times f, u)dt.$$

If the initial velocity is smooth, $\frac{1}{T}H(u_0) \to 0$ as $T \to \infty$. Further, if $\nabla \times f$ is square integrable, then, by the Cauchy–Schwarz inequality,

$$Lim_{T\to\infty} \left| \frac{1}{T}\int_0^T \frac{1}{|\Omega|}(\nabla \times f, u)dt \right| \leq \frac{FU}{L}(< \infty). \tag{5.1}$$

Thus, the following generalized limits satisfy

$$Lim_{T\to\infty} \left| \frac{1}{T}H(u)(T) + \frac{1}{T}\int_0^T \gamma(u)(t)dt \right| \leq \frac{FU}{L}(< \infty).$$

In the NSE case, Foias [F97] and Doering and Foias [DF02] prove, as a step to bounds on energy dissipation rates, the following intermediate result on the numerator of the RHS:

$$FU \leq \frac{U^3}{L} + \nu^{\frac{1}{2}}\frac{U}{L}\langle \varepsilon \rangle^{\frac{1}{2}}. \tag{5.2}$$

(For example, formally set $\delta = 0$ in the estimate (4.8).) Using this and the bound on energy dissipation in [F97], [DF02] that $\langle \varepsilon \rangle \leq 2\frac{U^3}{L} + \mathrm{Re}^{-1}\frac{U^3}{L}$, in (5.1) gives

$$Lim_{T\to\infty} \left| \frac{1}{T}H(u)(T) + \frac{1}{T}\int_0^T \gamma(u)(t)dt \right| \leq \frac{U^3}{L^2} + \sqrt{2}\,\mathrm{Re}^{-\frac{1}{2}}\frac{U^3}{L^2} + \mathrm{Re}^{-1}\frac{U^3}{L^2}. \tag{5.3}$$

If $H(T)$ is eventually bounded (as is expected physically but unknown mathematically), this gives the upper bound

$$|\langle \gamma \rangle| \leq \frac{U^3}{L^2} + \sqrt{2}\,\mathrm{Re}^{-\frac{1}{2}}\frac{U^3}{L^2} + \mathrm{Re}^{-1}\frac{U^3}{L^2}. \tag{5.4}$$

Thus, it is clear that obtaining a bound on the helicity dissipation, i.e., completing the step between (5.3) and (5.4), depends on proving boundedness of $H(T)$.

5.2. Helicity dissipation in ADM turbulence models. From the NSE case, it is clear that the key to completing the argument of section 5.1 for the ADM turbulence model will be proving $H_{ADM-N}(w)(T)$ is bounded. We begin by recalling the ADM turbulence model's helicity balance (the essential first ingredient in the analysis) discovered by Rebholz [R07]. Define the ADM helicity which is exactly conserved by the model if $\nu = 0$

$$H_{ADM-N}(w)(t) := \frac{1}{|\Omega|}[(w, \nabla \times w)_N + \delta^2(\nabla \times w, \nabla \times \nabla \times w)_N]. \tag{5.5}$$

For the ADM, it is known [R07] that the ADM helicity satisfies the balance equation

$$H_{ADM-N}(w)(T) + \int_0^T \gamma_{ADM-N}(w)(t)dt = H(w(0)) + \int_0^T \frac{1}{|\Omega|}(\nabla \times f, w)dt, \tag{5.6}$$

where $\gamma_{ADM-N}(w)$ is the model's helicity dissipation rate given by

$$\gamma_{ADM-N}(w) := \frac{\nu}{|\Omega|}[(\nabla \times w, \nabla \times \nabla \times w)_N + \delta^2(\nabla \times \nabla \times w, \nabla \times \nabla \times \nabla \times w)_N]. \tag{5.7}$$

Note that (in the zeroth order model, to simplify notation)

$$H_{ADM}(w) = H(w) + \delta^2 H(\nabla \times w) \text{ and}$$
$$\gamma_{ADM}(w) = \gamma(w) + \delta^2 \gamma(\nabla \times w).$$

5.3. Bounding model helicity. Arguing as in [LN07b], the zeroth order model, (1.4) with $N = 0$, is equivalent to $\nabla \cdot Aw = 0$ and

$$Aw_t + w \cdot \nabla w + \nabla Aq - \nu \triangle Aw = f. \tag{5.8}$$

LEMMA 5.1. *Let the data $f(x)$ and $u_0(x)$ be smooth, divergence-free functions. Suppose $\delta > 0$; then*

$$\sup_{0 \le T < \infty} [||w(T)|| + ||\nabla w(T)|| + ||\triangle w(T)||] \le C(data, \delta) < \infty. \tag{5.9}$$

Proof. First we note that it has been proven that the ADM turbulence model (1.4) has a unique strong solution that is as smooth as the problem data, so formal manipulations of the model are mathematically justified. The bound on w and ∇w follows from the energy inequality for the model in Proposition 3.1. Taking the inner product of (5.8) with Aw gives

$$\frac{1}{2}\frac{d}{dt}||Aw||^2 + \nu||\nabla Aw||^2 = (Af, w) - (w \cdot \nabla w, Aw).$$

Basic inequalities and the bounds on w and ∇w give

$$|(Af, w)| \le C(data),$$
$$|(w \cdot \nabla w, Aw)| \le C||\nabla w||^2||\nabla Aw||^2 \le C(data, \delta) + \frac{\nu}{2}||\nabla Aw||^2.$$

Thus,

$$\frac{1}{2}\frac{d}{dt}||Aw||^2 + \nu||\nabla Aw||^2 \leq C(data,\delta) + \frac{\nu}{2}||\nabla Aw||^2,$$

and the result follows by a Poincaré inequality and using an integrating factor. ∎

COROLLARY 5.2. *Let the data $f(x)$ and $u_0(x)$ be smooth, divergence-free functions and $N = 0$. If $\delta > 0$, then $|H_{ADM-0}(w)(T)|$ is uniformly bounded.*

Proof. There follows

$$|H_{ADM-0}(w)(T)| \leq C(||w||||\nabla w|| + \delta^2||\nabla w||||\triangle w||) \leq C(data,\delta). \qquad \square$$

In the general case, the analogous estimates hold by essentially the same proofs.

COROLLARY 5.3. *Let the data $f(x)$ and $u_0(x)$ be smooth, divergence-free functions. If $\delta > 0$, then, for all cases $N = 0,1,2,\dots$, $|H_{ADM-0}(w)(T)|$ is uniformly bounded.*

Proof (*sketch*). By using weighted norms, we show by the same proof as the above lemma that the estimate holds:

$$\sup_{0\leq T<\infty} [||w(T)||_N + ||\nabla w(T)||_N + ||\triangle w(T)||_N] \leq C(data,\delta,N) < \infty. \tag{5.10}$$

Using this estimate it follows that

$$|H_{ADM-N}(w)(T)| \leq C(||w||_N||\nabla w||_N + \delta^2||\nabla w||_N||\triangle w||_N) \leq C(data,\delta,N). \qquad \square$$

With these bounds, the final step lacking in the argument from section 5.1 in the NSE case can be carried through successfully.

THEOREM 5.4. *Let $f(x), u_0(x)$ be smooth, divergence-free functions and $\delta > 0$. Then for any $N = 0,1,2,\dots$*

$$|\langle\gamma_{ADM-N}(w)\rangle| \leq \frac{U_N^3}{L_N^2} + \sqrt{2}\,\mathrm{Re}^{-\frac{1}{2}}\left(1+\frac{\delta^2}{L_N^2}\right)^{\frac{1}{2}}\frac{U_N^3}{L_N^2} + \mathrm{Re}^{-1}\left(1+\frac{\delta^2}{L_N^2}\right)\frac{U_N^3}{L_N^2}.$$

Proof. We shall give the proof in the case $N = 0$. The proof for general N is the same as the $N = 0$ case with subscripts N added.

Time average the ADM turbulence model's helicity balance relation; both helicity terms drop out by Corollaries 5.2 ($N = 0$ case) and 5.3 ($N > 0$ case). Thus, when $N = 0$ we have

$$Lim_{T\to\infty}\left|\frac{1}{T}\int_0^T \gamma_{ADM-0}(w)(t)dt\right| \leq Lim_{T\to\infty}\left|\frac{1}{T}\int_0^T \frac{1}{|\Omega|}(\nabla\times f,u)dt\right| \leq \frac{FU}{L}.$$

Inserting the bounds on FU and $\langle\varepsilon_{ADM}(w)\rangle$ from section 4 gives

$$\begin{aligned}
|\langle\gamma_{ADM-0}(w)\rangle| &\leq \frac{FU}{L} \leq \frac{U^3}{L^2} + \frac{U}{L}\left(\frac{\nu}{L^2}+\frac{\nu\delta^2}{L^4}\right)^{\frac{1}{2}}\langle\varepsilon_{ADM-0}(w)\rangle^{\frac{1}{2}} \\
&\leq \frac{U^3}{L^2} + \mathrm{Re}^{-\frac{1}{2}}\left(\frac{U}{L}\right)^{\frac{3}{2}}(1+\delta^2)^{\frac{1}{2}}\left[2\frac{U^3}{L} + \mathrm{Re}^{-1}\frac{U^3}{L}\left(1+\frac{\delta^2}{L^2}\right)\right]^{\frac{1}{2}} \\
&\leq \frac{U^3}{L^2} + \sqrt{2}\,\mathrm{Re}^{-\frac{1}{2}}(1+\delta^2)^{\frac{1}{2}}\frac{U^3}{L^2} + \mathrm{Re}^{-1}(1+\delta^2)\frac{U^3}{L^2},
\end{aligned}$$

as claimed. When $N > 0$ the proof is the same with only subscripts N added. ∎

6. Conclusions. Similarity theories of cascades in homogeneous, isotropic turbulence are based on several assumptions which have yet to be verified directly from the NSE. Nevertheless, the predictions of these theories have been observed in many turbulent flows in nature. As analytic understanding advances, many of these predictions have also been proven directly from the Navier–Stokes equations.

The correctness of the predictions of turbulence models, however, can be very unclear, and simulations based on those models can be even more so. We have considered the energy and helicity dissipation rates of general solutions of one promising family of models. These rates are key quantities in physical theories of energy and helicity cascades. Rigorous upper bounds of the model's energy and helicity dissipation rates are derived. These estimates agree with those proven for energy for the NSE and derived by physical theories of homogeneous, isotropic turbulence. This analysis is based on a rigorous understanding of physical integral invariants of flow models and their corresponding dissipation rates. It gives important (even essential) analytic insight into the reliability of models and the predictions coming from them. Approximate deconvolution models have a systematic mathematical derivation which is reflected in their high-accuracy, exact conservation (when $\nu = 0$) of a model energy and helicity and the validity of their predictions of energy and helicity dissipation rates (when $\nu > 0$).

REFERENCES

[AS01] N. A. Adams and S. Stolz, *Deconvolution methods for subgrid-scale approximation in large eddy simulation*, Modern Simulation Strategies for Turbulent Flow, R. T. Edwards, Philadelphia, PA, 2001.

[AS02] N. A. Adams and S. Stolz, *A subgrid-scale deconvolution approach for shock capturing*, J. Comput. Phys., 178 (2002), pp. 391–426.

[AL77] J. C. Andre and M. Lesieur, *Influence of helicity on high Reynolds number isotropic turbulence*. J. Fluid Mech., 81 (1977), pp. 187–207.

[BB98] M. Bertero and B. Boccacci, *Introduction to Inverse Problems in Imaging*, Institute of Physics, Bristol, 1998.

[BIL06] L. C. Berselli, T. Iliescu, and W. Layton, *Large Eddy Simulation*, Springer, Berlin, 2006.

[BO97] V. Borue and S. Orszag, *Spectra in helical three-dimensional homogeneous isotropic turbulence*, Phys. Rev. E, 55 (1997), pp. 7005–7009.

[BFL73] A. Brissaud, U. Frisch, J. Leorat, M. Lesieur, and A. Mazure, *Helicity cascades in fully developed isotropic turbulence*, Phys. Fluids, 16 (1973), pp. 1366–1367.

[B78] F.H. Busse, *The optimum theory of turbulence*, Adv. Appl. Mech., 18 (1978), pp. 77–121.

[CCE03] Q. Chen, S. Chen, and G. Eyink, *The joint cascade of energy and helicity in three dimensional turbulence*, Phys. Fluids, 15 (2003), pp. 361–374.

[CCEH03] Q. Chen, S. Chen, G. Eyink, and D. Holm, *Intermittence in the joint cascade of energy and helicity*, Phys. Rev. Lett., 90 (2003), p. 214503.

[CDP07] A. Cheskidov, C. Doering, and N. Petrov, *Energy dissipation in fractal-forced flow*, J. Math. Phys., 48 (2007), pp. 1–10.

[CHOT05] A. Cheskidov, D. D. Holm, E. Olson, and E. S. Titi, *On a Leray-α model of turbulence*, Proc. Roy. Soc. Lond. Ser. A Math. Phys. Engrg. Sci., 461 (2005), pp. 629–649.

[CKG01] S. Childress, R. R. Kerswell, and A. D. Gilbert, *Bounds on dissipation for Navier-Stokes flows with Kolmogorov forcing*, Phys. D, 158 (2001), pp. 1–4.

[CD92] P. Constantin and C. Doering, *Energy dissipation in shear driven turbulence*, Phys. Rev. Lett., 69 (1992), pp. 1648–1651.

[DF02] C. Doering and C. Foias, *Energy dissipation in body-forced turbulence*, J. Fluid Mech., 467 (2002), pp. 289–306.

[DG95] C. Doering and J.D. Gibbon, *Applied Analysis of the Navier-Stokes Equations*, Cambridge University Press, Cambridge, 1995.

[DG01a] P. Ditlevsen and P. Giuliani, *Cascades in helical turbulence*, Phys. Rev. E, 63 (2001), p. 036304.

[DG01b] P. Ditlevsen and P. Giuliani, *Dissipation in helical turbulence*, Phys. Fluids, 13 (2001), pp. 3508–3509.

[DE06] A. Dunca and Y. Epshteyn, *On the Stolz-Adams deconvolution model for the large-eddy simulation of turbulent flows*, SIAM J. Math. Anal., 37 (2006), pp. 1890–1902.

[ELN07] V. Ervin, W. Layton, and M. Neda, *Numerical analysis of a higher order time relaxation model of fluids*, Int. J. Numer. Anal. Model., 4 (2007), pp. 648–670.

[F97] C. Foias, *What do the Navier-Stokes equations tell us about turbulence?*, Contemp. Math., 208 (1997), pp. 151–180.

[FHN06] C. Foias, L.T. Hoang, and B. Nicolenko, *On the Helicity in the* 3D *Navier-Stokes Equations,* I, Proc. London Math. Soc., to appear.

[Frisch] U. Frisch, *Turbulence*, Cambridge, Cambridge University Press, 1995.

[Gal95] G.P. Galdi, *An introduction to the Mathematical Theory of the Navier-Stokes Equations, Volume* I, Springer, Berlin, 1994.

[Ger86] M. Germano, *Differential filters of elliptic type*, Phys. Fluids, 29 (1986), pp. 1757–1758.

[Geu97] B. J. Geurts, *Inverse modeling for large eddy simulation*, Phys. Fluids, 9 (1997), p. 3585.

[Gue04] R. Guenanff, *Non-Stationary Coupling of Navier-Stokes/Euler for the Generation and Radiation of Aerodynamic Noises*, Ph.D. thesis, Department of Mathematics, Universite Rennes 1, Rennes, France, 2004.

[H72] L. N. Howard, *Bounds on flow quantities*, Annu. Rev. Fluid Mech., 4 (1972), pp. 473–494.

[J04] V. John, *Large Eddy Simulation of Turbulent Incompressible Flows*, Springer, Berlin, 2004.

[K98] R.R. Kerswell, *Unification of variational methods for turbulent shear flows: The background method of Doering-Constantin and the mean-flow method of Howard-Busse*, Phys. D, 121 (1998), pp. 175–192.

[LL03] W. Layton and R. Lewandowski, *A simple and stable scale similarity model for large eddy simulation: Energy balance and existence of weak solutions*, Appl. Math. Lett., 16 (2003), pp. 1205–1209.

[LL06a] W. Layton and R Lewandowski, *On a well posed turbulence model*, Discrete Contin. Dyn. Sys. Ser. B, 6 (2006), pp. 111–128.

[LL06b] W. Layton and R Lewandowski, *Residual stress of approximate deconvolution large eddy simulation models of turbulence*, J. Turbul., 46 (2006), pp. 1–21.

[LMNR06] W. Layton, C. Manica, M. Neda, and L. Rebholz, *The Joint Helicity-Energy Cascade for Homogeneous, Isotropic Turbulence Generated by Approximate Deconvolution Models*, Technical report, University of Pittsburgh, 2006, http://www.math.pitt.edu/techreports.html.

[LN07a] W. Layton and M. Neda, *Truncation of scales by time relaxation*, J. Math. Anal. Appl., 325 (2007), pp. 788–807.

[LN07b] W. Layton and M. Neda, *The energy cascade for homogeneous, isotropic turbulence generated by approximate deconvolution models*, J. Math. Anal. Appl., 333 (2007), pp. 416–429.

[Les97] M. Lesieur, *Turbulence in Fluids*, Kluwer Academic Publishers, Norwell, MA, 1997.

[LMC05] M. Lesieur, O. Metais, and P. Comte, *Large-Eddy Simulations of Turbulence*, Cambridge University Press, Cambridge, 2005.

[L67] D. K. Lilly, *The representation of small-scale turbulence in numerical simulation experiments*, in Proceedings of the IBM Scientific Computing Symposium on Environmental Sciences, Yorktown Heights, 1967.

[MM06] C. C. Manica and S. Kaya Merdan, *Convergence Analysis of the Finite Element Method for a Fundamental Model in Turbulence*, Technical report, University of Pittsburgh, http://www.math.pitt.edu/techreports.html, 2006, submitted.

[M84] H. Moffatt, *Simple topological aspects of turbulent vorticity dynamics*, in Proceedings of the IUTAM Symposium on Turbulence and Chaotic Phenomena in Fluids, T. Tatsumi, ed., Elsevier, New York, 1984, pp. 223–230,

[MT92] H. Moffatt and A. Tsoniber, *Helicity in laminar and turbulent flows*, Annu. Rev. Fluid Mech., 24 (1992), pp. 281–312.

[M61] J. J. Moreau, *Constantes d'unilot tourbilloinnaire en fluide parfait barotrope*, C.R. Acad. Sci. Paris, 252 (1961), pp. 2810–2812.

[Mus97] A. Muschinski, *A similarity theory of locally homogeneous and isotropic turbulence generated by a Smagorinsky-type LES*, J. Fluid. Mech., 325 (1996), pp. 239–260.

[P00] S. Pope, *Turbulent Flows,* Cambridge University Press, Cambridge, 2000.

[P06] C. Pruett, *Temporal large-eddy simulation: Theory and practice*, Theoret. Computat. Fluid Dyn., to appear.

[R07] L. Rebholz, *Conservation laws of turbulence models*, J. Math. Anal. Appl., 326 (2007), pp. 33–45.

[S01] P. Sagaut, *Large Eddy Simulation for Incompressible Flows*, Springer, Berlin, 2001.

[S84] K. R. Sreenivasan, *On the scaling of the turbulent energy dissipation rate*, Phys. Fluids, 27 (1984), pp. 1048–1051.

[S98] K. R. Sreenivasan, *An update on the energy dissipation rate in isotropic turbulence*, Phys. Fluids, 10 (1998), pp. 528–529.

[SA99] S. Stolz and N. A. Adams, *On the approximate deconvolution procedure for LES*, Phys. Fluids, 11 (1999), pp. 1699–1701.

[SAK01a] S. Stolz, N. A. Adams, and L. Kleiser, *The approximate deconvolution model for LES of compressible flows and its application to shock-turbulent-boundary-layer interaction*, Phys. Fluids, 13 (2001), p. 2985.

[SAK01b] S. Stolz, N. A. Adams, and L. Kleiser, *An approximate deconvolution model for large eddy simulation with application to wall-bounded flows*, Phys. Fluids, 13 (2001), p. 997.

[SAK02] S. Stolz, N. A. Adams, and L. Kleiser, *The approximate deconvolution model for compressible flows: Isotropic turbulence and shock-boundary-layer interaction*, in Advances in LES of Complex Flows, R. Friedrich and W. Rodi, eds., Kluwer, Dordrecht, 2002.

[W97] X. Wang, *The time averaged energy dissipation rates for shear flows*, Phys. D, 99 (1997), pp. 555–563.

SIAM J. MATH. ANAL.
Vol. 39, No. 3, pp. 932–964

SPATIAL DYNAMICS METHODS FOR SOLITARY GRAVITY-CAPILLARY WATER WAVES WITH AN ARBITRARY DISTRIBUTION OF VORTICITY*

M. D. GROVES† AND E. WAHLÉN‡

Abstract. This paper presents existence theories for several families of small-amplitude solitary-wave solutions to the classical two-dimensional water-wave problem in the presence of surface tension and with an arbitrary distribution of vorticity. Moreover, the established local bifurcation diagram for irrotational solitary waves is shown to remain qualitatively unchanged for any choice of vorticity distribution. The hydrodynamic problem is formulated as an infinite-dimensional Hamiltonian system in which the horizontal spatial direction is the timelike variable. A center-manifold reduction technique is employed to reduce the system to a locally equivalent Hamiltonian system with a finite number of degrees of freedom. Homoclinic solutions to the reduced system, which correspond to solitary water waves, are detected by a variety of dynamical systems methods.

Key words. water waves, vorticity, capillarity, bifurcation theory

AMS subject classifications. 76B15, 76B03, 76B45, 35Q35, 35J65, 47J15

DOI. 10.1137/060676040

1. Introduction. The *gravity-capillary water-wave problem* concerns the flow of a perfect fluid of unit density subject to the forces of gravity and surface tension; the fluid is bounded below by a rigid horizontal bottom $\{y = 0\}$ and above by a free surface $\{y = \eta(x, t)\}$, where η depends upon the horizontal spatial coordinate x and time t. *Traveling waves* are waves which propagate from left to right with constant speed c and without change of shape, so that $\eta(x, t) = \eta(x - ct)$. The two principal classes of traveling waves are *Stokes waves*, which are periodic in a frame of reference moving with the wave, and *solitary waves*, which have the property that $\eta(x - ct) \to 0$ as $x - ct \to \pm\infty$. In this paper we construct rigorous existence theories for solitary waves on flows with an arbitrary distribution of vorticity; the physical setting corresponds to waves originating from a distant storm advancing into a shear current.

Working in a frame of reference moving with the wave, let us describe the velocity field $(u(x, y), v(x, y))$ within the fluid domain $D_\eta = \{(x, y) : x \in \mathbb{R}, 0 < y < \eta(x)\}$ in terms of a stream function $\psi(x, y)$ which satisfies $\psi_x = -v$, $\psi_y = u - c$ and suppose that $u < c$, so that $\psi_y < 0$. The vorticity $\omega(x, y) = v_x(x, y) - u_y(x, y)$ is known under this condition to be a function of the stream function ψ, and we specify its distribution by prescribing a *vorticity function* γ such that $\omega = \gamma(\psi)$. The hydrodynamic problem is to solve the nonlinear elliptic equation

$$\Delta\psi = -\gamma(\psi), \qquad 0 < y < \eta(x), \tag{1}$$

subject to the Dirichlet boundary conditions

$$\psi(x, 0) = 0, \tag{2}$$

*Received by the editors November 27, 2006; accepted for publication (in revised form) May 14, 2007; published electronically September 19, 2007.
http://www.siam.org/journals/sima/39-3/67604.html

†Department of Mathematical Sciences, Loughborough University, Loughborough, LE11 3TU, UK (m.d.groves@lboro.ac.uk).
‡Department of Mathematics, Lund University, 22100 Lund, Sweden (ewahlen@maths.lth.se).

$$\psi(x, \eta(x)) = m_0, \tag{3}$$

where $m_0 < 0$ is the constant mass flux, the asymptotic conditions

$$\eta(x) \to d \qquad \text{as } x \to \pm\infty, \tag{4}$$

and the nonlinear boundary condition

$$\frac{1}{2}|\nabla\psi|^2 + g(y-d) - \sigma\left[\frac{\eta_x}{\sqrt{1+\eta_x^2}}\right]_x = \frac{\lambda}{2} \qquad \text{on } y = \eta(x), \tag{5}$$

where g, σ, and d are, respectively, the acceleration due to gravity, the coefficient of surface tension, and the asymptotic depth of the water, and λ is a constant called the Bernoulli constant (e.g., see Constantin and Strauss [11]).

Several existence theories for *irrotational* gravity-capillary solitary waves (where $\gamma = 0$) are available in the literature (see below), and many of them use *spatial dynamics* methods. The phrase "spatial dynamics" refers to an approach where a system of partial differential equations governing a physical problem is formulated as a (typically ill-posed) evolutionary equation

$$u_\xi = L(u) + N(u), \tag{6}$$

in which an unbounded spatial coordinate plays the role of the timelike variable ξ. The water-wave problem has one bounded or semibounded coordinate, namely, the vertical coordinate y; by contrast no restriction is placed upon the behavior of the waves in the horizontal coordinate x, and so this coordinate qualifies as "timelike." One therefore studies the problem using spatial dynamics by formulating it as an evolutionary system of the form (6), where $\xi = x$, in an infinite-dimensional phase space consisting of functions of y. Notice that the hydrodynamic problem is conservative and isotropic in x, and these symmetries manifest themselves in the fact that its spatial dynamics formulation is *Hamiltonian* and *reversible*. In section 2 we derive a formulation of the water-wave problem with an arbitrary choice of $\gamma \in L^2(m_0, 0)$ as a reversible Hamiltonian system and place it in a secure functional-analytic framework.

One particularly useful technique for finding solutions of (6) is known as *center-manifold reduction*. Supposing that L has a finite number of purely imaginary eigenvalues and that certain technical hypotheses are satisfied, one can show that (6) admits an invariant manifold called the *center manifold* which contains all its small, bounded solutions; the dimension of the center manifold is given by the number of purely imaginary eigenvalues. This reduction procedure is explained in detail by Mielke [32] and is shown in section 3.1 to be applicable to our spatial dynamics formulation of the gravity-capillary water-wave problem for fluid of finite depth and with any choice of $\gamma \in H^1(m_0, 0)$ which satisfies

$$\int_0^1 \frac{1}{\sqrt{2\Gamma(s) - 2\Gamma_{\min}}}\,\mathrm{d}s > 1, \qquad \Gamma_{\min} = \min_{s\in[0,1]} \Gamma(s), \tag{7}$$

where

$$\Gamma(s) = \frac{d^2}{m_0}\int_s^1 \gamma(m_0 u)\,\mathrm{d}u.$$

(The additional regularity requirement on γ simplifies the calculation of the spectrum of L.) An important aspect of the center-manifold reduction procedure is that it preserves symmetries of the initial evolutionary equation. This feature can be exploited in existence theories for water waves: Since our spatial dynamics formulation is Hamiltonian and reversible, the reduced system on the center manifold is a reversible Hamiltonian system with finitely many degrees of freedom.

Bifurcation phenomena obtained by varying a parameter can also be captured by the center-manifold reduction procedure. A bifurcation parameter ε may be introduced by perturbing physical parameters present in the problem (in the case of water waves, $\alpha = gd^3/m_0^2$ and $\beta = \sigma d/m_0^2$) around fixed reference values, and the reduction procedure delivers an ε-dependent manifold which captures the small-amplitude dynamics for small values of this parameter; the manifold is a true center manifold at criticality ($\varepsilon = 0$), so that its dimension is the number of purely imaginary eigenvalues of L at $\varepsilon = 0$. The reduction procedure is therefore especially helpful in detecting bifurcations which are associated with a change in the number of purely imaginary eigenvalues. In the case of irrotational waves there are three critical curves C_2, C_3, C_4 in the (β, α) parameter plane at which the number of purely imaginary eigenvalues of L changes (see Figure 1(a)), together with a fourth curve C_1 at which the number of real eigenvalues changes; an explicit parametrization of each of these curves is available (Kirchgässner [29]). Section 3.2 contains spectral theory which shows that this "bifurcation diagram" remains qualitatively unchanged for each choice of vorticity function γ; explicit formulas are, however, available only for $C_3 = \{(\beta, \alpha_\star) : \beta < \beta_\star\}$ and $C_4 = \{(\beta, \alpha_\star) : \beta > \beta_\star\}$, where $\beta_\star$, $\alpha_\star$ are positive constants determined by the choice of γ (the values of $\beta_\star$ and $\alpha_\star$ for $\gamma = 0$ are, respectively, 1/3 and 1).

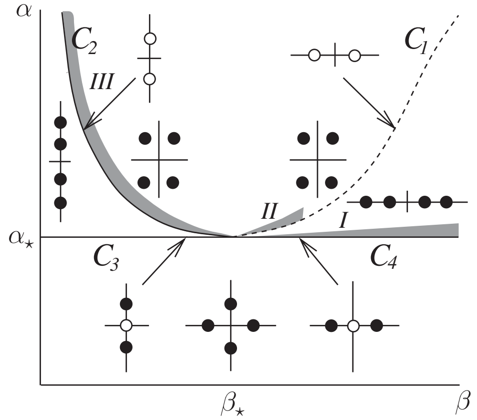
In irrotational water-wave theory the curves C_1, C_2, and C_4 are associated with *homoclinic bifurcation*, where solutions of the reduced Hamiltonian system which are asymptotically zero bifurcate from the trivial solution; these solutions correspond to solitary water waves. Figure 1 illustrates the regions I, II, and III adjacent to, respectively, C_4, C_1, and C_2 in which homoclinic solutions exist for $\gamma = 0$, and in the present paper we show that the same is true for any choice of $\gamma \in H^1(m_0, 0)$ satisfying (7). There are, however, physical differences in the corresponding solitary waves: A small-amplitude irrotational solitary wave is a perturbation of a uniform flow, while a solitary wave with $\gamma \neq 0$ "rides" a laminar flow in which the velocity field is horizontal but depth-dependent.

Irrotational waves in region I were studied by Kirchgässner [29] (see also Amick and Kirchgässner [1], Sachs [33], and Iooss and Kirchgässner [24]). A *Hamiltonian* 0^2*-resonance* takes place at C_4; that is, two imaginary eigenvalues collide at the origin and become real as one crosses the curve from below. Kirchgässner showed that the flow on the two-dimensional center manifold is controlled by the reversible, Hamiltonian equation

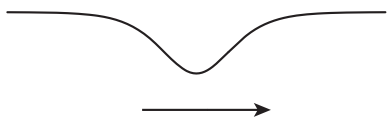
$$u_{xx} = u + \frac{3}{2}u^2 + O(\delta^{1/2}),$$

where $0 < \delta \ll 1$ is the bifurcation parameter $\alpha - \alpha_\star$. This equation admits a homoclinic solution which corresponds to a solitary wave of depression whose tail decays exponentially and monotonically. In section 4.1 we show that Kirchgässner's analysis and conclusions remain valid for an arbitrary choice of $\gamma \neq 0$ and give a geometric interpretation of his method.

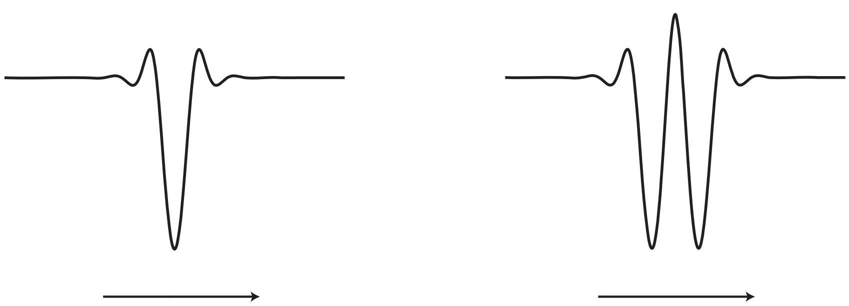
Region II lies on the "complex side" of the curve C_1, at points of which two pairs of small-magnitude real eigenvalues collide and become complex. The center manifold is four-dimensional, and it was shown by Buffoni, Groves, and Toland [8] that for $\gamma = 0$



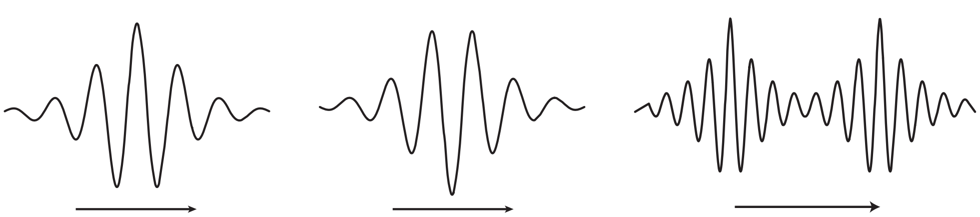
(a) *Bifurcation curves in the* (β, α)*-plane; the shaded regions indicate the parameter regimes in which homoclinic bifurcation is detected.*



(b) *A symmetric solitary wave of depression (left) is found in region* I.



(c) *Region* II *contains an infinite family of multitroughed solitary waves which decay in an oscillatory fashion.*



(d) *Symmetric unipulse modulated solitary waves (left and center) coexist with an infinite family of multipulse modulated solitary waves (right) in region* III.

FIG. 1. *Summary of the solitary waves whose existence is established in the present paper by center-manifold reduction and homoclinic bifurcation theory.*

the flow on the center manifold is controlled by the reversible, Hamiltonian equation

$$u_{xxxx} - 2(1+\delta)u_{xx} + u - u^2 = 0(\mu), \tag{8}$$

where $0 < \mu \ll 1$ measures the distance from the point $(\beta_\star, \alpha_\star)$ and $0 < \delta \ll 1$ is the bifurcation parameter (measuring the distance from C_1). This equation has an infinite family of *multipulse* homoclinic solutions which make several large excursions away from the origin. The corresponding water waves are solitary waves of depression with $2, 3, 4, \ldots$ large troughs separated by $2, 3, \ldots$ small oscillations; their tails are oscillatory and decay exponentially. In section 4.2 we compute the reduced Hamiltonian system for a general choice of γ using a method which is simpler than that employed by Buffoni, Groves, and Toland. We again arrive at (8), so that our hydrodynamic problem also admits a plethora of multipulse solitary waves in this parameter regime.

Region III was first examined by Iooss and Kirchgässner [23], who studied homoclinic bifurcation associated with the *Hamiltonian-Hopf bifurcation* at points of C_2 (two pairs of purely imaginary eigenvalues collide at nonzero points $\pm \mathrm{i}q$ and become complex). The center manifold is four-dimensional at Hamiltonian-Hopf points, and the two-degree-of-freedom reduced Hamiltonian system is conveniently studied using complex coordinates (A, B) and a normal-form transformation. Introducing a bifurcation parameter δ so that positive values of δ correspond to points on the complex side of C_2, one obtains the reduced Hamiltonian system

$$A_x = \frac{\partial H}{\partial \bar{B}}, \qquad B_x = -\frac{\partial H}{\partial \bar{A}},$$

$$H = \mathrm{i}q(A\bar{B} - \bar{A}B) + |B|^2 + H_{\mathrm{NF}}(|A|^2, \mathrm{i}(A\bar{B} - \bar{A}B), \delta) + O(|(A,B)|^2 |(\delta, A, B)|^{n_0}),$$

where H_{NF} is a real polynomial which satisfies $H_{\mathrm{NF}}(0,0,\delta) = 0$; it contains the terms of order $3, \ldots, n_0 + 1$ in the Taylor expansion of H. Supposing that the coefficients of certain terms in H_{NF} have the correct sign, one finds that the "truncated normal form" obtained by neglecting the remainder term admits a circle of homoclinic solutions, two of which persist when the remainder terms are reinstated (see Iooss and Pérouème [26]). The corresponding water waves are symmetric solitary waves which take the form of periodic wave trains modulated by exponentially decaying envelopes. Buffoni and Groves [7] strengthened this result by showing that the above Hamiltonian system in fact has an infinite number of geometrically distinct homoclinic solutions which generically resemble multiple copies of one of the homoclinic solutions found by Iooss and Kirchgässner. The relevant normal-form coefficients were computed for irrotational waves by Buffoni and Groves [7, Appendix B]. Although such explicit formulas are not available, for a general choice of γ it is possible to prove that the coefficients have the correct signs in the local part of region III near $(\beta_\star, \alpha_\star)$; this procedure is carried out in section 4.3.

A different homoclinic bifurcation phenomenon occurs at C_3, where a *Hamiltonian* $0^2\mathrm{i}\omega$ *resonance* takes place; that is, two imaginary eigenvalues collide at the origin and become real as C_3 is crossed from above, while a second pair of eigenvalues remains on the imaginary axis. Lombardi [30] has proved the existence of irrotational *generalized solitary waves* in the region just below C_3; their pulselike profile decays at infinity to a periodic ripple whose amplitude is exponentially small compared to that of the pulse. It is still an open question whether genuine solitary waves exist in this parameter regime, although Sun [34] has recently proved that they do not exist for values of β

close to $\beta_\star$, and there is strong numerical evidence that the same is true for all values of $\beta < \beta_\star$ (Champneys, Vanden-Broeck, and Lord [9]). The corresponding discussion for $\gamma \neq 0$ is beyond the scope of the present paper.

Spatial dynamics techniques are also used in a variety of existence proofs for other types of irrotational gravity-capillary water waves, notably in center-manifold methods for three-dimensional traveling waves (Groves and Mielke [18], Groves [15], Groves and Haragus [16], Groves and Sandstede [19]). On the other hand, center-manifold reduction is not available in other situations, for example, in the existence theories for three-dimensional solitary waves by Groves, Haragus, and Sun [17] and for two-dimensional solitary waves on water of infinite depth by Iooss and Kirrmann [25]; in these cases other methods are used to find solutions of the spatial dynamics formulation of the hydrodynamic problem as an infinite-dimensional evolutionary equation.

The present contribution is one of a series of papers in which established existence theories for irrotational water waves are generalized to flows with arbitrary distributions of vorticity. This program of research began with Constantin and Strauss [11], who generalized Keady and Norbury's [28] proof of the existence of a connected global branch of irrotational symmetric gravity Stokes waves containing waves whose speeds at the crest are arbitrarily small. Wahlén [35] has recently generalized the bifurcation theory for small-amplitude gravity-capillary Stokes waves by Jones [27] (see also Zeidler [36, 37]), while Hur [21] has extended Beale's [2] construction of small-amplitude gravity solitary waves.

2. Formulation as a Hamiltonian system. In this section we formulate the hydrodynamic problem as a reversible Hamiltonian system, the irrotational version of which was outlined by Groves [14] (see also Benjamin [3, Appendix B]). Note that the irrotational version differs from the Hamiltonian system by Groves and Toland [20] which is employed in the center-manifold reduction methods for irrotational water waves described above.

We begin by writing the hydrodynamic problem (1)–(5) in terms of the dimensionless variables

$$(x', y') = \frac{1}{d}(x, y), \quad \eta'(x') = \frac{1}{d}\eta(x), \quad \psi'(x', y') = -\frac{1}{m_0}\psi(x, y)$$

and dimensionless vorticity function

$$\gamma'(\psi') = -\frac{d^2}{m_0}\gamma(\psi).$$

One finds that

$$\Delta\psi = -\gamma(\psi), \qquad 0 < y < \eta(x), \tag{9}$$

$$\psi(x, 0) = 0, \tag{10}$$

$$\psi(x, \eta(x)) = -1, \tag{11}$$

$$\eta(x) \to 1 \qquad \text{as } x \to \pm\infty, \tag{12}$$

and

$$\frac{1}{2}|\nabla\psi|^2 + \alpha(y-1) - \beta\left[\frac{\eta_x}{\sqrt{1+\eta_x^2}}\right]_x = \frac{\mu}{2} \qquad \text{on } y = \eta(x), \tag{13}$$

in which

$$\alpha = \frac{gd^3}{m_0^2}, \quad \beta = \frac{\sigma d}{m_0^2}, \quad \mu = \frac{\lambda d^2}{m_0^2}$$

are dimensionless parameters and the primes have been dropped for notational simplicity. The next step is to map the unknown fluid domain D_η into a fixed strip $\mathbb{R} \times (0,1)$ using a transformation devised by Dubreil-Jacotin [13]. We define $s = -\psi(x,y)$, $h = y$, and treat $(x,s) \in \mathbb{R} \times (0,1)$ as independent variables and $\eta(x)$, $h(x,s)$ as dependent variables. A straightforward calculation shows that (9)–(13) are transformed into

$$\text{(14)} \quad \left[\frac{h_x}{h_s}\right]_x - \left[\frac{1+h_x^2}{2h_s^2}\right]_s + \gamma(-s) = 0, \qquad 0 < s < 1,$$

$$\text{(15)} \quad h(x,0) = 0,$$

$$\text{(16)} \quad h(x,1) = \eta,$$

$$\text{(17)} \quad \eta(x) \to 1 \qquad \text{as } x \to \pm\infty,$$

$$\text{(18)} \quad \frac{1+h_x^2}{2h_s^2} + \alpha(h-1) - \beta\left[\frac{h_x}{\sqrt{1+h_x^2}}\right]_x = \frac{\mu}{2}, \qquad s = 1,$$

and we seek solutions with $h_s > 0$, a condition which is implied by the assumption $\psi_y < 0$ (Constantin and Strauss [11]). The following proposition, which is proved by straightforward arguments from the theory of elliptic boundary-value problems, relates solutions of the transformed equations to those of (9)–(13).

PROPOSITION 2.1. *Define $I = (x_1, x_2)$, $I' = (x_1', x_2')$, with $x_1 < x_1' < x_2' < x_2$, and let $D_{\eta,I} = \{(x,y) : x \in I,\ 0 < y < \eta(x)\}$.*

(i) *Suppose that $\gamma \in L^2(-1,0)$. Any solution $h \in H^2(I \times (0,1)) \cap C^1(I \times [0,1])$ and $\eta = h|_{\{s=1\}} \in C^2(I)$ of (14)–(16), (18) defines a solution $\psi \in H^2(D_{\eta,I'}) \cap C^1(\overline{D_{\eta,I'}})$, $\eta \in C^2(I)$ of (9)–(11), (13).*

(ii) *The additional regularity $\gamma \in C^{k,\alpha}[-1,0]$ and $\eta \in C^{k+2,\alpha}(I)$ for some $\alpha \in (0,1)$ and some nonnegative integer k implies that $\psi \in C^{k+2,\alpha}(\overline{D_{\eta,I'}})$.*

Observe that (14)–(16), (18) follow from the formal variational principle

$$\delta\mathcal{J} = 0,$$

where

$$\text{(19)} \quad \mathcal{J} = \int \left\{ \int_0^1 \left(-\frac{1+h_x^2}{2h_s^2} + \alpha(h-1) - \frac{\mu}{2} - \Gamma(s) \right) h_s \,\mathrm{d}s + \beta\sqrt{1+\eta_x^2} \right\} \mathrm{d}x,$$

$$\Gamma(s) = -\int_s^1 \gamma(-u)\,du, \qquad s \in [0,1],$$

and the variations are taken with respect to $\eta(x)$ and $h(x,s)$ such that $h(x,1) = \eta(x)$ and $h(x,0) = 0$. (The corresponding variational principle for gravity waves was given by Constantin, Sattinger, and Strauss [10].) We exploit this variational principle by regarding $\mathcal{J}$ as an action functional of the form

$$\mathcal{J} = \int J(\eta, h, \eta_x, h_x)\,\mathrm{d}x,$$

in which J is the integrand on the right-hand side of (19), and deriving a Hamiltonian formulation of (14)–(16), (18) by means of the Legendre transform. To this end, let us introduce new variables ω and w by the formulas

$$\omega = \frac{\delta \mathcal{J}}{\delta \eta_x} = \frac{\beta \eta_x}{\sqrt{1+\eta_x^2}}, \qquad w = \frac{\delta \mathcal{J}}{\delta h_x} = -\frac{h_x}{h_s}, \tag{20}$$

in which the variational derivatives are taken in, respectively, $L^2(\mathbb{R})$ and $L^2(\mathbb{R}\times(0,1))$, and define the Hamiltonian function by

$$\begin{aligned} &H(\eta,\omega,h,w) \\ &\quad = \int_0^1 w h_x \,\mathrm{d}s + \omega\eta_x - J(\eta,h,\eta_x,h_x) \\ &\quad = \int_0^1 \left\{ \frac{1}{2}\left(\frac{1}{h_s} - h_s w^2\right) + \Gamma(s) h_s \right\} \mathrm{d}s - \frac{1}{2}\alpha(\eta-1)^2 + \frac{1}{2}\alpha + \frac{\mu}{2}\eta - \sqrt{\beta^2-\omega^2}. \end{aligned} \tag{21}$$

This procedure suggests that the equations

$$\eta_x = \frac{\delta H}{\delta \omega}, \quad \omega_x = -\frac{\delta H}{\delta \eta}, \quad h_x = \frac{\delta H}{\delta w}, \quad w_x = -\frac{\delta H}{\delta h}$$

formally represent Hamilton's equations for a formulation of the hydrodynamic problem (14)–(16), (18) as a Hamiltonian system.

In order to make the above suggestion rigorous, we define the Hilbert spaces

$$\begin{aligned} X &= \{(\eta,\omega,h,w) \in \mathbb{R}\times\mathbb{R}\times H^1(0,1)\times L^2(0,1) : h(0)=0,\ h(1)=\eta\}, \\ Y &= \{(\eta,\omega,h,w) \in \mathbb{R}\times\mathbb{R}\times H^2(0,1)\times H^1(0,1) : h(0)=0,\ h(1)=\eta\} \end{aligned}$$

and consider the symplectic manifold (X,Ω), where Ω is the position-independent 2-form on X given by

$$\Omega|_{(\eta,\omega,h,w)}((\eta_1,\omega_1,h_1,w_1),(\eta_2,\omega_2,h_2,w_2)) = \int_0^1 (w_2 h_1 - w_1 h_2)\,\mathrm{d}s + \omega_2\eta_1 - \omega_1\eta_2$$

(the canonical 2-form with respect to the $\mathbb{R}\times\mathbb{R}\times L^2(0,1)\times L^2(0,1)$-inner product). Choose $\gamma \in L^2(-1,0)$, so that $\Gamma \in H^1(-1,0)$, and observe that the set

$$M = \{(\eta,\omega,h,w) \in Y : |\omega| < \beta,\ \eta > 0,\ h_s(s) > 0 \text{ for each } s \in [0,1]\}$$

is a manifold domain of X and that the function H given by (21) belongs to $C^\infty(M,\mathbb{R})$; a direct calculation shows that

$$\begin{aligned} \mathrm{d}H|_m(v_1|_m) = \int_0^1 \left\{ -\frac{1}{2}\left(w^2 + \frac{1}{h_s^2}\right) + \Gamma(p) \right\} h_{1s}\,\mathrm{d}s - \int_0^1 h_s w w_1 \,\mathrm{d}s \\ - \alpha(\eta-1)\eta_1 + \frac{\mu}{2}\eta_1 + \frac{\omega\omega_1}{\sqrt{\beta^2-\omega^2}} \end{aligned}$$

for $m = (\eta,\omega,h,w) \in M$ and $v_1|_m = (\eta_1,\omega_1,h_1,w_1) \in TM|_m \cong Y$. The triple (X,Ω,H) is therefore a Hamiltonian system.

Recall that the point $m \in M$ belongs to the domain $\mathcal{D}(v_H)$ of the Hamiltonian vector field v_H associated with (X,Ω,H), with $v_H|_m = \bar{v}|_m$, if and only if

$$\Omega_m(\bar{v}|_m, v_1|_m) = \mathrm{d}H|_m(v_1|_m)$$

for all tangent vectors $v_1|_m \in TM|_m \subset TX|_m$. Using this criterion and the above expression for $\mathrm{d}H|_m(v_1|_m)$ and integrating by parts, one finds that

$$\mathcal{D}(v_H) = \left\{ (\eta, \omega, h, w) \in M : \ w(0) = 0, \ h_s(1)w(1) = -\frac{\omega}{\sqrt{\beta^2 - \omega^2}} \right\}$$

and that Hamilton's equations

$$u_x = v_H(u)$$

are given explicitly by

$$\eta_x = \frac{\omega}{\sqrt{\beta^2 - \omega^2}}, \tag{22}$$

$$\omega_x = \frac{1}{2}\left(w^2(1) + \frac{1}{h_s^2(1)} \right) + \alpha(\eta - 1) - \frac{\mu}{2}, \tag{23}$$

$$h_x = -h_s w, \tag{24}$$

$$w_x = -\frac{1}{2}\left(w^2 + \frac{1}{h_s^2} \right)_s + \gamma(-s). \tag{25}$$

Observe that Hamilton's equations are reversible; the reverser $S : X \to X$ is defined by $S(\eta, \omega, h, w) = (\eta, -\omega, h, -w)$.

Proposition 2.2. *Suppose that $(\eta, \omega, h, w) \in C(I, \mathcal{D}(v_H)) \cap C^1(I, X)$, $I = (x_1, x_2)$ solves Hamilton's equations, and let $I' = (x_1', x_2')$, with $x_1 < x_1' < x_2' < x_2$. The functions $\tilde{h}, \tilde{w}$ defined by*

$$\tilde{h}(x, s) = h(x)(s), \qquad \tilde{w}(x, s) = w(x)(s)$$

belong to, respectively, $H^2(D_{\eta,I'}) \cap C^1(\overline{D_{\eta,I'}})$ and $H^1(D_{\eta,I'}) \cap C(\overline{D_{\eta,I'}})$, while η and ω belong to, respectively, $C^2(\overline{I'})$ and $C^1(\overline{I'})$. These functions satisfy $\tilde{h}(x, s) > 0$ in $\overline{D_{\eta,I'}}$, $|\omega| < \beta$ in $\overline{I'}$, the equations

$$\tilde{h}_x = -\tilde{h}_s \tilde{w}, \qquad \tilde{w}_x = -\frac{1}{2}\left(\tilde{w}^2 + \frac{1}{\tilde{h}_s^2} \right)_s + \gamma(-s)$$

in $D_{\eta,I'}$ with boundary conditions

$$\tilde{h}(x, 0) = \tilde{w}(x, 0) = 0, \qquad \tilde{h}(x, 1) = \eta(x), \qquad \tilde{h}_s(x, 1)\tilde{w}(x, 1) = -\frac{\omega}{\sqrt{\beta^2 - \omega^2}},$$

and the equations

$$\eta_x = \frac{\omega}{\sqrt{\beta^2 - \omega^2}},$$

$$\omega_x = \frac{1}{2}\left(\tilde{w}^2 + \frac{1}{\tilde{h}_s^2} \right)\Bigg|_{s=1} + \alpha(\eta - 1) - \frac{\mu}{2}$$

in I'.

The above proposition is proved using the methods given by Groves and Toland [20]. Eliminating ω and $\tilde{w}$ between the above equations, we find that $\tilde{h}$ and η satisfy (14)–(16), (18), and Proposition 2.1 yields a solution of the hydrodynamic problem (9)–(11), (13). Note that the additional regularity $\gamma \in C^{k,\alpha}[0,1]$ and $u \in C^{k+2}(I, \mathcal{D}(v_H)) \cap C^{k+3}(I, X)$ for some $\alpha \in (0,1)$ and some nonnegative integer k

implies that $\psi \in C^{k+2,\alpha}(\overline{D_{\eta,I'}})$. In the remainder of this article we take $\gamma \in H^1(0,1)$ rather than $\gamma \in L^2(0,1)$ in order to simplify the spectral theory presented in section 3.2.

We proceed by seeking solutions $(\eta, \omega, h, w) \in C(\mathbb{R}, \mathcal{D}(v_H)) \cap C^1(\mathbb{R}, X)$ of Hamilton's equations which satisfy $\eta(x) \to 1$ as $x \to \pm\infty$. These solutions take the form of perturbations of equilibrium (that is, x-independent) solutions $(\eta_0, \omega_0, h_0(s), w_0(s))$, where necessarily $\eta_0 = 1$ and $\omega_0 = 0$, $w_0 = 0$ (see (20)); our solitary waves therefore ride a horizontal laminar flow (which is, in general, not uniform). The requirement that the hydrodynamic problem admits a horizontal laminar flow for a given vorticity function fixes the value of the Bernoulli constant μ.

LEMMA 2.3. *Suppose that*

$$\int_0^1 \frac{1}{\sqrt{2\Gamma(s) - 2\Gamma_{\min}}} \, \mathrm{d}s > 1, \tag{26}$$

where

$$\Gamma_{\min} = \min_{s \in [0,1]} \Gamma(s), \qquad \Gamma_{\max} = \max_{s \in [0,1]} \Gamma(s).$$

There exists a unique value $\mu^\star > -2\Gamma_{\min}$ of μ for which Hamilton's equations (22)–(25) *admit a solution of the form*

$$(\eta, \omega, h, w) = (1, 0, \theta(s), 0) \tag{27}$$

for all $\beta, \alpha > 0$. The function $\theta(s)$ is given by the formula

$$\theta(s) = \int_0^s a^{-1}(u) \, \mathrm{d}u, \qquad a(s) = \sqrt{\mu^\star + 2\Gamma(s)}.$$

Proof. Any solution of (22)–(25) of the form (27) satisfies

$$\frac{1}{2}\left(\frac{1}{\theta_s^2}\right)_s = \gamma(-s), \qquad \frac{1}{\theta_s^2(1)} = \mu, \qquad \theta(0) = 0,$$

so that

$$\theta(s) = \int_0^s \frac{1}{\sqrt{\mu + 2\Gamma(u)}} \, \mathrm{d}u,$$

where we have assumed that $\mu > -2\Gamma_{\min}$. It remains to satisfy the boundary condition $h(1) = \eta$; that is, $\theta(1) = 1$. Observe that $\theta(1)$ is a strictly decreasing function of $\mu \in [-2\Gamma_{\min}, \infty)$ which satisfies $\theta(1) \to 0$ as $\mu \to \infty$ and $\theta(1)|_{\mu = -2\Gamma_{\min}} > 1$ under the assumption (26). It follows that there exists a unique value $\mu^\star$ of μ such that $\theta(1) = 1$. $\square$

In accordance with Lemma 2.3 we take $\mu = \mu^\star$ and seek solutions of Hamilton's equations for (X, Ω, H) of the form

$$\eta = 1 + \rho, \qquad h = \theta + \phi,$$

where $\rho > -1$ and $\phi_s(s) > -a^{-1}(s)$ for $s \in [0,1]$. Let us write

$$(\beta, \alpha) = (\beta_0 + \varepsilon_1, \alpha_0 + \varepsilon_2),$$

where (β_0, α_0) is fixed and $(\varepsilon_1, \varepsilon_2)$ lies in a neighborhood Λ of the origin in $\mathbb{R}^2$, and consider solutions (ρ, ω, ϕ, w) which lie in a neighborhood V of the origin in Y; here Λ and V are chosen small enough so that

$$|\varepsilon_1| < \frac{\beta_0}{4}, \quad \rho > -\frac{1}{2} > -1, \quad |\omega| < \frac{\beta_0}{2} < \beta_0 + \varepsilon_1,$$
$$\phi_s(s) > -\frac{1}{2}(\mu^\star + 2\Gamma_{\max})^{-1/2} > -a^{-1}(s)$$

for each $s \in [0,1]$. This change of variable transforms (X, Ω, H) into $(X, \Omega, H^\varepsilon)$, where $H^\varepsilon \in C^\infty(V, \mathbb{R})$ is defined by the formula

$$\begin{aligned} H^\varepsilon(\rho, \omega, \phi, w) = &\int_0^1 \left\{ \frac{1}{2} \left(\frac{1 - (a^{-1}(s) + \phi_s)^2 w^2}{a^{-1}(s) + \phi_s} \right) - \frac{1}{2} a(s) + \Gamma(s)\phi_s \right\} \mathrm{d}s \\ &- \frac{1}{2}(\alpha_0 + \varepsilon_2)\rho^2 + \frac{1}{2}\mu^\star \rho + (\beta_0 + \varepsilon_1) - \sqrt{(\beta_0 + \varepsilon_1)^2 - \omega^2} \end{aligned}$$

(a constant term has also been added to the Hamiltonian to ensure that $H^\varepsilon(0) = 0$). Hamilton's equations (22)–(25) are transformed into

$$\rho_x = \frac{\omega}{\sqrt{(\beta_0 + \varepsilon_1)^2 - \omega^2}}, \tag{28}$$

$$\omega_x = \frac{1}{2} \left(w^2(1) + \frac{a^2(1)}{(1 + a(1)\phi_s(1))^2} \right) + (\alpha_0 + \varepsilon_2)\rho - \frac{\mu^\star}{2}, \tag{29}$$

$$\phi_x = -(a^{-1}(s) + \phi_s)w, \tag{30}$$

$$w_x = -\frac{1}{2} \left(w^2 + \frac{a^2(s)}{(1 + a(s)\phi_s)^2} \right)_s + \gamma(-s), \tag{31}$$

the domain $\mathcal{D}(v_{H^\varepsilon})$ of the Hamiltonian vector field on the right-hand side of this system of equations is the set of elements $(\rho, \omega, \phi, w) \in V$ which satisfy

$$\begin{aligned} w(0) &= 0, \\ (\phi_s(1) + a^{-1}(1))w(1) &= -\frac{\omega}{\sqrt{(\beta + \varepsilon_1)^2 - \omega^2}}, \end{aligned} \tag{32}$$

and the action of the reverser $S : X \to X$ is given by $S(\rho, \omega, \phi, w) = (\rho, -\omega, \phi, -w)$. Our task is to find homoclinic solutions of the above equations, that is, solutions $(\rho, \omega, \phi, w) \in C(\mathbb{R}, \mathcal{D}(v_{H^\varepsilon})) \cap C^1(\mathbb{R}, X)$ which satisfy

$$(\rho(x), \omega(x), \phi(x), w(x)) \to (0, 0, 0, 0)$$

as $x \to \pm\infty$.

3. Center-manifold reduction.

3.1. Application of the reduction theorem. Our strategy in finding solutions to Hamilton's equations (28)–(31) for $(X, \Omega, H^\varepsilon)$ consists in applying a reduction principle which asserts that $(X, \Omega, H^\varepsilon)$ is locally equivalent to a finite-dimensional Hamiltonian system. The key result is the following theorem, which is a parametrized, Hamiltonian version of a reduction principle for quasi-linear evolutionary equations presented by Mielke [31, Theorem 4.1] (see Buffoni, Groves, and Toland [8, Theorem 4.1]).

THEOREM 3.1. *Consider the reversible differential equation*

$$u_x = \mathcal{L}u + \mathcal{N}(u;\lambda), \tag{33}$$

which is Hamilton's equation for the Hamiltonian system $(X, \Omega^\lambda, H^\lambda)$. *Here* u *belongs to a Hilbert space* $\mathcal{X}$, $\lambda \in \mathbb{R}^\ell$ *is a parameter, and* $\mathcal{L} : \mathcal{D}(\mathcal{L}) \subset \mathcal{X} \to \mathcal{X}$ *is a densely defined, closed linear operator. Regarding* $\mathcal{D}(\mathcal{L})$ *as a Hilbert space equipped with the graph norm, suppose that* 0 *is an equilibrium point of* (33) *when* $\lambda = 0$ *and that*

(H1) *the part of the spectrum* $\sigma(\mathcal{L})$ *of* $\mathcal{L}$ *which lies on the imaginary axis consists of a finite number of eigenvalues of finite multiplicity and is separated from the rest of* $\sigma(\mathcal{L})$ *in the sense of Kato, so that* $\mathcal{X}$ *admits the decomposition* $\mathcal{X} = \mathcal{X}_1 \oplus \mathcal{X}_2$, *where* $\mathcal{X}_1 = \mathcal{P}(\mathcal{X})$, $\mathcal{X}_2 = (I - \mathcal{P})(\mathcal{X})$, *and* $\mathcal{P}$ *is the spectral projection corresponding the purely imaginary part of* $\sigma(\mathcal{L})$;

(H2) *the operator* $\mathcal{L}_2 = \mathcal{L}|_{\mathcal{X}_2}$ *satisfies the estimate*

$$\|(\mathcal{L}_2 - \mathrm{i}\xi I)^{-1}\|_{\mathcal{X}_2 \to \mathcal{X}_2} \le \frac{C}{1 + |\xi|}, \qquad \xi \in \mathbb{R}$$

for some constant C *that is independent of* ξ;

(H3) *there exists a natural number* k *and neighborhoods* $\Lambda \subset \mathbb{R}^\ell$ *of* 0 *and* $U \subset \mathcal{D}(\mathcal{L})$ *of* 0 *such that* $\mathcal{N}$ *is* $(k+1)$ *times continuously differentiable on* $U \times \Lambda$, *its derivatives are bounded and uniformly continuous on* $U \times \Lambda$, *and* $\mathcal{N}(0,0) = 0$, $\mathrm{d}_1\mathcal{N}[0,0] = 0$.

Under these hypotheses there exist neighborhoods $\tilde{\Lambda} \subset \Lambda$ *of* 0 *and* $\tilde{U}_1 \subset U \cap \mathcal{X}_1$, $\tilde{U}_2 \subset U \cap \mathcal{X}_2$ *of* 0 *and a reduction function* $r : \tilde{U}_1 \times \tilde{\Lambda} \to \tilde{U}_2$ *with the following properties. The reduction function* r *is* k *times continuously differentiable on* $\tilde{U}_1 \times \tilde{\Lambda}$, *its derivatives are bounded and uniformly continuous on* $\tilde{U}_1 \times \tilde{\Lambda}$, *and* $r(0;0) = 0$, $\mathrm{d}_1 r[0;0] = 0$. *The graph* $\tilde{X}^\lambda = \{u_1 + r(u_1;\lambda) \in \tilde{U}_1 \times \tilde{U}_2 : u_1 \in \tilde{U}_1\}$ *is a Hamiltonian center manifold for* (33), *so that*

(i) $\tilde{X}^\lambda$ *is a locally invariant manifold of* (33): *Through every point in* $\tilde{X}^\lambda$ *there passes a unique solution of* (33) *that remains on* $\tilde{X}^\lambda$ *as long as it remains in* $\tilde{U}_1 \times \tilde{U}_2$;

(ii) *every small bounded solution* $u(x)$, $x \in \mathbb{R}$ *of* (33) *satisfying* $(u_1(x), u_2(x)) \in \tilde{U}_1 \times \tilde{U}_2$ *lies completely in* $\tilde{X}^\lambda$;

(iii) *every solution* $u_1 : (x_1, x_2) \to \tilde{U}_1$ *of the reduced equation*

$$u_{1x} = \mathcal{L}u_1 + \mathcal{P}\mathcal{N}(u_1 + r(u_1;\lambda);\lambda) \tag{34}$$

generates a solution

$$u(x) = u_1(x) + r(u_1(x);\lambda) \tag{35}$$

of the full equation (33);

(iv) $\tilde{X}^\lambda$ *is a symplectic submanifold of* X, *and the flow determined by the Hamiltonian system* $(\tilde{X}^\lambda, \tilde{\Omega}^\lambda, \tilde{H}^\lambda)$, *where the tilde denotes restriction to* $\tilde{X}^\lambda$, *coincides with the flow on* $\tilde{X}^\lambda$ *determined by* $(X, \Omega^\lambda, H^\lambda)$. *The reduced equation* (34) *is reversible and represents Hamilton's equations for* $(\tilde{X}^\lambda, \tilde{\Omega}^\lambda, \tilde{H}^\lambda)$.

The center manifold $\tilde{X}^\lambda$ is equipped with the single coordinate chart $\tilde{U}_1 \subset \mathcal{X}_1$ and coordinate map $\chi : \tilde{X}^\lambda \to \tilde{U}_1$ defined by $\chi^{-1}(u_1) = u_1 + r(u_1;\lambda)$. It is, however, more convenient to use an alternative coordinate map for calculations. According to the parameter-dependent version of Darboux's theorem presented by Buffoni and Groves [7, Theorem 4], there exists a near-identity change of variable

$$\tilde{u}_1 = u_1 + \Theta(u_1;\lambda) \tag{36}$$

of class C^{k-1} which transforms $\tilde{\Omega}^\lambda$ into Ψ, where

$$\Psi(v_1, v_2) = \Omega^0|_0(v_1, v_2).$$

Define the function $\tilde{r} : \tilde{U}_1 \times \tilde{\Lambda} \to \tilde{U}_1 \times \tilde{U}_2$ (which, in general, has components in $\mathcal{X}_1$ and $\mathcal{X}_2$) by the formula

$$\tilde{u}_1 + \tilde{r}(\tilde{u}_1; \lambda) = \mathcal{T}(u_1 + r(u_1; \lambda); \lambda), \qquad \mathcal{T}(u_1, u_2; \lambda) = (u_1 + \Theta(u_1; \lambda), u_2),$$

where $\tilde{r}(0; 0) = 0$, $\mathrm{d}_1\tilde{r}[0; 0] = 0$, and equip $\tilde{X}^\lambda$ with the coordinate map $\tilde{\chi} : \tilde{X}^\lambda \to \tilde{U}_1$ given by $\tilde{\chi}^{-1}(\tilde{u}_1) = \tilde{u}_1 + \tilde{r}(\tilde{u}_1; \lambda)$. One can always choose a basis for $\mathcal{X}_1$ so that Ψ is the canonical symplectic 2-form Υ in this coordinate system (a "symplectic basis"). The choice of $\tilde{\chi}$ as coordinate map therefore yields a finite-dimensional canonical Hamiltonian system.

Theorem 3.1 cannot be applied directly to (28)–(29) because of the nonlinear boundary condition (32). We overcome this difficulty by using the change of variable $(\rho, \zeta, \phi, z) = G^\varepsilon(\rho, \omega, \phi, w)$, where

$$\zeta = -\frac{a(1)\omega}{\sqrt{(\beta + \varepsilon_1)^2 - \omega^2}}, \qquad z(s) = (1 + a(s)\phi_s(s))w(s),$$

which transforms the nonlinear boundary condition into the linear condition $z(1) = \zeta$. The following lemma shows that G^ε defines a valid change of variable.

LEMMA 3.2.

(i) *For each $\varepsilon \in \Lambda$ the mapping G^ε is a smooth diffeomorphism from the neighborhood V of 0 in Y onto itself.*

(ii) *For each $(v, \varepsilon) \in V \times \Lambda$ the operators $\mathrm{d}G^\varepsilon[v]$, $\mathrm{d}((G^\varepsilon)^{-1})[v] : Y \to Y$ extend to mappings $X \to X$ that depend smoothly upon $(v, \varepsilon) \in V \times \Lambda$.*

Proof. These results follow from the explicit formula

$$G^{-1}(\rho, \zeta, \phi, z) = (\rho, \omega, \phi, w),$$

where

$$\omega = -\frac{(\beta_0 + \varepsilon_1)\zeta}{\sqrt{\zeta^2 + a^2(1)}}, \qquad w(s) = \frac{z(s)}{1 + a(s)\phi_s(s)},$$

and

$$\begin{aligned}
&\mathrm{d}G^\varepsilon[\rho, \omega, \phi, w](\rho_1, \omega_1, \phi_1, w_1)\\
&\quad = \left(\rho_1, -\frac{a(1)(\beta_0 + \varepsilon_1)^2\omega_1}{((\beta_0 + \varepsilon_1)^2 - \omega^2)^{3/2}}, \phi_1, (1 + a(s)\phi_s)w_1 + a(s)w\phi_{1s}\right),\\
&\mathrm{d}((G^\varepsilon)^{-1}))[\rho, \zeta, \phi, z](\rho_1, \zeta_1, \phi_1, z_1)\\
&\quad = \left(\rho_1, -\frac{(\beta_0 + \varepsilon_1)a^2(1)\zeta_1}{(a^2(1) + \zeta^2)^{3/2}}, \phi_1, \frac{z_1}{1 + a(s)\phi_s} - \frac{a(s)z\phi_{1s}}{(1 + a(s)\phi_s)^2}\right). \qquad \square
\end{aligned}$$

An explicit calculation shows that the above change of variable transforms (28)–(31) into

$$\rho_x = -a^{-1}(1)\zeta, \tag{37}$$

$$\zeta_x = -\frac{(a^2(1) + \zeta^2)^{3/2}}{a^2(1)(\beta_0 + \varepsilon_1)}\left(\frac{z^2(1) + a^2(1)}{2(1 + a(1)\phi_s(1))^2} + (\alpha_0 + \varepsilon_2)\rho - \frac{\mu^\star}{2}\right), \tag{38}$$

$$\phi_x = -a^{-1}(s)z, \tag{39}$$

$$z_x = -\frac{a(s)z(a^{-1}(s)z)_s}{1 + a(s)\phi_s} - (1 + a(s)\phi_s)\left(\left(\frac{z^2 + a^2(s)}{2(1 + a(s)\phi_s)^2}\right)_s - \gamma(-s)\right); \tag{40}$$

these equations hold in the phase space X, and the domain of the vector field on their right-hand side is a neighborhood of the origin in the linear space

$$Z = \{(\rho, \zeta, \phi, z) \in Y : \ z(0) = 0, \ z(1) = \zeta\}.$$

Notice that the change of variable preserves the reversibility; the action of the reverser $S : X \to X$ is given by $S(\rho, \zeta, \phi, z) = (\rho, -\zeta, \phi, -z)$. Equations (37)–(40) represent Hamilton's equations for the Hamiltonian system $(X, \Phi^\varepsilon, K^\varepsilon)$, where

$$\begin{aligned}\Phi^\varepsilon|_{(\rho,\zeta,\phi,z)}&((\rho_1, \zeta_1, \phi_1, z_1), (\rho_2, \zeta_2, \phi_2, z_2)) \\ &= \int_0^1 \left\{ \frac{z_2\phi_1 - z_1\phi_2}{1 + a(s)\phi_s} - \frac{a(s)z}{(1 + a(s)\phi_s)^2}(\phi_{2s}\phi_1 - \phi_{1s}\phi_2) \right\} \mathrm{d}s \\ &\quad - \frac{(\beta_0 + \varepsilon_1)a^2(1)}{(\zeta^2 + a^2(1))^{3/2}}(\zeta_2\rho_1 - \zeta_1\rho_2)\end{aligned}$$

and $K^\varepsilon \in C^\infty(V, \mathbb{R})$ is defined by the formula

$$\begin{aligned}K^\varepsilon(\rho, \zeta, \phi, z) = \int_0^1 &\left\{ \frac{a(s) - a^{-1}(s)z^2}{2(1 + a(s)\phi_s)} + \Gamma(s)\phi_s - \frac{a(s)}{2} \right\} \mathrm{d}s \\ &- \frac{1}{2}(\alpha_0 + \varepsilon_2)\rho^2 + \frac{1}{2}\mu^\star\rho - \frac{a(1)(\beta_0 + \varepsilon_1)}{\sqrt{a^2(1) + \zeta^2}} + (\beta_0 + \varepsilon_1).\end{aligned}$$

The next step is to verify that (37)–(40) satisfy the hypotheses of Theorem 3.1. We write these equations as

$$u_x = Lu + N^\varepsilon(u), \tag{41}$$

in which the linear operator $L : \mathcal{D}(L) \subset X \to X$, with $\mathcal{D}(L) = Z$, is given by

$$L \begin{pmatrix} \rho \\ \zeta \\ \phi \\ z \end{pmatrix} = \begin{pmatrix} -a^{-1}(1)\zeta \\ a^4(1)\beta_0^{-1}\phi_s(1) - a(1)\alpha_0\beta_0^{-1}\rho \\ -a^{-1}(s)z \\ (a^3(s)\phi_s)_s \end{pmatrix} \tag{42}$$

(the linearization of the Hamiltonian vector field v_{K^ε} at $\varepsilon = 0$). It follows from Proposition 3.3 and Lemma 3.4, the former of which is proved using the elementary theory of ordinary differential equations, that L satisfies hypotheses (H1) and (H2); hypothesis (H3) is clearly satisfied for an arbitrary value of k.

Proposition 3.3. *The spectrum of the operator $L : \mathcal{D}(L) \subset X \to X$ consists of isolated, geometrically simple eigenvalues of finite algebraic multiplicity.*

Lemma 3.4. *There exist real constants C, $\xi_0 > 0$ such that each solution $v \in Y$ of the resolvent equation*

$$(L - \mathrm{i}\xi I)v = f^\star, \tag{43}$$

in which $f^\star$ belongs to X and ξ is a real number with $|\xi| > \xi_0$, satisfies the estimates

$$\|v\|_Y \le C\|f^\star\|_X, \qquad \|v\|_X \le \frac{C}{|\xi|}\|f^\star\|_X.$$

Proof. Let us write the resolvent equation (43) in the form

$$-a^{-1}(1)\zeta - \mathrm{i}\xi\rho = \rho^\star, \tag{44}$$

$$a^4(1)\beta_0^{-1}\phi_s(1) - a(1)\alpha_0\beta_0^{-1}\rho - \mathrm{i}\xi\zeta = \zeta^\star, \tag{45}$$

$$-a^{-1}(s)z - \mathrm{i}\xi\phi = \phi^\star, \tag{46}$$

$$(a^3(s)\phi_s)_s - \mathrm{i}\xi z = z^\star, \tag{47}$$

where $(\rho^\star, \omega^\star, \phi^\star, z^\star) \in X$ and $(\rho, \omega, \phi, z) \in Z$, so that $\phi^\star(0) = 0$, $\phi^\star(1) = \rho^\star$, and $\phi(0) = 0$, $\phi(1) = \rho$, $z(0) = 0$, $z(1) = \zeta$.

The first step is to differentiate (46) and multiply by $a^{3/2}(s)$, and multiply (47) by $a^{-1/2}(s)$, so that

$$\begin{aligned} -a^{3/2}(s)(a^{-1}(s)z)_s - \mathrm{i}\xi a^{3/2}(s)\phi_s &= a^{3/2}(s)\phi_s^\star, \\ a^{-1/2}(s)(a^3(s)\phi_s)_s - \mathrm{i}\xi a^{-1/2}(s)z &= a^{-1/2}(s)z^\star; \end{aligned}$$

squaring and adding these equations, one finds that

$$\begin{aligned} &a^3(s)|(a^{-1}(s)z)_s|^2 + a^{-1}(s)|(a^3(s)\phi_s)_s| + \xi^2(a^3(s)|\phi_s|^2 + a^{-1}(s)|z|^2) \\ &\quad + 2\xi\,\mathrm{Im}\,\{(a^{-1}(s)z)_s a^3(s)\overline{\phi}_s + a^{-1}(s)z(a^3(s)\overline{\phi}_s)_s\} = a^3(s)|\phi_s^\star|^2 + a^{-1}(s)|z^\star|^2. \end{aligned}$$

Let $\|\cdot\|$ and $\|\cdot\|_w$ denote, respectively, the usual $L^2(0,1)$-norm and the $L^2(0,1)$-norm with weight function w. We integrate the above equation over $(0,1)$, so that

$$\begin{aligned} &\|(a^{-1}(s)z)_s\|_{a^3}^2 + \|(a^3(s)\phi_s)_s\|_{a^{-1}}^2 + \xi^2(\|\phi_s\|_{a^3}^2 + \|z\|_{a^{-1}}^2) \\ &\qquad = \|\phi_s^\star\|_{a^3}^2 + \|z^\star\|_{a^{-1}}^2 - 2\xi a^2(1)\mathrm{Im}\,\{z(1)\overline{\phi}_s(1)\}, \end{aligned} \tag{48}$$

and multiplying (45) by $\overline{\phi}_s(1)$ and taking real parts, we find that

$$\begin{aligned} \xi\,\mathrm{Im}\,\{z(1)\overline{\phi}_s(1)\} &= -\frac{a^4(1)}{\beta_0}|\phi_s(1)|^2 + \frac{a(1)\alpha_0}{\beta_0}\mathrm{Re}\,\{\phi(1)\overline{\phi}_s(1)\} + \mathrm{Re}\,\{\zeta^\star\overline{\phi}_s(1)\} \\ &\le C_1(|\phi_s(1)|^2 + |\phi(1)|^2 + |\zeta^\star|^2). \end{aligned} \tag{49}$$

To estimate $|\phi_s(1)|^2$ we multiply (47) by s^n, where n is a positive integer, and integrate over $(0,1)$, so that

$$a^3(1)\phi_s(1) = \int_0^1 a^3(s)\phi_s(s)ns^{n-1}\,\mathrm{d}s + \mathrm{i}\xi\int_0^1 z(s)s^n\,\mathrm{d}s + \int_0^1 z^\star(s)s^n\,\mathrm{d}s,$$

and an application of the Cauchy–Schwarz inequality yields

$$|\phi_s(1)|^2 \le C_2\left(\frac{n^2}{2n-1}\|\phi_s\|_{a^6}^2 + \frac{\xi^2}{2n+1}\|z\|^2 + \frac{1}{2n+1}\|z^\star\|^2\right). \tag{50}$$

A straightforward combination of estimates (48)–(50), the inequality $|\phi(1)|^2 \le \|\phi_s\|^2$, and the fact that $\|\cdot\|$ and $\|\cdot\|_w$ are equivalent norms for any positive $w \in C[0,1]$ shows that

$$\begin{aligned} &\|\phi_{ss}\|^2 + \|\phi_s\|^2 + \|z_s\|^2 + \|z\|^2 + \xi^2(\|\phi_s\|^2 + \|z\|^2) \\ &\qquad \le C_3\left(\frac{n^2}{2n-1}\|\phi_s\|^2 + \frac{\xi^2}{2n+1}\|z\|^2 + \|\phi_s^\star\|^2 + \|z^\star\|^2 + |\zeta^\star|^2\right). \end{aligned}$$

Choosing n large enough so that $C_3/(2n+1) \leq 1/2$, we therefore find that

$$\|\phi_{ss}\|^2 + \|\phi_s\|^2 + \|z_s\|^2 + \|z\|^2 + \frac{\xi^2}{2}(\|\phi_s\|^2 + \|z\|^2) \leq C_3(|\zeta^\star|^2 + \|\phi_s^\star\|^2 + \|z^\star\|^2)$$

and hence that

$$\|\phi\|_{H^2}^2 + \|z\|_{H^1}^2 + \frac{\xi^2}{2}(\|\phi\|_{H^1}^2 + \|z\|^2) \leq C_4(|\zeta^\star|^2 + \|\phi^\star\|_{H^1}^2 + \|z^\star\|^2)$$

for sufficiently large ξ, where we have exploited the fact that $\phi(0)$, $\phi^\star(0)$, and $z(0)$ all vanish.

It follows from (45) that

$$\xi^2|\zeta|^2 \leq C_4(|\zeta^\star|^2 + |\phi(1)|^2 + |\phi_s(1)|^2) \leq C_5(|\zeta^\star|^2 + \|\phi\|_{H^1}^2 + \xi^2\|z\|^2 + \|z^\star\|^2)$$

and clearly $|\rho|^2 = |\phi(1)|^2 \leq \|\phi\|_{H^1}^2$, so that altogether

$$\|\phi\|_{H^2}^2 + \|z\|_{H^1}^2 + \frac{\xi^2}{2}(|\rho|^2 + |\zeta|^2 + \|\phi\|_{H^1}^2 + \|z\|^2) \leq C_6(|\zeta^\star|^2 + \|\phi^\star\|_{H^1}^2 + \|z^\star\|^2). \qquad \square$$

3.2. Eigenvalues of the linearized problem. In this section we examine the spectrum of L in more detail, in particular the qualitative dependence of its eigenvalues upon (β_0, α_0). Eliminating ρ, ζ, and z, we find that the eigenvalue problem $Lu = \kappa u$ is equivalent to

$$-a^{-1}(s)(a^3(s)\phi_s)_s = \kappa^2\phi, \qquad 0 < s < 1, \tag{51}$$

$$-\frac{a^3(1)}{\beta}\phi_s(1) + \frac{\alpha}{\beta}\phi(1) = \kappa^2\phi(1), \tag{52}$$

$$\phi(0) = 0 \tag{53}$$

(for convenience we drop the subscript 0 in this section). The change of variable

$$y = \int_0^s a^{-1}(u)\,\mathrm{d}u, \qquad v(y) = a(s)\phi(s)$$

transforms the above equations into the equivalent non-self-adjoint Sturm–Liouville problem

$$-v_{yy} + Q(y)v = \nu v, \tag{54}$$

$$\frac{v_y(1)}{v(1)} = \hat{\alpha} - \hat{\beta}\nu, \tag{55}$$

$$v(0) = 0, \tag{56}$$

where $\nu = \kappa^2$, $Q(y) = -\gamma'(-s)$, $\hat{\alpha} = a'(1) + a^{-2}(1)\alpha$, and $\hat{\beta} = a^{-2}(1)\beta$; detailed spectral results for problems of this type have been presented by Binding et al. [4].

The Sturm–Liouville problem (54)–(56) has a countable number of geometrically simple eigenvalues which can be listed as $\nu_0, \nu_1, \nu_2, \ldots,$ where $\operatorname{Re}\nu_n \leq \operatorname{Re}\nu_{n+1}$ (entries are repeated according to their algebraic multiplicity); they occur in complex-conjugate pairs and satisfy

$$\nu_n = n^2\pi^2 + \int_0^1 Q(y)\,\mathrm{d}y - \frac{2\hat{\alpha}}{\hat{\beta}} + o\left(\frac{1}{n}\right), \qquad n \in \mathbb{N}_0. \tag{57}$$

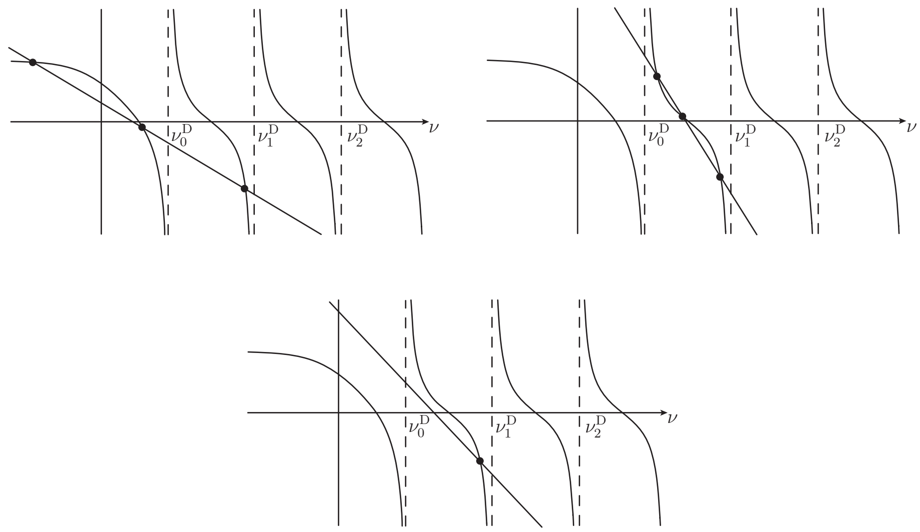


FIG. 2. *Geometric characterization of the eigenvalues ν_n as the points of intersection of the curve $s = B(\nu)$ with the straight line $s = \hat{\alpha} - \hat{\beta}\nu$; one real eigenvalue lies in each interval $(\nu_{n-1}^{\mathrm{D}}, \nu_n^{\mathrm{D}})$, $n \in \mathbb{N}$. Clockwise from top left: two additional real eigenvalues to the left of ν_0^{D}; two additional real eigenvalues in the interval $(\nu_{n-1}^{\mathrm{D}}, \nu_n^{\mathrm{D}})$ for some $n \in \mathbb{N}$; two additional complex eigenvalues.*

Observe that the real eigenvalues of the spectral problem (54)–(56) correspond to the intersections in the (ν, s) plane of the line $s = \hat{\alpha} - \hat{\beta}\nu$ and the curve $s = B(\nu)$, where $B(\nu) = v_y(1;\nu)/v(1;\nu)$ and $v(y;\nu)$ solves the initial-value problem

$$-v_{yy} + Q(y)v = \nu v, \qquad v(0;\nu) = 0;$$

a tangent intersection indicates that the eigenvalue has algebraic multiplicity 2. The function $B(\nu) = v_y(1;\nu)/v(1;\nu)$ has poles exactly at the *Dirichlet eigenvalues* ν_n^{D} (the necessarily real, positive, and algebraically simple eigenvalues of the self-adjoint problem in which (55) is replaced by $v(1) = 0$); it is strictly decreasing from $+\infty$ to $-\infty$ in each interval $(-\infty, \nu_0^{\mathrm{D}})$ and $(\nu_{k-1}^{\mathrm{D}}, \nu_k^{\mathrm{D}})$, $k \in \mathbb{N}$. It follows that (54)–(56) have at least one real eigenvalue in each interval $(\nu_{k-1}^{\mathrm{D}}, \nu_k^{\mathrm{D}})$, $k \in \mathbb{N}$ (see Figure 2).

Further information concerning the distribution of the eigenvalues ν_n is obtained by comparing (57) with the corresponding formula

$$\nu_n^{\mathrm{D}} = (n+1)^2\pi^2 + \int_0^1 Q(y)\,\mathrm{d}y + o\left(\frac{1}{n}\right), \qquad n \in \mathbb{N}_0$$

for the Dirichlet eigenvalues; in particular one finds that

$$\nu_{n-1}^{\mathrm{D}} < \operatorname{Re}\nu_{n+1} < \nu_n^{\mathrm{D}}$$

for sufficiently large n. It follows that for sufficiently large n the eigenvalue ν_{n+1} is real and located in the interval $(\nu_{n-1}^{\mathrm{D}}, \nu_n^{\mathrm{D}})$. Using this observation and the above geometrical characterization of real eigenvalues, one concludes that

(i) each interval $(\nu_{n-1}^{\mathrm{D}}, \nu_n^{\mathrm{D}})$, $n \in \mathbb{N}$ contains a simple real eigenvalue;

(ii) there are precisely two additional eigenvalues (counted according to algebraic multiplicity) which are either

  (a) real and located to the left of ν_0^{D} (Figure 2 (top left)),

  (b) real and located in the interval $(\nu_{n-1}^{\mathrm{D}}, \nu_n^{\mathrm{D}})$ for some $n \in \mathbb{N}$ (Figure 2 (top right)),

(c) a complex-conjugate pair (with a nonvanishing imaginary part) whose real part is smaller than ν_0^{D} (Figure 2 (bottom)),

(d) a complex-conjugate pair (with a nonvanishing imaginary part) whose real part lies in the interval $(\nu_{n-1}^{\mathrm{D}}, \nu_n^{\mathrm{D}})$ for some $n \in \mathbb{N}$ (Figure 2 (bottom)).

The eigenvalues κ of L are recovered from the above analysis by the formula $\nu = \kappa^2$, so that in particular they occur in plus-minus pairs. Clearly L has a real eigenvalue in each interval $((\nu_{n-1}^{\mathrm{D}})^{1/2}, (\nu_n^{\mathrm{D}})^{1/2})$ and $(-(\nu_n^{\mathrm{D}})^{1/2}, -(\nu_{n-1}^{\mathrm{D}})^{1/2})$, $n \in \mathbb{N}$ (see point (i) above), and it follows from point (ii) that there are four additional eigenvalues (counted according to algebraic multiplicity). The four additional eigenvalues could be real with magnitude greater than $(\nu_0^{\mathrm{D}})^{1/2}$ (case (b)) or a plus-minus, complex-conjugate quartet with nonvanishing real and imaginary parts (cases (c) and (d)). Observe that only the remaining case (a) can lead to purely imaginary eigenvalues of L, and we now examine this case in detail.

Case (a) has eight subcases, each of which is illustrated in Figure 3. Clearly a positive eigenvalue ν yields a plus-minus pair of real eigenvalues of L, while a negative eigenvalue ν yields a complex-conjugate pair of purely imaginary eigenvalues of L; an algebraically double positive or negative eigenvalue ν yields a pair of algebraically double real or purely imaginary eigenvalues. A zero eigenvalue of (54)–(56) similarly leads to a zero eigenvalue of L, the algebraic multiplicity of which is readily determined by studying the equation $Lu = 0$ directly. A straightforward calculation shows that zero is an eigenvalue of L if and only if $\alpha = \alpha_\star$; the eigenvalue has algebraic multiplicity 2 when $\beta \neq \beta_\star$ and algebraic multiplicity 4 when $\beta = \beta_\star$, where

$$\alpha_\star = \left(\int_0^1 a^{-3}(s)\,\mathrm{d}s\right)^{-1}, \qquad \beta_\star = \alpha_\star^2 \int_0^1 a(s)\left(\int_0^s a^{-3}(t)\,\mathrm{d}t\right)^2 \mathrm{d}s.$$

The generalized eigenvectors w_j, where $Lw_1 = 0$, $Lw_2 = w_1$, and $Lw_j = w_{j-1}$, $j = 3, 4$, for $\beta = \beta_\star$ are given by

$$w_1 = \begin{pmatrix} \int_0^1 a^{-3}(s)\,\mathrm{d}s \\ 0 \\ \int_0^s a^{-3}(t)\,\mathrm{d}t \\ 0 \end{pmatrix}, \qquad w_2 = \begin{pmatrix} 0 \\ -a(1)\int_0^1 a^{-3}(s)\,\mathrm{d}s \\ 0 \\ -a(s)\int_0^s a^{-3}(t)\,\mathrm{d}t \end{pmatrix}, \tag{58}$$

$$w_3 = \begin{pmatrix} -\int_0^1 a^{-3}(s)\int_0^s a(t)\int_0^t a^{-3}(u)\,\mathrm{d}u\,\mathrm{d}t\,\mathrm{d}s \\ 0 \\ -\int_0^s a^{-3}(t)\int_0^t a(u)\int_0^u a^{-3}(v)\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}t \\ 0 \end{pmatrix}, \tag{59}$$

$$w_4 = \begin{pmatrix} 0 \\ a(1)\int_0^1 a^{-3}(s)\int_0^s a(t)\int_0^t a^{-3}(u)\,\mathrm{d}u\,\mathrm{d}t\,\mathrm{d}s \\ 0 \\ a(s)\int_0^s a^{-3}(t)\int_0^t a(u)\int_0^u a^{-3}(v)\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}t \end{pmatrix}. \tag{60}$$

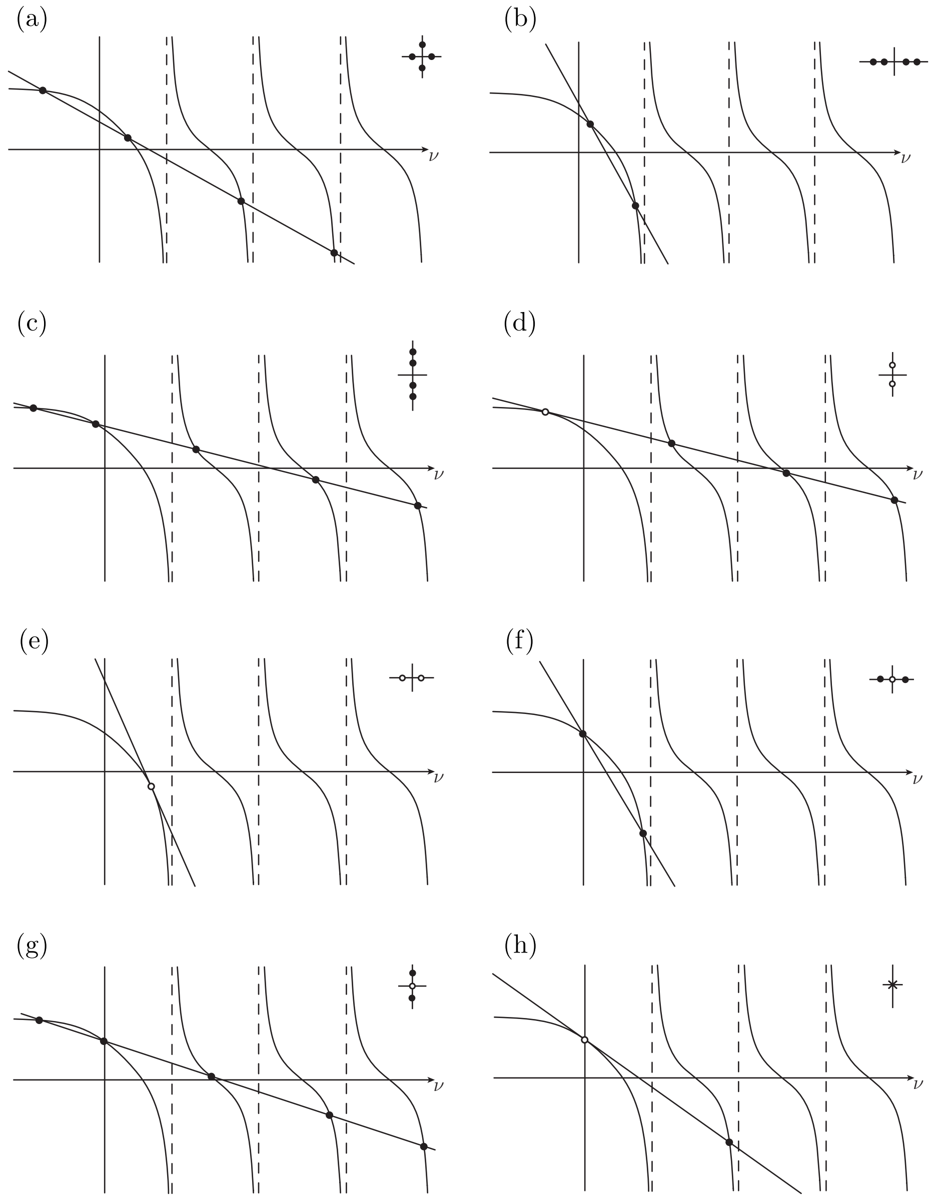


FIG. 3. *The eight cases in which the Sturm–Liouville problem* (54)–(56) *has two real eigenvalues (counted according to algebraic multiplicity) which are less than ν_0^{D}; solid and hollow dots denote, respectively, algebraically simple and algebraically double eigenvalues. The insets show the corresponding eigenvalues of the linear operator L; the cross denotes an eigenvalue of algebraic multiplicity* 4.

Recall that L has four eigenvalues (counted according to algebraic multiplicity) with a real part in the interval $(-(\nu_0^{\mathrm{D}})^{1/2}, (\nu_0^{\mathrm{D}})^{1/2})$. An algebraically simple zero eigenvalue of (54)–(56) (Figure 3(f) or (g)) therefore corresponds to the case $\alpha = \alpha_\star$, $\beta \neq \beta_\star$, while an algebraically double zero eigenvalue (Figure 3(h)) corresponds to the case $\alpha = \alpha_\star$, $\beta = \beta_\star$; it follows that

$$B(0) = a'(1) + a^{-2}(1)\alpha_\star, \qquad B'(0) = -a^{-2}(1)\beta_\star.$$

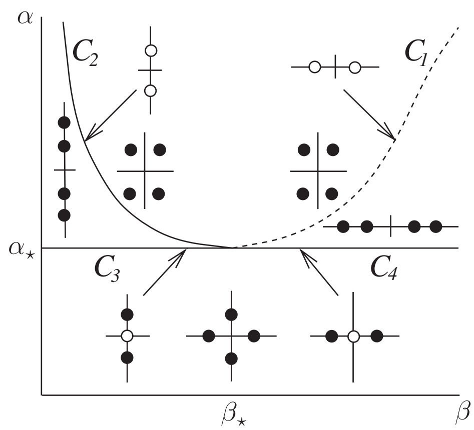


FIG. 4. *Eigenvalues of L whose real parts lie in $(-(\nu_0^{\mathrm{D}})^{1/2}, (\nu_0^{\mathrm{D}})^{1/2})$; solid and hollow dots denote, respectively, algebraically simple and double eigenvalues. The curves C_j, $j = 1, \ldots, 4$, consist of points in (β, α) parameter space at which the qualitative nature of the eigenvalue picture changes; the real parts of the four complex eigenvalues just above $C_1 \cup C_2$ move out of the range $(-(\nu_0^{\mathrm{D}})^{1/2}, (\nu_0^{\mathrm{D}})^{1/2})$ as one moves away from $C_1 \cup C_2$.*

From this observation we conclude that Figure 3(f) corresponds to the case $\alpha = \alpha_\star$, $\beta > \beta_\star$, while Figure 3(g) corresponds to the case $\alpha = \alpha_\star$, $\beta < \beta_\star$; furthermore, Figure 3(a) corresponds to the case $\alpha < \alpha_\star$, Figures 3(b) and (e) arise when $\alpha > \alpha_\star$, $\beta > \beta_\star$, and Figures 3(c) and (d) arise when $\alpha > \alpha_\star$, $\beta < \beta_\star$.

Figure 4 summarizes the behavior of the four eigenvalues of L whose real parts lie in $(-(\nu_0^{\mathrm{D}})^{1/2}, (\nu_0^{\mathrm{D}})^{1/2})$. The values of β and α determine, respectively, the gradient of the straight line in Figure 3 and its point of intersection with the vertical axis: Increasing β makes it steeper, while increasing α moves it upwards. Suppose, for example, that $\alpha < \alpha_\star$ (Figure 3(a)), and we increase the value of α while keeping β fixed; according to whether $\beta < \beta_\star$ or $\beta > \beta_\star$ we pass through the sequence (a), (g), (c), (d) or (a), (f), (b), (e). Figure 4 follows from these observations. Of particular interest are the four bifurcation curves C_j, $j = 1, 4$, at points of which the qualitative nature of the eigenvalue picture changes. In contrast to $C_4 = \{(\beta, \alpha_\star) : \beta > \beta_\star\}$ and $C_3 = \{(\beta, \alpha_\star) : \beta < \beta_\star\}$, explicit parametrizations of C_2 and C_1 are available only in certain special cases, in particular for irrotational flows (Kirchgässner [29]).

Recall that L also has a countably infinite number of eigenvalues whose real parts lie outside the range $(-(\nu_0^{\mathrm{D}})^{1/2}, (\nu_0^{\mathrm{D}})^{1/2})$; these eigenvalues are all real and simple at points in the (β, α) parameter plane below the curve $C_1 \cup C_2$, while all but possibly four of the complete set of eigenvalues are real above $C_1 \cup C_2$. The region above $C_1 \cup C_2$ contains a further countably infinite family of bifurcation curves, points of the nth of which correspond to the transition between Figure 2 (top right) and Figure 2 (bottom): Four real eigenvalues with magnitudes in the interval $((\nu_{n-1}^{\mathrm{D}})^{1/2}, (\nu_n^{\mathrm{D}})^{1/2})$ become complex by colliding in pairs on the real axis. At each point in (β, α) parameter space above $C_1 \cup C_2$ there is therefore an infinite number of real, simple eigenvalues and possibly one pair of algebraically double real eigenvalues or one plus-minus, complex-conjugate quartet of complex eigenvalues.

The spectral theory presented above can also be formulated in terms of self-adjoint operators on Pontryagin spaces (see, e.g., Bognár [5] and Iohvidov, Krein, and Langer [22]). Introduce the π_1-space $P = L^2(0,1) \times \mathbb{R}$ with indefinite inner product

$$[(\phi_1, b_1), (\phi_2, b_2)] = \langle a\phi_1, \phi_2 \rangle_{L^2(0,1)} - \beta b_1 b_2$$

and the linear operator $T : \mathcal{D}(T) \subset P \to P$ defined by

$$T(\phi, b) = (-a^{-1}(s)(a^3(s)\phi_s)_s, -a^3(1)\beta^{-1}\phi_s(1) + \alpha\beta^{-1}\phi(1)),$$

with

$$\mathcal{D}(T) = \{(\phi, b) : \phi \in H^2(0,1),\ \phi(0) = 0,\ \phi(1) = b\}.$$

Observe that T is densely defined (with respect to the topology of $L^2(0,1) \times \mathbb{R}$) and symmetric (with respect to $[\cdot\,,\cdot]$), so that it is a self-adjoint operator on P, and it follows from (51)–(53) that κ is an eigenvalue of L if and only if κ^2 is an eigenvalue of T. This framework, which was used by Wahlén [35] in his study of periodic water waves, yields a convenient method of calculating the local parts of the curves C_1 and C_2 in Figure 4 near the point $(\beta_\star, \alpha_\star)$.

A point $(\beta_\kappa, \alpha_\kappa)$ of C_1 is characterized by the fact that κ^2 is a geometrically simple eigenvalue of $T_\kappa := T|_{(\beta,\alpha)=(\beta_\kappa,\alpha_\kappa)}$ with algebraic multiplicity 2, and since the algebraic multiplicity of this eigenvalue exceeds its geometric multiplicity the corresponding eigenspace E_{κ^2} is neutral; that is, $[\tilde{v}, \tilde{v}]_\kappa := [\tilde{v}, \tilde{v}]_{\beta=\beta_\kappa}$ vanishes for every $\tilde{v} = (v, v(1)) \in E_{\kappa^2}$ (see Wahlén [35, p. 12]). One obtains expressions for v_j, β_j, α_j by inserting the expansions

$$T_\kappa = T_0 + \kappa^2 T_2 + \kappa^4 T_4 + \cdots, \qquad [\cdot\,,\cdot]_\kappa = [\cdot\,,\cdot]_0 + \kappa^2[\cdot\,,\cdot]_2 + \kappa^4[\cdot\,,\cdot]_4 + \cdots$$

and

$$v_\kappa = v_0 + \kappa^2 v_2 + \kappa^4 v_4 + \cdots,$$

$$\beta_\kappa = \beta_\star + \kappa^2\beta_2 + \kappa^4\beta_4 + \cdots, \qquad \alpha_\kappa = \alpha_\star + \kappa^2\alpha_2 + \kappa^4\alpha_4 + \cdots$$

into

$$T_\kappa \tilde{v}_\kappa = \kappa^2 \tilde{v}_\kappa, \qquad [\tilde{v}_\kappa, \tilde{v}_\kappa]_\kappa = 0,$$

and equating coefficients of κ^2. In particular we find that

$$v_0 = \int_0^s a^{-3}(s)\,\mathrm{d}t, \qquad v_2 = -\int_0^s a^{-3}(t)\int_0^t a(u)\int_0^u a^{-3}(v)\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}t,$$

$$\beta_2 = \frac{2[\tilde{v}_2, \tilde{v}_0]_0}{v_0^2(1)}, \qquad \alpha_2 = 0, \quad \alpha_4 = \frac{[\tilde{v}_2, \tilde{v}_0]_0}{v_0^2(1)},$$

and hence that

$$\alpha_4 = \alpha_\star^2 d_1, \qquad \beta_2 = 2\alpha_\star^2 d_1, \tag{61}$$

where

$$\begin{aligned} d_1 = &-\alpha_\star \left(\int_0^1 a^{-3}(s)\int_0^s a(t)\int_0^t a^{-3}(u)\,\mathrm{d}u\,\mathrm{d}t\,\mathrm{d}s\right)^2 \\ &+ \int_0^1 a^{-3}(s)\left(\int_0^s a(t)\int_0^t a^{-3}(u)\,\mathrm{d}u\,\mathrm{d}t\right)^2\mathrm{d}s > 0. \end{aligned} \tag{62}$$

A similar calculation shows that the corresponding parametrization of the local part of C_2 near $(\beta_\star, \alpha_\star)$ is

$$\beta_\lambda = \beta_\star - 2\alpha_\star^2 d_1\lambda^2 + O(\lambda^4), \qquad \alpha_\lambda = \alpha_\star + \alpha_\star^2 d_1\lambda^4 + O(\lambda^6)$$

(a point $(\beta_\lambda, \alpha_\lambda)$ of C_2 is characterized by the fact that $-\lambda^2$ is a geometrically simple eigenvalue of $T_\lambda := T|_{(\beta,\alpha)=(\beta_\lambda,\alpha_\lambda)}$ with algebraic multiplicity 2).

4. The reduced Hamiltonian systems. Our existence theory for solitary waves is completed by showing that the reduced Hamiltonian system on the center manifold admits homoclinic solutions. Irrotational solitary waves have been found at points in (β, α) parameter space near C_1, C_2, and C_4 in this fashion, and in this section we examine the corresponding reduced systems in our more general context.

4.1. Homoclinic bifurcation at C_4. A *Hamiltonian* 0^2 *resonance* takes place at points of the curve C_4 in Figure 4: Two real eigenvalues become purely imaginary by colliding at the origin and forming a Jordan chain of length 2. This resonance is associated with the bifurcation of a branch of homoclinic solutions into the region with real eigenvalues (the parameter regime marked I in Figure 1(a)). Let us therefore fix reference values $(\beta_0, \alpha_0) \in C_4$, so that $\beta_0 > \beta_\star$, $\alpha_0 = \alpha_\star$, and introduce a bifurcation parameter by choosing $(\varepsilon_1, \varepsilon_2) = (0, \delta)$, where $0 < \delta \ll 1$.

Formulas for the generalized eigenvectors w_1, w_2, where $Lw_1 = 0$, $Lw_2 = w_1$, are given in (58), and one finds that

$$\Psi(w_1, w_2) = \alpha_\star^{-2}(\beta_0 - \beta_\star),$$

where

$$\begin{aligned}\Psi((\rho_1, \zeta_1, \phi_1, z_1), (\rho_2, \zeta_2, \phi_2, z_2)) &= \Phi^0|_0((\rho_1, \zeta_1, \phi_1, z_1), (\rho_2, \zeta_2, \phi_2, z_2)) \\ &= -\beta_0 a^{-1}(1)(\zeta_2\rho_1 - \zeta_1\rho_2) + \int_0^1 (z_2\phi_1 - z_1\phi_2)\,\mathrm{d}s.\end{aligned}$$

It follows that $\{e, f\}$, where

$$e = \alpha_\star(\beta_0 - \beta_\star)^{-1/2} w_1, \qquad f = \alpha_\star(\beta_0 - \beta_\star)^{-1/2} w_2,$$

is a symplectic basis for the central subspace $X_1 = P(X)$ of X defined by the spectral projection $P : X \to X$ corresponding to the purely imaginary part of L. The coordinates q, p in the e and f directions are canonical coordinates for X_1, and the action of the reverser S on this space is given by

$$S(q, p) = (q, -p).$$

Modeling the center manifold $\tilde{X}^\varepsilon$, $\varepsilon = (0, \delta)$, upon the single coordinate chart $\tilde{U}_1$ and choosing the coordinate map according to the recipe given in the paragraph below Theorem 3.1, we can identify $(\tilde{X}^\varepsilon, \tilde{\Phi}^\varepsilon, \tilde{H}^\varepsilon)$ with the two-dimensional canonical Hamiltonian system $(\mathcal{M}, \Upsilon, \tilde{H}^\varepsilon)$, where $\mathcal{M}$ is a neighborhood of the origin in $\mathbb{R}^2$,

$$\Upsilon((q^1, p^1), (q^2, p^2)) = q^1 p^2 - p^1 q^2$$

and

$$\tilde{H}^\varepsilon(q, p) = K^\varepsilon(\tilde{u}_1 + \tilde{r}(\tilde{u}_1; \varepsilon)), \qquad \tilde{u}_1 = qe + pf.$$

A direct calculation shows that

$$\tilde{H}_2^{0,0}(q, p) = K_2^{0,0}[\tilde{u}_1, \tilde{u}_1] = \frac{1}{2}p^2,$$

where $\varepsilon_1^i \varepsilon_2^j \tilde{H}_k^{i,j}(\tilde{u}_1)$ denotes the part of the Taylor expansion of $\tilde{H}^\varepsilon(\tilde{u}_1)$ which is homogeneous of order i in ε_1, j in ε_2, and k in $\tilde{u}_1 \cong (q, p)$ and $K_k^{i,j}$ denotes the symmetric,

k-linear operator $X_1^k \to \mathbb{R}$ which defines the corresponding coefficient in the Taylor expansion of K^ε. Anticipating the scaling $q \sim \delta Q$, $p \sim \delta^{3/2} P$, we write

$$\tilde{H}^\varepsilon(q,p) = \frac{1}{2}p^2 + c_1\delta q^2 + c_2 q^3 + O(|p||(q,p)||(\delta,q,p)|) + O(|(q,p)|^2|(\delta,q,p)|^2),$$

so that the first three terms on the right-hand side of the above equation are $O(\delta^3)$ and the remainder is of higher order. The coefficients c_1 and c_2 are obtained from the calculations

$$\begin{aligned} c_1 &= K_2^{0,1}[e,e] + 2K_2^{0,0}[e,\tilde{r}_{10}^{0,1}] \\ &= K_2^{0,1}[e,e] + \Psi(Le,\tilde{r}_{10}^{0,1}) \\ &= K_2^{0,1}[e,e] \\ &= -\frac{1}{2}(\beta_0 - \beta_\star)^{-1} \end{aligned}$$

and

$$c_2 = K_3^{0,0}[e,e,e] + 2K_2^{0,0}[e,\tilde{r}_{20}^{0,0}] = -\frac{1}{2}c_0(\beta_0-\beta_\star)^{-3/2}, \qquad c_0 = \alpha_\star^3 \int_0^1 a^{-5}(s)\,\mathrm{d}s,$$

in which $\tilde{r}_{k\ell}^{i,j}$ denotes the coefficient of $\varepsilon_1^i \varepsilon_2^j q^k p^\ell$ in the Taylor expansion of $\tilde{r}$ and we have made use of the identity

$$\Psi(L\tilde{u}_1^1, \tilde{u}_1^2) = 2K_2^{0,0}[\tilde{u}_1^1, \tilde{u}_2^2]. \tag{63}$$

Hamilton's equations for $(\mathcal{M}, \Upsilon, \tilde{H}^\varepsilon)$ are

$$q_x = p + \mathcal{R}_1(q,p,\delta), \tag{64}$$

$$p_x = \delta(\beta_0-\beta_\star)^{-1}q + \frac{3}{2}c_0(\beta_0-\beta_\star)^{-3/2}q^2 + \mathcal{R}_2(q,p,\delta), \tag{65}$$

where $\mathcal{R}_1$, $\mathcal{R}_2$ are, respectively, odd and even in their second arguments and

$$\mathcal{R}_1 = O(|p||(\delta,q,p)|) + O(|(q,p)||(\delta,q,p)|^2), \qquad \mathcal{R}_2 = O(|(q,p)||(\delta,q,p)|).$$

Introducing the scaled variables

$$X = \delta^{1/2}(\beta_0-\beta_\star)^{-1/2}x, \quad q(x) = c_0^{-1}\delta(\beta_0-\beta_\star)^{1/2}Q(X), \quad p(x) = c_0^{-1}\delta^{3/2}P(X),$$

one finds from (64), (65) that

$$Q_X = P + \mathcal{R}_3(Q,P,\delta), \tag{66}$$

$$P_X = Q + \frac{3}{2}Q^2 + \mathcal{R}_4(Q,P,\delta), \tag{67}$$

where the remainder terms $\mathcal{R}_3$ and $\mathcal{R}_4$ are $O(\delta^{1/2})$ and, respectively, odd and even in their second arguments. In the limit $\delta \to 0$, (66), (67) are equivalent to

$$\begin{aligned} Q_X &= P, \\ P_X &= Q + \frac{3}{2}Q^2, \end{aligned}$$

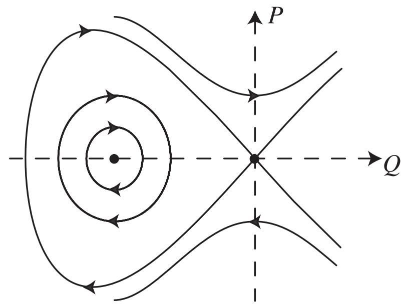


FIG. 5. *Phase portrait of the scaled reduced system of equations.*

whose phase portrait is easily calculated by elementary methods and is depicted in Figure 5. Notice in particular that it has a nonzero equilibrium $(-2/3, 0)$, surrounded by the symmetric homoclinic orbit

$$Q(X) = -\mathrm{sech}^2\left(\frac{X}{2}\right), \qquad P(X) = \mathrm{sech}^2\left(\frac{X}{2}\right)\tanh\left(\frac{X}{2}\right).$$

One can exploit the reversibility of (66), (67) to deduce that it has a symmetric homoclinic orbit for small positive values of δ. For $\delta = 0$ the stable manifold W_{s}^0 of the zero equilibrium is known explicitly (it consists of the points on the homoclinic orbit), and since $TW_{\mathrm{s}}^0|_{(-1,0)} = \{Q = 0\}$ it intersects the symmetric section Fix S $= \{P = 0\}$ transversally in the point $(-1, 0)$. The stable manifold theorem states that W_{s}^δ depends uniformly smoothly upon δ, and because the symmetric section is independent of δ it follows that W_{s}^δ and Fix S intersect transversally in a point near $(-1, 0)$ for sufficiently small positive values of δ. One concludes that the phase portrait of (66), (67) has a reversible homoclinic orbit in the left half-plane for sufficiently small positive values of δ.

Tracing back the various changes of variable, one finds that the surface profile of the water corresponding to the homoclinic orbit detected above is given by

$$\rho(x) = -c_0^{-1}\delta\,\mathrm{sech}^2\left(\frac{\delta^{1/2}x}{2(\beta_0 - \beta_\star)^{1/2}}\right) + O(\delta^{3/2}).$$

We therefore obtain a symmetric solitary wave of depression which decays exponentially and monotonically to a horizontal laminar flow as $x \to \pm\infty$ and is sketched in Figure 1(b).

4.2. Homoclinic bifurcation at C_1. A *Hamiltonian real* $1:1$ *resonance* occurs at points of the curve C_1 in Figure 4: Two pairs of real eigenvalues become complex by colliding at nonzero points on the real axis and forming two Jordan chains of length 2. Of particular interest here is the local part of C_1 near the point $(\beta_\star, \alpha_\star)$, since we can access this curve using the center-manifold reduction technique with reference value $(\beta_0, \alpha_0) = (\beta_\star, \alpha_\star)$. We choose values of the bifurcation parameter $\varepsilon = (\varepsilon_1, \varepsilon_2)$ in a fashion which enables us to access this curve effectively, namely, by writing

$$\varepsilon_1 = \beta_2(1+\delta)\mu^2, \qquad \varepsilon_2 = \alpha_4\mu^4, \tag{68}$$

where β_2, α_4 are the first nonvanishing coefficients in the parametrization

$$\beta_\kappa = \beta_\star + \beta_2\kappa^2 + O(\kappa^4), \qquad \alpha_\kappa = \alpha_\star + \alpha_4\kappa^4 + O(\kappa^6)$$

of the local part of C_1 near $(\beta_\star, \alpha_\star)$; explicit formulas for β_2 and α_4 are given in (61). Notice that μ indicates the distance from the point $(\beta_\star, \alpha_\star)$, while δ plays the role of a bifurcation parameter (varying δ through zero from above we cross the critical curve C_1 in parameter space from above); the parameter regime marked II in Figure 1(a) corresponds to small, positive values of δ and μ.

The point $(\beta_0, \alpha_0) = (\beta_\star, \alpha_\star)$ in parameter space is associated with a *Hamiltonian* 0^4 *resonance* (L has a zero eigenvalue with a Jordan chain of length 4); formulas for the generalized eigenvectors w_j, $j = 1, \ldots 4$, where $Lw_1 = 0$ and $Lw_j = w_{j-1}$, $j = 2, 3, 4$, are given in (58)–(60). One finds that

$$\Psi(w_1, w_4) = -d_1, \quad \Psi(w_2, w_3) = d_1, \quad \Psi(w_3, w_4) = d_2,$$

and the symplectic product of any other combination of these vectors is zero; here $d_1 > 0$ is given by (62) and

$$d_2 = \beta_\star \left(\int_0^1 a^{-3}(s) \int_0^s a(t) \int_0^t a^{-3}(u)\, \mathrm{d}u\, \mathrm{d}t\, \mathrm{d}s \right)^2 - \int_0^1 a(s) \left(\int_0^s a^{-3}(t) \int_0^t a(u) \int_0^u a^{-3}(v)\, \mathrm{d}v\, \mathrm{d}u\, \mathrm{d}t \right)^2 \mathrm{d}s.$$

It follows that $\{e_1, e_2, f_1, f_2\}$, where

$$e_1 = d_1^{-1/2}(w_4 + d_3 w_2), \quad e_2 = d_1^{-1/2} w_2, \qquad f_1 = d_1^{-1/2} w_1, \quad f_2 = d_1^{-1/2} w_3,$$

and $d_3 = d_2/d_1$, is a symplectic basis for the central subspace X_1. The coordinates q_1, q_2, p_1, and p_2 in the e_1, e_2, f_1, and f_2 directions are canonical coordinates for X_1, and the action of the reverser S on this space is given by

$$S(q_1, q_2, p_1, p_2) = (-q_1, -q_2, p_1, p_2).$$

Modeling the center manifold $\tilde{X}^\varepsilon$ upon the single coordinate chart $\tilde{U}_1$ and choosing the coordinate map according to the recipe given in the paragraph below Theorem 3.1, we can identify $(\tilde{X}^\varepsilon, \tilde{\Phi}^\varepsilon, \tilde{H}^\varepsilon)$ with the four-dimensional canonical Hamiltonian system $(\mathcal{M}, \Upsilon, \tilde{H}^\varepsilon)$, where $\mathcal{M}$ is a neighborhood of the origin in $\mathbb{R}^4$,

$$\Upsilon((q_1^1, q_2^1, p_1^1, p_1^1), (q_2^2, q_2^2, p_2^2, p_2^2)) = q_1^1 p_1^2 + q_2^1 p_2^2 - p_1^1 q_1^2 - p_2^1 q_2^2$$

and

$$\tilde{H}^\varepsilon(q_1, q_2, p_1, p_2) = K^\varepsilon(\tilde{u}_1 + \tilde{r}(\tilde{u}_1; \varepsilon)), \qquad \tilde{u}_1 = q_1 e_1 + q_2 e_2 + p_1 f_1 + p_2 f_2.$$

The quadratic part of the reduced Hamiltonian is readily computed; one finds that

$$\tilde{H}_2^{0,0}(q_1, q_2, p_1, p_2) = K_2^{0,0}[\tilde{u}_1, \tilde{u}_1] = -\frac{d_3}{2} q_1^2 - q_1 q_2 + \frac{1}{2} p_2^2,$$

and anticipating the scaling $q_1 \sim \mu^7 Q_1$, $q_2 \sim \mu^5 Q_2$, $p_1 \sim \mu^4 P_1$, $p_2 \sim \mu^6 P_2$ and the parametrization (68), we write

$$\begin{aligned} \tilde{H}^\varepsilon(q_1, q_2, p_1, p_2) &= \frac{1}{2} p_2^2 - q_1 q_2 \\ &+ c_1^{1,0} \varepsilon_1 p_1^2 + c_2^{1,0} \varepsilon_1 p_1 p_2 + c_6^{1,0} \varepsilon_1 q_2^2 + c_1^{0,1} \varepsilon_2 p_1^2 + c_1^{2,0} \varepsilon_1^2 p_1^2 + c p_1^3 + \mathcal{R}(q_1, q_2, p_1, p_2), \end{aligned} \tag{69}$$

so that the third term on the right-hand side of the above equation is $O(\mu^{10})$, the remainder term $\mathcal{R}$ is $O(\mu^{14})$ (note that $\tilde{H}^\varepsilon(-q_1,-q_2,p_1,p_2) = \tilde{H}^\varepsilon(q_1,q_2,p_1,p_2)$ because of the reversibility), and all other terms are $O(\mu^{12})$.

PROPOSITION 4.1. *The Darboux transformation used to construct* $(\mathcal{M},\Upsilon,\tilde{H}^\varepsilon)$ *(see* (36)*) can be chosen so that* $c_2^{1,0}=0$.

Proof. Observe that the change of variable

$$\bar{q}_1 = q_1 - c_2^{1,0}\varepsilon_1 q_2, \quad \bar{q}_2 = q_2, \quad \bar{p}_1 = p_1, \quad \bar{p}_2 = p_2 + c_2^{1,0}\varepsilon_1 p_1 \tag{70}$$

is symplectic and transforms $(\mathcal{M},\Upsilon,\tilde{H}^\varepsilon)$ into $(\mathcal{M},\Upsilon,\bar{H}^\varepsilon)$, where

$$\begin{aligned}\bar{H}^\varepsilon(\bar{q}_1,\bar{q}_2,\bar{p}_1,\bar{p}_2) &= \frac{1}{2}\bar{p}_2^2 - \bar{q}_1\bar{q}_2\\ &\quad + c_1^{1,0}\varepsilon_1\bar{p}_1^2 + \bar{c}_6^{1,0}\varepsilon_1\bar{q}_2^2 + c_1^{0,1}\varepsilon_2\bar{p}_1^2 + \bar{c}_1^{2,0}\varepsilon_1^2\bar{p}_1^2 + c\bar{p}_1^3 + \bar{\mathcal{R}}(\bar{q}_1,\bar{q}_2,\bar{p}_1,\bar{p}_2)\end{aligned}$$

and $\bar{c}_6^{1,0} = c_6^{1,0} - c_2^{1,0}$, $\bar{c}_1^{2,0} = c_1^{2,0} - \frac{1}{2}(c_2^{1,0})^2$ (the remainder term $\bar{\mathcal{R}}$ is $O(\mu^{14})$ in the sense explained above). The result follows by replacing the Darboux transformation used in the construction by its composition with the change of variable (70). ◻

To calculate the remaining coefficients on the right-hand side of (69), we exploit the identity

$$\begin{aligned}&L\tilde{r}(\tilde{u}_1;\varepsilon) - \mathrm{d}_1\tilde{r}[\tilde{u}_1;\varepsilon](L\tilde{u}_1)\\ &\qquad = -N^\varepsilon(\tilde{u}_1 + \tilde{r}(\tilde{u}_1;\varepsilon)) + \mathrm{d}_1\tilde{r}[\tilde{u}_1;\varepsilon](P^\varepsilon(\tilde{u}_1)) + P^\varepsilon(\tilde{u}_1),\end{aligned} \tag{71}$$

in which $P^\varepsilon(\tilde{u}_1)$ is the nonlinear part of the reduced Hamiltonian vector field $v_{\tilde{H}^\varepsilon}$; this identity is derived by substituting $u = \tilde{u}_1 + \tilde{r}(\tilde{u}_1;\varepsilon)$ and $\tilde{u}_{1x} = L\tilde{u}_1 + P^\varepsilon(\tilde{u}_1)$ into (41). Let us write

$$\tilde{H}_2^{1,0}(q_1,q_2,p_1,p_2) = c_1^{1,0}p_1^2 + c_3^{1,0}p_2^2 + c_4^{1,0}q_1^2 + c_5^{1,0}q_1q_2 + c_6^{1,0}q_2^2$$

(the coefficients of the remaining terms vanish because of the reversibility), so that

$$P_1^{1,0}(\tilde{u}_1) = 2c_1^{1,0}p_1e_1 + 2c_3^{1,0}p_2e_2 - (2c_4^{1,0}q_1 + c_5^{1,0}q_2)f_1 - (2c_6^{1,0}q_2 + c_5^{1,0}q_1)f_2,$$

and equating coefficients of $\varepsilon_1\tilde{u}_1$ on both sides of (71), we find that

$$L\tilde{r}_{1000}^{1,0} - \tilde{r}_{0001}^{1,0} - d_3\tilde{r}_{0010}^{1,0} = -N_1^{1,0}[e_1] - 2c_4^{1,0}f_1 - c_5^{1,0}f_2, \tag{72}$$

$$L\tilde{r}_{0100}^{1,0} - \tilde{r}_{0010}^{1,0} = -N_1^{1,0}[e_2] - c_5^{1,0}f_1 - 2c_6^{1,0}f_2, \tag{73}$$

$$L\tilde{r}_{0010}^{1,0} = -N_1^{1,0}[f_1] + 2c_1^{1,0}e_1, \tag{74}$$

$$L\tilde{r}_{0001}^{1,0} - \tilde{r}_{0100}^{1,0} = -N_1^{1,0}[f_2] + 2c_3^{1,0}e_2. \tag{75}$$

Here the notation $N_k^{i,j}$ and $\tilde{r}_k^{i,j}$, $P_k^{i,j}$ is defined analogously to the notation $K_k^{i,j}$ and $\tilde{H}_k^{i,j}$, while $\tilde{r}_{k_1k_2k_3k_4}^{i,j}$ is the coefficient of $\varepsilon_1^i\varepsilon_2^jq_1^{k_1}q_2^{k_2}p_1^{k_3}p_2^{k_4}$ in the Taylor expansion of $\tilde{r}$. Explicit calculations show that $N_1^{1,0}(e_1)$, $N_1^{1,0}(e_2)$, $N_1^{1,0}(f_1)$ vanish while

$$N_1^{1,0}(f_2) = \left(0, d_1^{-1/2}\beta_\star^{-1}\alpha_\star^{-1}a(1), 0, 0\right).$$

It follows by elementary linear algebra that the system of equations (72)–(75) is solvable if and only if

$$c_1^{1,0} = 0, \qquad c_6^{1,0} = -\frac{1}{2d_1\alpha_\star^2},$$

and $\tilde{r}_{0010}^{1,0} = \gamma f_1$ for some $\gamma \in \mathbb{R}$.

The value of the coefficient $c_1^{0,1}$ is obtained from the calculation

$$\begin{aligned} c_1^{0,1} &= K_2^{0,1}[f_1,f_1] + 2K_2^{0,0}[f_1,\tilde{r}_{0010}^{0,1}] \\ &= K_2^{0,1}[f_1,f_1] + \Psi(Lf_1,\tilde{r}_{0010}^{0,1}) \\ &= K_2^{0,1}[f_1,f_1] \\ &= -\frac{1}{2d_1\alpha_\star^2}. \end{aligned}$$

Equating coefficients of $\varepsilon^2 p_1$ on both sides of (71), we find that

$$L\tilde{r}_{0010}^{2,0} = -N_1^{2,0}[f_1] - N_1^{1,0}[\tilde{r}_{0010}^{1,0}] + 2c_1^{2,0}e_1 + c_2^{2,0}e_2, \tag{76}$$

where the notation for the coefficients in $\tilde{H}_1^{0,1}$ is analogous to that used for $\tilde{H}_1^{1,0}$. An explicit calculation shows that

$$N_1^{2,0}[f_1] = 0, \qquad N_1^{1,0}[\tilde{r}_{0010}^{1,0}] = \gamma_1 N_1^{1,0}[f_1] = 0,$$

and it follows from (76) that

$$c_1^{2,0} = \frac{1}{2}\Psi(L\tilde{r}_{0010}^{2,0}, f_1) = -\frac{1}{2}\Psi(\tilde{r}_{0010}^{2,0}, Lf_1) = 0$$

(the "skew orthogonality" of L with respect to Ψ follows from (63)). Similarly, one finds from the p_1^3 component of (71), namely,

$$L\tilde{r}_{0020}^{0,0} = -N_2^{0,0}[f_1,f_1] + 3ce_1,$$

that

$$3c - \Psi(N_2^{0,0}[f_1,f_1],f_1) = \Psi(L\tilde{r}_{2000}^{0,0}, f_1) = -\Psi(\tilde{r}_{2000}^{0,0}, Lf_1) = 0$$

and hence that

$$c = \frac{1}{3}\Psi(N_2^{0,0}[f_1,f_1],f_1) = -\frac{1}{2d_1^{3/2}}\int_0^1 a^{-5}(s)\,\mathrm{d}s.$$

Observe that

$$c_6^{1,0} = -\frac{1}{\beta_2}, \qquad c_1^{0,1} = -\frac{1}{2\alpha_4}.$$

Choosing ε_1, ε_2 according to (68) and introducing the scaled variables $X = \mu x$ and

$$q_1(x) = \mu^7 Q_1(X), \quad q_2(x) = \mu^5 Q_2(X), \quad p_1(x) = \mu^4 P_1(X), \quad p_2(x) = \mu^6 P_2(X),$$

we therefore find that

$$\tilde{H}^\varepsilon(q_1,q_2,p_1,p_2) = \mu^{12}\left[-\frac{1}{2}P_1^2 - (1+\delta)Q_2^2 + \frac{1}{2}P_2^2 - Q_1Q_2 + cP_1^3\right] + O(\mu^{14})$$

and that Hamilton's equations for $(\mathcal{M},\Upsilon,\tilde{H}^\varepsilon)$ are

$$Q_{1X} = -P_1 + 3cP_1^2 + O(\mu), \tag{77}$$

$$Q_{2X} = P_2 + O(\mu), \tag{78}$$

$$P_{1X} = Q_2 + O(\mu), \tag{79}$$

$$P_{2X} = 2(1+\delta)Q_2 + Q_1 + O(\mu). \tag{80}$$

In the limit $\mu \to 0$ this dynamical system is equivalent to the single fourth-order ordinary differential equation

$$\partial_X^4 u - 2(1+\delta)\partial_X^2 u + u - u^2 = 0 \tag{81}$$

for the variable $u = 3c\,P_1$.

It was shown by Buffoni, Champneys, and Toland [6, section 2] that for $\delta = 0$ (81), and hence the system (77)–(80), has a homoclinic solution which corresponds to a transverse intersection (relative to the zero energy surface) of the stable and unstable manifolds of the zero equilibrium. Since transversality is an open condition, it follows that the same is true of the system (77)–(80) for sufficiently small positive values of δ and μ, for which (77)–(80) is a four-dimensional Hamiltonian system whose linearization has a plus-minus, complex-conjugate quartet of complex eigenvalues. The work of Devaney [12] therefore implies that there is a Smale horseshoe in the dynamics within the zero energy surface, and implicit in this construction is the existence of infinitely many homoclinic orbits which pass several times through a neighborhood of the "primary" transverse homoclinic orbit. These "multipulse" homoclinic orbits resemble multiple copies of the primary homoclinic orbit, between which there are distributed smaller local maxima and minima, and have an exponentially decaying oscillatory tail at infinity.

Tracing back the various changes of variable, one finds that the surface profile of the water is given by

$$\rho(x) = d_1^{-1/2}\mu^4 P_1(\mu x) + O(\mu^5).$$

The primary homoclinic orbit $u(x)$ of (81) satisfies $\max_{x\in\mathbb{R}} u(x) > 0$, so that the homoclinic orbits of (77)–(80) correspond to multitroughed solitary waves of depression which decay exponentially to a horizontal laminar flow as $x \to \pm\infty$ (see Figure 1(c)).

4.3. Homoclinic bifurcation at C_2. A *Hamiltonian Hopf bifurcation* takes place at points of the curve C_2 in Figure 4: Two pairs of purely imaginary eigenvalues become complex by colliding at nonzero points $\pm \mathrm{i}q$ on the imaginary axis and forming two Jordan chains of length 2. This resonance is associated with the bifurcation of a branch of homoclinic solutions into the region with complex eigenvalues (the parameter regime marked III in Figure 1(a)). Let us therefore fix reference values $(\beta_0, \alpha_0) \in C_2$ and introduce a bifurcation parameter by choosing $(\varepsilon_1, \varepsilon_2) = (0, \delta)$, where $0 < \delta \ll 1$.

By normalizing e, f and modifying f by the addition of a suitable multiple of e if necessary, we may suppose that the generalized eigenvectors e, f, where

$$Le = \mathrm{i}qe, \quad L\bar{e} = -\mathrm{i}q\bar{e}, \qquad (L - \mathrm{i}qI)f = e, \quad (L + \mathrm{i}qI)\bar{f} = \bar{e}, \tag{82}$$

satisfy $Se = \bar{e}$, $Sf = -\bar{f}$ and $\Psi(e, \bar{f}) = 1$, $\Psi(f, \bar{e}) = -1$, and the symplectic products of all other combinations are zero (note that Ψ acts bilinearly on pairs of complex vectors). It follows that $\{e, f, \bar{e}, \bar{f}\}$ is a symplectic basis for the central subspace X_1 (so that the coordinates A, B, $\bar{A}$, and $\bar{B}$ in the e, f, $\bar{e}$, and $\bar{f}$ directions are canonical coordinates for X_1), and the action of the reverser S on this space is given by

$$S(A, B) = (\bar{A}, -\bar{B}).$$

Modeling the center manifold $\tilde{X}^\varepsilon$, $\varepsilon = (0,\delta)$, upon the single coordinate chart $\tilde{U}_1$ and choosing the coordinate map according to the recipe given in the paragraph below Theorem 3.1, we can identify $(\tilde{X}^\varepsilon, \tilde{\Phi}^\varepsilon, \tilde{H}^\varepsilon)$ with the four-dimensional canonical Hamiltonian system $(\mathcal{M}, \Upsilon, \tilde{H}^\varepsilon)$, where $\mathcal{M}$ is a neighborhood of the origin in $\mathbb{R}^4$,

$$\Upsilon((A^1, B^1, \overline{A^1}, \overline{B^1}), (A^2, B^2, \overline{A^2}, \overline{B^2})) = A^1\overline{B^2} - A^2\overline{B^1} + \overline{A^1}B^2 - \overline{A^2}B^1$$

and

$$\tilde{H}^\varepsilon(A,B) = K^\varepsilon(\tilde{u}_1 + \tilde{r}(\tilde{u}_1;\varepsilon)), \qquad \tilde{u}_1 = Ae + Bf + \bar{A}\bar{e} + \bar{B}\bar{f}.$$

The flow of the above four-dimensional Hamiltonian system can be analyzed using the theory developed by Iooss and Pérouème [26] and Buffoni and Groves [7]. The Birkhoff normal-form theory states that for each $n_0 \geq 2$ there is a near-identity, analytic, symplectic change of coordinates with the property that

$$\begin{aligned}\tilde{H}^\varepsilon(A,B) = {} & \mathrm{i}q(A\bar{B} - \bar{A}B) + |B|^2 \\ & + H_{\mathrm{NF}}(|A|^2, \mathrm{i}(A\bar{B} - \bar{A}B), \delta) + O(|(A,B)|^2|(\delta, A, B)|^{n_0})\end{aligned}$$

in the new coordinates; the function H_{NF} is a real polynomial of order n_0+1 which satisfies

$$H_{\mathrm{NF}}(|A|^2, \mathrm{i}(A\bar{B} - \bar{A}B), \delta) = O(|(A,B)|^2|(\delta, A, B)|),$$

and in these coordinates Hamilton's equations for the reduced system are given by

$$A_x = \mathrm{i}qA + B + \mathrm{i}A\partial_2 H_{\mathrm{NF}}(|A|^2, \mathrm{i}(A\bar{B} - \bar{A}B), \delta) + O(|(A,B)||(\delta, A, B)|^{n_0}), \tag{83}$$

$$\begin{aligned}B_x = {} & \mathrm{i}qB + \mathrm{i}B\partial_2 H_{\mathrm{NF}}(|A|^2, \mathrm{i}(A\bar{B} - \bar{A}B), \delta) \\ & - A\partial_1 H_{\mathrm{NF}}(|A|^2, \mathrm{i}(A\bar{B} - \bar{A}B), \delta) + O(|(A,B)||(\delta, A, B)|^{n_0}).\end{aligned} \tag{84}$$

The theory by Iooss and Pérouème and Buffoni and Groves demands that the coefficients c_1 and c_3 in the expansion

$$\begin{aligned}H_{\mathrm{NF}} = {} & \delta c_1|A|^2 + \delta \mathrm{i}c_2(A\bar{B} - \bar{A}B) + c_3|A|^4 \\ & + \mathrm{i}c_4|A|^2(A\bar{B} - \bar{A}B) - c_5(A\bar{B} - \bar{A}B)^2 + \delta^2 c_6|A|^2 + \delta^2 \mathrm{i}c_7(A\bar{B} - \bar{A}B) + \cdots\end{aligned}$$

are, respectively, negative and positive; the methods explained in section 4.2 show that

$$c_1 = \Psi(N_1^{0,1}[e], \bar{e})$$

and

$$c_3 = \Psi(N_2^{0,0}[e, \tilde{r}_{1010}^{0,0}], \bar{e}) + \Psi(N_2^{0,0}[\bar{e}, \tilde{r}_{2000}^{0,0}], \bar{e}) + \frac{3}{2}\Psi(N_3^{0,0}[e,e,\bar{e}], \bar{e}),$$

where $\tilde{r}_{1010}^{0,0}$ and $\tilde{r}_{2000}^{0,0}$ are the unique solutions of the equations

$$\begin{aligned}L\tilde{r}_{1010}^{0,0} &= -2N_2^{0,0}[e, \bar{e}], \\ (L - 2\mathrm{i}qI)\tilde{r}_{2000}^{0,0} &= -N_2^{0,0}[e,e].\end{aligned}$$

THEOREM 4.2. *Suppose that* $c_1 < 0$ *and* $c_3 > 0$.

(i) *(Iooss and Pérouème) For each sufficiently small, positive value of* δ *the two-degree-of-freedom Hamiltonian system* (83), (84) *has two distinct symmetric homoclinic solutions.*

(ii) *(Buffoni and Groves) For each sufficiently small, positive value of* δ *the two-degree-of-freedom Hamiltonian system* (83), (84) *has an infinite number of geometrically distinct homoclinic solutions which generically resemble multiple copies of one of the homoclinic solutions in part* (i).

The homoclinic solutions identified above correspond to envelope solitary waves of amplitude $O((-c_1\delta)^{1/2})$ *which decay exponentially to a horizontal flow as* $x \to \pm\infty$*; they are sketched in Figure* 1(c).

Explicit formulas for c_1 and c_3 were computed for irrotational waves by Buffoni and Groves [7, Appendix B], and for general vorticity distributions one can prove that $c_1 < 0$, $c_3 > 0$ for values of (β_0, α_0) on the local part of C_2 near $(\beta_\star, \alpha_\star)$. To this end suppose that $(\beta_0, \alpha_0) = (\beta_\mu, \alpha_\mu)$, where

$$\beta_\mu = \beta_\star - 2\alpha_\star^2 d_1 \mu^2 + O(\mu^4), \qquad \alpha_\mu = \alpha_\star + \alpha_\star^2 d_1 \mu^4 + O(\mu^6),$$

so that $q = \mu$. We find that

$$Le_0 = \mathrm{i}\mu e_0, \qquad (L - \mathrm{i}\mu I) f_0 = e_0,$$

where

$$e_0 = \begin{pmatrix} \mathrm{i}\mu\alpha_\star^{-1} + \mathrm{i}\mu^3 \int_0^1 a^{-3}(s) \int_0^s a(t) \int_0^t a^{-3}(u)\,\mathrm{d}u\,\mathrm{d}t\,\mathrm{d}s + O_\mathrm{i}(\mu^5) \\ \mu^2 a(1)\alpha_\star^{-1} + \mu^4 a(1) \int_0^1 a^{-3}(s) \int_0^s a(t) \int_0^t a^{-3}(u)\,\mathrm{d}u\,\mathrm{d}t\,\mathrm{d}s + O_\mathrm{r}(\mu^6) \\ \mathrm{i}\mu \int_0^s a^{-3}(t)\,\mathrm{d}t + \mathrm{i}\mu^3 \int_0^s a^{-3}(t) \int_0^t a(u) \int_0^u a^{-3}(v)\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}t + O_\mathrm{i}(\mu^5) \\ \mu^2 a(s) \int_0^s a^{-3}(t)\,\mathrm{d}t + \mu^4 a(s) \int_0^s a^{-3}(t) \int_0^t a(u) \int_0^u a^{-3}(v)\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}t + O_\mathrm{r}(\mu^6) \end{pmatrix},$$

$$f_0 = \begin{pmatrix} 2\mu^2 \int_0^1 a^{-3}(s) \int_0^s a(t) \int_0^t a^{-3}(u)\,\mathrm{d}u\,\mathrm{d}t\,\mathrm{d}s + O_\mathrm{r}(\mu^4) \\ -\mathrm{i}\mu a(1)\alpha_\star^{-1} - 3\mathrm{i}\mu^3 a(1) \int_0^1 a^{-3}(s) \int_0^s a(t) \int_0^t a^{-3}(u)\,\mathrm{d}u\,\mathrm{d}t\,\mathrm{d}s + O_\mathrm{i}(\mu^5) \\ 2\mu^2 \int_0^s a^{-3}(t) \int_0^t a(u) \int_0^u a^{-3}(v)\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}t + O_\mathrm{r}(\mu^4) \\ -\mathrm{i}\mu a(s) \int_0^s a^{-3}(t)\,\mathrm{d}t - 3\mathrm{i}\mu^3 a(1) \int_0^s a^{-3}(t) \int_0^t a(u) \int_0^u a^{-3}(v)\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}t + O_\mathrm{i}(\mu^5) \end{pmatrix},$$

and the symbols $O_\mathrm{r}(\mu^n)$, $O_\mathrm{i}(\mu^n)$ denote quantities of $O(\mu^n)$ which are, respectively, real and purely imaginary. Observe that $e_0 = \mathrm{i}\mu(\phi_\mu(1), -\mathrm{i}\mu a(1)\phi_\mu(1), \phi_\mu, -\mathrm{i}\mu a(s)\phi_\mu)^\mathrm{T}$, where

$$T_\mu \tilde{\phi}_\mu = -\mu^2 \tilde{\phi}_\mu, \qquad [\tilde{\phi}_\mu, \tilde{\phi}_\mu]_\mu = 0.$$

The next step is to normalize the generalized eigenvectors. We have that

$$\Psi(e_0, \overline{f_0}) = -\Psi(f_0, \overline{e_0}) = d_4, \qquad \Psi(f_0, \overline{f_0}) = \mathrm{i}d_5,$$

where

$$d_4 = 4d_1\mu^4 + O_\mathrm{r}(\mu^6), \qquad d_5 = -4d_1\mu^3 + O_\mathrm{r}(\mu^5),$$

and the calculations

$$\Psi(e_0, \overline{e_0}) = 2\mathrm{i}\mu^3 [\tilde{\phi}_\mu, \tilde{\phi}_\mu]_\mu$$

and

$$\mathrm{i}\mu\Psi(e_0,f_0)=\Psi(Le_0,f_0)=-\Psi(e_0,Lf_0)=-\Psi(e_0,\mathrm{i}\mu f_0+e_0)=-\mathrm{i}\mu\Psi(e_0,f_0)$$

imply that $\Psi(e_0,\bar{e}_0)=\Psi(e_0,f_0)=\Psi(\bar{e}_0,\bar{f}_0)=0$. It follows that

$$e=\frac{\mathrm{i}}{d_4^{1/2}}e_0,\qquad f=\frac{\mathrm{i}}{d_4^{1/2}}\left(f_0-\frac{\mathrm{i}d_5}{2d_4}e_0\right)$$

satisfy (82) and the normalization requirements given below it.

Performing a series of lengthy calculations, one finds that

$$c_1=\frac{1}{d_4}\Psi(N_1^{0,1}[e_0],\overline{e_0})=-\frac{1}{4d_1}\alpha_\star^{-2}\mu^{-2}+O_\mathrm{r}(1)<0$$

and

$$\begin{aligned}c_3&=\frac{1}{d_4^2}\Psi(N_2^{0,0}[e_0,\tilde{R}_{1010}^{0,0}],\overline{e_0})+\frac{1}{d_4^2}\Psi(N_2^{0,0}[\overline{e_0},\tilde{R}_{2000}^{0,0}],\overline{e_0})+\frac{3}{2d_4^2}\Psi(N_3^{0,0}[e_0,e_0,\overline{e_0}],\overline{e_0})\\&=\frac{19}{64d_1^3}\left(\int_0^1 a^{-5}(s)\,\mathrm{d}s\right)^2\mu^{-8}+O_\mathrm{r}(\mu^{-6})>0,\end{aligned}$$

where

$$\begin{aligned}\tilde{R}_{1010}^{0,1}&=-2L^{-1}N_2^{0,0}[e_0,\overline{e}_0]\\&=\begin{pmatrix}-3d_1^{-1}\alpha_\star^{-1}\mu^{-2}\int_0^1 a^{-5}(s)\,\mathrm{d}s+O_\mathrm{r}(1)\\0\\-3d_1^{-1}\mu^{-2}\int_0^1 a^{-5}(s)\,\mathrm{d}s\int_0^s a^{-3}(t)\,\mathrm{d}t+O_\mathrm{r}(1)\\0\end{pmatrix},\end{aligned}$$

$$\begin{aligned}\tilde{R}_{2000}^{0,1}&=-(L-2\mathrm{i}\mu I)^{-1}N_2^{0,0}[e_0,e_0]\\&=\begin{pmatrix}\frac{1}{6}d_1^{-1}\alpha_\star^{-1}\mu^{-2}\int_0^1 a^{-5}(s)\,\mathrm{d}s+O_\mathrm{r}(1)\\-\frac{1}{3}\mathrm{i}d_1^{-1}\alpha_\star^{-1}\mu^{-1}a(1)\int_0^1 a^{-5}(s)\,\mathrm{d}s+O_\mathrm{i}(\mu)\\\frac{1}{6}d_1^{-1}\mu^{-2}\int_0^1 a^{-5}(s)\,\mathrm{d}s\int_0^s a^{-3}(t)\,\mathrm{d}s+O_\mathrm{r}(\mu)\\-\frac{1}{3}\mathrm{i}d_1^{-1}\mu^{-1}a(p)\int_0^1 a^{-5}(s)\,\mathrm{d}s\int_0^s a^{-3}(t)\,\mathrm{d}t+O_\mathrm{i}(\mu)\end{pmatrix}.\end{aligned}$$

REFERENCES

[1] C. J. Amick and K. Kirchgässner, *A theory of solitary water waves in the presence of surface tension*, Arch. Ration. Mech. Anal., 105 (1989), pp. 1–49.

[2] J. T. Beale, *The existence of solitary water waves*, Commun. Pure Appl. Math., 30 (1977), pp. 373–389.

[3] T. B. Benjamin, *Verification of the Benjamin-Lighthill conjecture about steady water waves*, J. Fluid Mech., 295 (1995), pp. 337–356.

[4] P. A. Binding, P. J. Browne, W. J. Code, and B. A. Watson, *Transformation of Sturm-Liouville problems with decreasing affine boundary conditions*, Proc. Edinb. Math. Soc., 47 (2004), pp. 533–552.

[5] J. Bognár, *Indefinite Inner Product Spaces*, Springer-Verlag, New York, 1974.
[6] B. Buffoni, A. R. Champneys, and J. F. Toland, *Bifurcation and coalescence of a plethora of homoclinic orbits for a Hamiltonian system*, J. Dynam. Differential Equations, 8 (1996), pp. 221–279.
[7] B. Buffoni and M. D. Groves, *A multiplicity result for solitary gravity-capillary waves in deep water via critical-point theory*, Arch. Ration. Mech. Anal., 146 (1999), pp. 183–220.
[8] B. Buffoni, M. D. Groves, and J. F. Toland, *A plethora of solitary gravity-capillary water waves with nearly critical Bond and Froude numbers*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 354 (1996), pp. 575–607.
[9] A. R. Champneys, J.-M. Vanden-Broeck, and G. J. Lord, *Do true elevation gravity-capillary solitary waves exist? A numerical investigation*, J. Fluid Mech., 454 (2002), pp. 403–417.
[10] A. Constantin, D. H. Sattinger, and W. Strauss, *Variational formulations for steady water waves with vorticity*, J. Fluid Mech., 548 (2006), pp. 151–163.
[11] A. Constantin and W. Strauss, *Exact steady periodic water waves with vorticity*, Commun. Pure Appl. Math., 57 (2004), pp. 481–527.
[12] R. L. Devaney, *Homoclinic orbits in Hamiltonian systems*, J. Differential Equations, 21 (1976), pp. 431–438.
[13] M. L. Dubreil-Jacotin, *Sur la détermination rigoureuse des ondes permanentes périodiques d'ampleur finite*, J. Math. Pures Appl., 13 (1934), pp. 217–291.
[14] M. D. Groves, *A new Hamiltonian formulation of the steady water-wave problem*, in Structure and Dynamics of Nonlinear Waves in Fluids, A. Mielke and K. Kirchgässner, eds., World Scientific, Singapore, 1995, pp. 259–267.
[15] M. D. Groves, *An existence theory for three-dimensional periodic travelling gravity-capillary water waves with bounded transverse profiles*, Phys. D, 152–153 (2001), pp. 395–415.
[16] M. D. Groves and M. Haragus, *A bifurcation theory for three-dimensional oblique travelling gravity-capillary water waves*, J. Nonlinear Sci., 13 (2003), pp. 397–447.
[17] M. D. Groves, M. Haragus, and S.-M. Sun, *A dimension-breaking phenomenon in the theory of gravity-capillary water waves*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 360 (2002), pp. 2189–2243.
[18] M. D. Groves and A. Mielke, *A spatial dynamics approach to three-dimensional gravity-capillary steady water waves*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 83–136.
[19] M. D. Groves and B. Sandstede, *A plethora of three-dimensional periodic travelling gravity-capillary water waves with multipulse transverse profiles*, J. Nonlinear Sci., 14 (2004), pp. 297–340.
[20] M. D. Groves and J. F. Toland, *On variational formulations for steady water waves*, Arch. Ration. Mech. Anal., 137 (1997), pp. 203–226.
[21] V. M. Hur, *Exact solitary water waves with vorticity*, Arch. Ration. Mech. Anal., to appear.
[22] I. S. Iohvidov, M. G. Krein, and H. Langer, *Introduction to the Spectral Theory of Operators in Spaces with an Indefinite Metric*, Akademie-Verlag, Berlin, 1982.
[23] G. Iooss and K. Kirchgässner, *Bifurcation d'ondes solitaires en présence d'une faible tension superficielle*, C. R. Acad. Sci. Paris, Sér. 1, 311 (1990), pp. 265–268.
[24] G. Iooss and K. Kirchgässner, *Water waves for small surface tension: An approach via normal form*, Proc. Roy. Soc. Edinburgh Sect. A, 122 (1992), pp. 267–299.
[25] G. Iooss and P. Kirrmann, *Capillary gravity waves on the free surface of an inviscid fluid of infinite depth. Existence of solitary waves*, Arch. Ration. Mech. Anal., 136 (1996), pp. 1–19.
[26] G. Iooss and M. C. Pérouème, *Perturbed homoclinic solutions in reversible* $1:1$ *resonance vector fields*, J. Differential Equations, 102 (1993), pp. 62–88.
[27] M. C. W. Jones, *Small amplitude gravity-capillary waves in a channel of finite depth*, Glasg. Math. J., 31 (1989), pp. 141–160.
[28] G. Keady and J. Norbury, *On the existence theory for irrotational water waves*, Math. Proc. Cambridge Philos. Soc., 83 (1978), pp. 137–157.
[29] K. Kirchgässner, *Nonlinearly resonant surface waves and homoclinic bifurcation*, Adv. Appl. Mech., 26 (1988), pp. 135–181.
[30] E. Lombardi, *Orbits homoclinic to exponentially small periodic orbits for a class of reversible systems. Application to water waves*, Arch. Ration. Mech. Anal., 137 (1997), pp. 227–304.
[31] A. Mielke, *Reduction of quasilinear elliptic equations in cylindrical domains with applications*, Math. Methods Appl. Sci., 10 (1988), pp. 51–66.
[32] A. Mielke, *Hamiltonian and Lagrangian Flows on Center Manifolds*, Springer-Verlag, Berlin, 1991.
[33] R. L. Sachs, *On the existence of small amplitude solitary waves with strong surface-tension*, J. Differential Equations, 90 (1991), pp. 31–51.

[34] S. M. Sun, *Nonexistence of truly solitary waves with small surface tension*, Proc. R. Soc. Lond. Ser. A, Math. Phys. Eng. Sci., 455 (1999), pp. 2191–2228.
[35] E. Wahlén, *Steady periodic capillary-gravity waves with vorticity*, SIAM J. Math. Anal., 38 (2006), pp. 921–943.
[36] E. Zeidler, *Existenzbeweis für cnoidal waves unter Berücksichtigung der Oberflächenspannung*, Arch. Ration. Mech. Anal., 41 (1971), pp. 81–107.
[37] E. Zeidler, *Existenzbeweis für permanente Kapillar-Schwerewellen mit allgemeinen Wirbelverteilungen*, Arch. Ration. Mech. Anal., 50 (1973), pp. 34–72.

SIAM J. MATH. ANAL.
Vol. 39, No. 3, pp. 965–986

WELL-POSEDNESS THEORY FOR A NONCONSERVATIVE BURGERS-TYPE SYSTEM ARISING IN DISLOCATION DYNAMICS*

AHMAD EL HAJJ†

Abstract. In this work we study a system of nonconservative Burgers type in one space dimension, arising in modeling the dynamics of dislocation densities in crystals. Starting from physically relevant initial data that are of a special form, namely nondecreasing, periodic plus linear functions, we prove the global existence and uniqueness of a solution in $H^1_{loc}(\mathbb{R} \times [0, +\infty))$ that preserves the nature of the initial data. The approach is made by adding some viscosity to the system, obtaining energy estimates, and passing to the limit for vanishing viscosity. A comparison principle is shown for this system as well as an application in the case of the classical Burgers equation.

Key words. system of Burgers equations, system of nonlinear transport equations, nonlinear hyperbolic system, dynamics of dislocation densities

AMS subject classifications. 35L45, 35Q53, 35Q72, 74H20, 74H25

DOI. 10.1137/060672170

1. Introduction.

1.1. Physical motivations and presentation of the model. Real crystals comprise certain defects in the organization of their crystalline structure called dislocations. In a particular case where these defects are parallel straight lines in the three-dimensional space, they can be viewed as points in a plan. Under the effect of exterior constraints, dislocations can move in a certain crystallographic direction called the slip direction. This slip direction is given by a vector called the "Burgers vector." The norm of this vector represents the amplitude of the generated deformation. (We refer the reader to [12] for further physical explanation.)

In this work, we are interested in the study of a one-dimensional submodel of a problem introduced by Groma and Balogh [11], initially proposed in the two-dimensional case. In fact, this one-dimensional submodel was defined by El Hajj and Forcadel [8, Lemme 3.1].

This two-dimensional model is characterized by the fact that dislocations propagate in the plane (x_1, x_2) following the two Burgers vectors $\pm\vec{b}$ with $\vec{b} = (1, 0)$. In this one-dimensional submodel we suppose also that dislocation densities depend only on the variable $x = x_1 + x_2$, which transforms the two-dimensional into a one-dimensional model (see El Hajj and Forcadel [8] for more modeling details).

More precisely this one-dimensional model is given by the following coupled

*Received by the editors October 12, 2006; accepted for publication (in revised form) April 9, 2007; published electronically October 19, 2007. This work was supported by contract JC 1025 called "ACI jeunes chercheuses et jeunes chercheurs" of the French Ministry of Research (2003–2005).
http://www.siam.org/journals/sima/39-3/67217.html
†Ecole Nationale des Ponts et Chaussées, CERMICS, 6 et 8 avenue Blaise Pascal, Cité Descartes Champs-sur-Marne, 77455 Marne-la-Vallée Cedex 2, France (elhajj@cermics.enpc.fr).

equations of nonconservative Burgers type:

(1)
$$\left\{\begin{array}{rcl} \dfrac{\partial\rho^+}{\partial t}(x,t) & = & -\left(a(t)+(\rho^+-\rho^-)(x,t)+\alpha\displaystyle\int_0^1(\rho^+-\rho^-)(y,t)dy\right)\dfrac{\partial\rho^+}{\partial x}(x,t) \\ & & \text{in } \mathcal{D}'(\mathbb{R}\times(0,T)), \\ \dfrac{\partial\rho^-}{\partial t}(x,t) & = & \left(a(t)+(\rho^+-\rho^-)(x,t)+\alpha\displaystyle\int_0^1(\rho^+-\rho^-)(y,t)dy\right)\dfrac{\partial\rho^-}{\partial x}(x,t) \\ & & \text{in } \mathcal{D}'(\mathbb{R}\times(0,T)). \end{array}\right.$$

The unknowns ρ^+ and ρ^- are scalar-valued functions, which we denote for simplicity by $\rho^\pm$. Their spatial derivatives $\frac{\partial\rho^\pm}{\partial x}$ are the dislocation densities of the Burgers vector $\pm\vec{b}=\pm(1,0)$. The function $a=a(t)$, representing the field of the imposed exterior constraint, is supposed to be independent of x, and the constant α depends on the elastic coefficients and the material size.

We consider the following initial conditions for (1):

$$\rho^\pm(x,t=0)=\rho_0^\pm(x)=\rho_0^{\pm,per}(x)+L_0x, \quad x\in\mathbb{R}, \tag{2}$$

where $\rho_0^{\pm,per}$ are 1-periodic functions. We thus modelize a periodic distribution for the $\pm$ dislocations, with a spatial period of length 1. Note that each type of $\pm$ dislocation has a mean density equal to L_0. In fact, the use of the periodic boundary conditions is a way of regarding what is going on in the interior of the material away from its boundary.

1.2. A brief review of some related literature. From a mathematical point of view, system (1) is related to other similar models such as transport equations based on vector fields with low regularity. Such equations were, for instance, studied by DiPerna and Lions in [7]. They proved the existence and uniqueness of a solution (in the renormalized sense) for vector fields in $L^1((0,+\infty);W^{1,1}_{loc}(\mathbb{R}^N))$ whose divergence is in $L^1((0,+\infty);L^\infty(\mathbb{R}^N))$. This study was generalized by Ambrosio [3], who considered vector fields in $L^1((0,+\infty);BV_{loc}(\mathbb{R}^N))$ with bounded divergence. In the present paper, we work in dimension $N=1$ and prove the existence and uniqueness of solutions of the system (1)–(2) with a vector field (i.e., the velocity) only in $L^\infty((0,+\infty),H^1_{loc}(\mathbb{R}))$.

We also refer the reader to the works of LeFloch [13] and LeFloch and Liu [14], in which they considered the study in the framework of functions of bounded variation for a system of the form

$$\left\{\begin{array}{ll} \dfrac{\partial u}{\partial t}(x,t)+A(u)\dfrac{\partial u}{\partial x}(x,t)=0, & u(x,t)\in U,\ x\in\mathbb{R},\ t\in(0,T), \\ u(x,0)=u_0(x), & x\in\mathbb{R}, \end{array}\right. \tag{3}$$

where the space of states U is an open subset of $\mathbb{R}^p$, and A is a $(p\times p)$ matrix which is of class C^1 on U. Moreover, $A(u)$ have p scalar distinct eigenvalues that we denote by $\lambda_1(u)<\lambda_2(u)<\cdots<\lambda_p(u)$. We remark that this condition on the eigenvalues does not enter into our framework even in the case where $\alpha=a=0$, because we have not sign property on $\rho^+-\rho^-$. LeFloch and Liu proved that if the initial condition u_0 is sufficiently close to a constant state, and if the total variation $TV(u_0)$ is assumed to be small enough, then system (3) admits a unique solution in

$L^\infty(\mathbb{R} \times (0, +\infty)) \cap BV(\mathbb{R} \times (0, +\infty))$, in the sense of weak entropy solutions with respect to admissible function (see LeFloch [13, Definition 3.2]).

When the system is hyperbolic and symmetric, this corresponds to the case $\alpha = a = 0$ in our system (1); it is proved in Serre [17, Thm. 3.6.1] and [18] to be a result of local existence and uniqueness in $C([0,T); H^s(\mathbb{R}^N)) \cap C^1([0,T); H^{s-1}(\mathbb{R}^N))$, with $s > \frac{N}{2} + 1$, this result being only local in time, even in dimension $N = 1$.

The assumptions of increasing initial conditions were also considered in the study of the Euler equation for compressible fluids in dimension one. With regard to these studies, we refer the reader to Chen and Wang [6, Thm. 3.1] for an existence and uniqueness result in $C^1(\mathbb{R} \times [0, +\infty))$ based on the method of characteristic. The result of Chen and Wang shows that the Euler equation of compressible fluids does not create shocks for suitable increasing and $C^1(\mathbb{R})$ initial conditions. In our case, we already knew that solutions of (1) are Lipschitz continuous; see El Hajj and Forcadel [8]. Even if this regularity question is not addressed in the present paper, we may expect some $C^1(\mathbb{R} \times [0, +\infty))$ regularity of the solution for $C^1(\mathbb{R})$ initial data.

1.3. Main result. The main result of this paper is the existence and uniqueness of global in time solutions for the system (1)–(2), modeling the dynamics of dislocation densities. This result ensures the mathematical well-posedness of the Groma–Balogh model [11] in the particular case of our interest.

THEOREM 1.1 (existence and uniqueness). *For all $T, L_0 \geq 0$, $\alpha \in \mathbb{R}$, and $\rho_0^\pm \in H^1_{loc}(\mathbb{R})$, and under the assumptions*

(H1) $\rho_0^\pm(x+1) = \rho_0^\pm(x) + L_0$ *(1-periodic function + linear function),*

(H2) $\dfrac{\partial \rho_0^\pm}{\partial x} \geq 0$ *a.e. in $\mathbb{R}$ ($\rho_0^\pm$ nondecreasing),*

(H3) $a \in L^\infty(0,T)$,

the system (1)–(2) *admits a unique solution $\rho^\pm \in H^1_{loc}(\mathbb{R} \times [0,T))$ such that, for a.e. $t \in (0,T)$, the function $\rho^\pm(.,t) : x \longmapsto \rho^\pm(x,t)$ verifies* (H1) *and* (H2).

The preceding theorem gives a global existence and uniqueness result of the system (1). Its proof is based on the following steps. First, we regularize the system (1); then we show a uniform a priori estimate in $L^\infty((0,T); H^1_{loc}(\mathbb{R}))$ for this regularized system. These estimates lead to a result of existence for long time solution and ensure the passage to the limit by compactness. Finally, the demonstration of uniqueness is done in a direct way.

THEOREM 1.2 (comparison principle for (1) with $\alpha = 0$). *Let $a(\cdot)$ satisfy* (H3) *and $\rho_1^{\pm,}$, $\rho_2^\pm \in H^1_{loc}(\mathbb{R}\times[0,T))$ be two solutions of the system* (1) *with $\alpha = 0$. Moreover, let $\rho_1^\pm(.,t)$, $\rho_2^\pm(.,t)$ verify* (H1) *and* (H2) *for a.e. $t \in (0,T)$. Then, if $\rho_1^\pm(\cdot,0) \leq \rho_2^\pm(\cdot,0)$ in $\mathbb{R}$, we have $\rho_1^\pm \leq \rho_2^\pm$ a.e. in $\mathbb{R} \times (0,T)$.*

This comparison result was crucial in a previous work [8], for the demonstration of existence and uniqueness of a Lipschitz solution to problem (1), in the sense of viscosity solution, for Lipschitz initial conditions. Here the interest of this result is a little bit secondary. Indeed, thanks to this comparison principle, we have been able to obtain indirectly $H^1_{loc}(\mathbb{R} \times [0,T))$ estimates. These estimates in turn lead to a result of existence in $H^1_{loc}(\mathbb{R} \times [0,T))$.

Our work focuses on the study of the dynamics of dislocation densities. In a different direction, let us quote some recent results on the dynamics of dislocation lines, taken individually, that are represented by nonlocal Hamilton–Jacobi equations (see [2, 9] and [1, 4] for local and global in time results, respectively).

Remark 1.3 (existence and uniqueness for the Burgers equation). We remark that these techniques can be applied to the case of the classical Burgers equations in

$W^{1,p}_{loc}(\mathbb{R}\times[0,T))$ for all $1\le p<+\infty$.

Indeed, if we consider for a given function f and initial data u_0 the equation

$$(4)\qquad \begin{cases} \dfrac{\partial u}{\partial t}+\dfrac{\partial}{\partial x}(f(u))=0 & \text{in}\quad \mathcal{D}'(\mathbb{R}\times(0,T)),\\ u(x,0)=u_0(x), & x\in\mathbb{R},\end{cases}$$

then we have the following theorem.

THEOREM 1.4. *Let $p\in[1,+\infty)$ and f be locally Lipschitz and convex; then, for all $T,L_0\ge 0$, and $u_0\in W^{1,p}_{loc}(\mathbb{R})$, they satisfy* (H1) *and* (H2). *Equation* (4) *admits a solution $u\in W^{1,p}_{loc}(\mathbb{R}\times[0,T))$, unique in the class of solutions satisfying* (H1) *and* (H2), *a.e. $t\in(0,T)$.*

1.4. Organization of the paper. In section 2, we regularize the function $a(\cdot)$ and the initial conditions and prove that the system (1)–(2) modified by the term $(\varepsilon\{\frac{\partial^2\rho^\pm}{\partial x^2}\})$ admits local in time solutions (in the "mild" sense). This will be achieved by using an application of a fixed point theorem in the space of functions in $C([0,T);H^1_{loc}(\mathbb{R}))$ and verifying (H1) for all $t\in(0,T)$. In section 3, we prove that the obtained solutions are regular and verify (H2) for all $t\in(0,T)$, with initial conditions verifying (H2). In section 4, we prove some uniform a priori estimates on the regularized solution obtained in section 3. Thanks to these estimates, we also prove the existence of global in time solutions. In section 5, we give the demonstration of Theorem 1.1, and in section 6 we prove a comparison principle result of the system (1) in the case $\alpha=0$. Finally, in section 7 we give an application of the previous results in the case of the classical Burgers equation.

2. Existence of solutions for an approximated system. In this section, we prove a theorem of existence of solutions, local in time, for the system (1) modified by the term $\varepsilon\{\frac{\partial^2\rho^\pm}{\partial x^2}\}$ after the regularization of the function $a(\cdot)$ and the initial conditions. This approximation brings us back to the study, for every $0<\varepsilon<1$, of the following system:

(5)
$$\begin{cases} \dfrac{\partial\rho^{+,\varepsilon}}{\partial t}-\varepsilon\dfrac{\partial^2\rho^{+,\varepsilon}}{\partial x^2}= & -\left(a^\varepsilon(t)+(\rho^{+,\varepsilon}-\rho^{-,\varepsilon})+\alpha\displaystyle\int_0^1(\rho^{+,\varepsilon}-\rho^{-,\varepsilon})(y,t)dy\right)\dfrac{\partial\rho^{+,\varepsilon}}{\partial x}\\ & \text{in }\mathcal{D}'(\mathbb{R}\times(0,T)),\\ \dfrac{\partial\rho^{-,\varepsilon}}{\partial t}-\varepsilon\dfrac{\partial^2\rho^{-,\varepsilon}}{\partial x^2}= & \left(a^\varepsilon(t)+(\rho^{+,\varepsilon}-\rho^{-,\varepsilon})+\alpha\displaystyle\int_0^1(\rho^{+,\varepsilon}-\rho^{-,\varepsilon})(y,t)dy\right)\dfrac{\partial\rho^{-,\varepsilon}}{\partial x}\\ & \text{in }\mathcal{D}'(\mathbb{R}\times(0,T)),\end{cases}$$

where $a^\varepsilon=\tilde a*\eta_\varepsilon$, with $\eta_\varepsilon(\cdot)=\frac{1}{\varepsilon}\eta(\frac{\cdot}{\varepsilon})$, such that $\eta\in C^\infty_c(\mathbb{R})$, η is positive, and $\int_{\mathbb{R}}\eta=1$. The function $\tilde a(\cdot)$ is an extension in $\mathbb{R}$ of the function $a(\cdot)$ by 0.

We also consider the regularized initial conditions of the system (5):

$$(6)\qquad \rho^{\pm,\varepsilon}(x,0)=\rho_0^{\pm,\varepsilon}(x)=\rho_0^{\pm,\varepsilon,per}(x)+L_0x=\rho_0^{\pm,per}*_{\mathbb{T}}\eta_\varepsilon(x)+L_0x.$$

We have the following local in time existence result for the approximated system.

THEOREM 2.1 (short time existence). *Assume* (H1) *and* (H3). *For all $\alpha\in\mathbb{R}$ and $\rho_0^\pm\in H^1_{loc}(\mathbb{R})$ there exists*

$$T^\star(\|\rho_0^{\pm,per}\|_{H^1(\mathbb{T})},\|a\|_{L^\infty(0,T)},L_0,\alpha,\varepsilon)>0$$

such that the system (5)–(6) *admits a solution* $\rho^{\pm,\varepsilon} \in C([0,T^\star); H^1_{loc}(\mathbb{R}))$ *with* $\rho^{\pm,\varepsilon}(.,t)$ *verifying* (H1).

For the proof of this theorem, see subsection 2.3. Before going on, we need to give some notation and preliminary results that will be used throughout the paper.

2.1. Notation. In what follows, we are going to use the following notation:

1. $\rho^\varepsilon = \rho^{+,\varepsilon} - \rho^{-,\varepsilon}$.
2. $\rho^{\pm,\varepsilon,per} = \rho^{\pm,\varepsilon} - L_0 x$.
3. $\mathbb{T} = (\mathbb{R}/\mathbb{Z})$ is the $[0,1)$-periodic interval.
4. Let $f = (f_1, f_2)$ be a vector such that $f_i \in H^1(\mathbb{T})$ for $i \in \{1,2\}$. The norm of f in $\left(H^1(\mathbb{T})\right)^2$ will be defined by $\|f\|_{H^1(\mathbb{T})} = \max(\|f_1\|_{H^1(\mathbb{T})}, \|f_2\|_{H^1(\mathbb{T})})$.
5. Let f be a function from $\mathbb{R} \times (0,T)$ to $\mathbb{R}$. We note by $f(t) = f(.,t) : x \longmapsto f(x,t)$.

Remark 2.2 (periodicity). According to (H1)–(H2), it is clear that ρ^ε, $\rho^{\pm,\varepsilon,per}$, and $\frac{\partial \rho^{\pm,\varepsilon}}{\partial x}$ are 1-periodic in space functions.

Under the notation of section 2.1, we know that the system (5) is equivalent to

$$\frac{\partial \rho^{\pm,\varepsilon,per}}{\partial t} - \varepsilon \frac{\partial^2 \rho^{\pm,\varepsilon,per}}{\partial x^2} = \overbrace{\mp C_\alpha[\rho^\varepsilon(t)] \frac{\partial \rho^{\pm,\varepsilon,per}}{\partial x}}^{\text{bilinear term}} \overbrace{\mp a^\varepsilon(t) \frac{\partial \rho^{\pm,\varepsilon,per}}{\partial x} \mp L_0 C_\alpha[\rho^\varepsilon(t)]}^{\text{linear term}} \mp L_0 a^\varepsilon(t) \quad \text{in } \mathbb{T} \times (0,T), \tag{7}$$

where $C_\alpha[\rho^\varepsilon(t)](x) = (\rho^\varepsilon(x,t) + \alpha \int_0^1 \rho^\varepsilon(y,t)dy)$, with the periodic initial conditions

$$\rho^{\pm,\varepsilon,per}(x,0) = \rho_0^{\pm,\varepsilon,per}(x) \quad \text{in} \quad \mathbb{T}. \tag{8}$$

2.2. Preliminary results.

LEMMA 2.3 (properties of the regularized sequence). *Under hypotheses* (H1) *and* (H3) *and for every* $\rho_0^{\pm} \in H^1_{loc}(\mathbb{R})$*, we have the following:*

1. *The functions* $\rho_0^{\pm,\varepsilon,per} \in C^\infty(\mathbb{T})$ *and verify the following estimate:*
$$\|\rho_0^{\pm,\varepsilon,per}\|_{H^1(\mathbb{T})} \le C \|\rho_0^{\pm,per}\|_{H^1(\mathbb{T})}.$$
2. *The function* $a^\varepsilon(\cdot) \in C^\infty(\mathbb{R}) \cap L^\infty(\mathbb{R})$ *and verifies the following estimate:*
$$\|a^\varepsilon\|_{L^\infty(\mathbb{R})} \le \|a\|_{L^\infty(0,T)}.$$
3. *The sequence* $a^\varepsilon(\cdot)$ *strongly converges to* $a(\cdot)$ *in* $L^2(0,T)$*. The sequences* $\rho_0^{\pm,\varepsilon,per}$ *strongly converge to* $\rho_0^{\pm,per}$ *in* $H^1(\mathbb{T})$.

The proof of this lemma is a classical property of the regularizing sequence $(\eta_\varepsilon)_\varepsilon$.

LEMMA 2.4 (mild solution). *Assume* (H3). *For every* $T \ge 0$*, if* $\rho^{\pm,\varepsilon,per} \in C([0,T); H^1(\mathbb{T}))$ *are solutions of the equation*

$$\begin{aligned} \rho^{\pm,\varepsilon,per}(x,t) &= S_\varepsilon(t)\rho_0^{\pm,\varepsilon,per} \\ &\mp L_0 \int_0^t a^\varepsilon(s)ds \mp \int_0^t S_\varepsilon(t-s)\left(C_\alpha[\rho^\varepsilon(s)] \frac{\partial \rho^{\pm,\varepsilon,per}}{\partial x}(s)\right)ds \\ &\mp \int_0^t S_\varepsilon(t-s)\left(L_0 C_\alpha[\rho^\varepsilon(s)] + a(t)\frac{\partial \rho^{\pm,\varepsilon,per}}{\partial x}(s)\right)ds, \end{aligned} \tag{9}$$

where $S_\varepsilon(t) = e^{\varepsilon t\Delta}$ *is the heat semigroup, then* $\rho^{\pm,\varepsilon,per}$ *is a solution of the system* (7)–(8) *in the sense of distributions.*

For the proof of this lemma, we refer the reader to Pazy [16, Thm. 5.2, p. 146].

LEMMA 2.5 (fixed point). *Let* E *be a Banach space,* B *be a continuous bilinear application from* $E \times E$ *to* E*, and* L *be a continuous linear application from* E *to* E *such that*

$$\|B(x,y)\|_E \leq \lambda \|x\|_E \|y\|_E \quad \textit{for all} \quad x, y \in E,$$

$$\|L(x)\|_E \leq \mu \|x\|_E \quad \textit{for all} \quad x \in E,$$

where $\lambda > 0$ *and* $\mu \in (0,1)$ *are given constants. Then, for all* $x_0 \in E$ *such that*

$$\|x_0\|_E < \frac{1}{4\lambda}(\mu - 1)^2,$$

the equation $x = x_0 + B(x,x) + L(x)$ *admits a solution in* E*.*

For the proof of this lemma we refer the reader to Cannone [5, Lem. 4.2.14].

In order to show the existence of a solution within the framework of Lemma 2.4, we apply Lemma 2.5 in the space $E = \left(L^\infty((0,T); H^1(\mathbb{T}))\right)^2$, where x_0, B, and L are defined, for $u = (u_1, u_2),\ v = (v_1, v_2) \in E$, by

(10)
$$x_0 = S_\varepsilon(t)\rho^\varepsilon_{0,vec} + L_0 \vec{i} \int_0^t a^\varepsilon(s) ds, \quad \text{where} \quad \rho^\varepsilon_{0,vec} = (\rho_0^{+,\varepsilon,per}, \rho_0^{-,\varepsilon,per}),\ \vec{i} = \begin{pmatrix} -1 \\ 1 \end{pmatrix},$$

(11)
$$B(u,v)(t) = \bar{I}_1 \int_0^t S_\varepsilon(t-s) \left(C_\alpha [u_1(s) - u_2(s)] \frac{\partial v}{\partial x}(s) \right) ds, \text{ where } \bar{I}_1 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},$$

(12)
$$L(u)(t) = L_0 \vec{i} \int_0^t S_\varepsilon(t-s) C_\alpha [u_1(s) - u_2(s)] ds + \bar{I}_1 \int_0^t S_\varepsilon(t-s) \left(a^\varepsilon(s) \frac{\partial u}{\partial x}(s) \right) ds.$$

LEMMA 2.6 (decreasing estimates). *If* $f \in L^q(\mathbb{T})$ *with* $2 \leq q \leq +\infty$ *and* $g \in L^2(\mathbb{T})$*, then for all* $t > 0$ *we have the following estimates:*

(i)

$$\|S_\varepsilon(t)(fg)\|_{L^\infty(\mathbb{T})} \leq C t^{-\frac{1}{2}} \|f\|_{L^2(\mathbb{T})} \|g\|_{L^2(\mathbb{T})}.$$

(ii)

$$\left\| \frac{\partial}{\partial x}(S_\varepsilon(t) f) \right\|_{L^2(\mathbb{T})} \leq C t^{-\frac{1}{2}} \|S_\varepsilon \left(\frac{t}{2} \right) f\|_{L^2(\mathbb{T})}.$$

(iii)

$$\left\| \frac{\partial}{\partial x} \left(S_\varepsilon(t)(fg) \right) \right\|_{L^2(\mathbb{T})} \leq C t^{-\frac{1}{2}(1+\frac{1}{q})} \|f\|_{L^q(\mathbb{T})} \|g\|_{L^2(\mathbb{T})},$$

where $C = C(\varepsilon)$ is a positive constant depending on ε.

For the proof of this lemma, see Pazy [16, Lem. 1.1.8 and Thm. 6.4.5].

PROPOSITION 2.7 (bilinear operator). *Let $F_T = \left(L^\infty((0,T); H^1(\mathbb{T}))\right)^2$. Then for every $T \geq 0$, $\alpha \in \mathbb{R}$, $u = (u_1, u_2) \in F_T$, and $v = (v_1, v_2) \in F_T$ the bilinear operator B defined in* (11) *is continuous from $F_T \times F_T$ to F_T. Moreover, there exists a positive constant $C = C(\alpha, \varepsilon)$ such that for all $u, v \in F_T$ we have*

$$\|B(u,v)\|_{F_T} \leq CT^{\frac{1}{2}} \|u\|_{F_T} \|v\|_{F_T}.$$

Proof of Proposition 2.7. First, we know that

$$\begin{aligned} \|B(u,v)(t)\|_{H^1(\mathbb{T})} &= \left\| \bar{I}_1 \int_0^t S_\varepsilon(t-s) \left(C_\alpha [u_1(s) - u_2(s)] \frac{\partial v}{\partial x}(s) \right) ds \right\|_{H^1(\mathbb{T})} \\ &\leq \int_0^t \left\| S_\varepsilon(t-s) \left(C_\alpha [u_1(s) - u_2(s)] \frac{\partial v}{\partial x}(s) \right) \right\|_{H^1(\mathbb{T})} ds. \end{aligned}$$

Then, since $L^\infty(\mathbb{T}) \hookrightarrow L^2(\mathbb{T})$, we have

$$\begin{aligned} \|B(u,v)(t)\|_{H^1(\mathbb{T})} &\leq \int_0^t \left\| S_\varepsilon(t-s) \left(C_\alpha [u_1(s) - u_2(s)] \frac{\partial v}{\partial x}(s) \right) \right\|_{L^\infty(\mathbb{T})} ds \\ &+ \int_0^t \left\| \frac{\partial}{\partial x} S_\varepsilon(t-s) \left(C_\alpha [u_1(s) - u_2(s)] \frac{\partial v}{\partial x}(s) \right) \right\|_{L^2(\mathbb{T})} ds. \end{aligned}$$

Using Lemma 2.6(i) for the first term and Lemma 2.6(iii) with $q = \infty$ for the second term, we can conclude that

$$\begin{aligned} \|B(u,v)(t)\|_{H^1(\mathbb{T})} &\leq C \int_0^t \frac{1}{(t-s)^{\frac{1}{2}}} \|C_\alpha [u_1(s) - u_2(s)]\|_{L^\infty(\mathbb{T})} \left\| \frac{\partial v}{\partial x}(s) \right\|_{L^2(\mathbb{T})} ds \\ &\leq C \sup_{0 \leq t < T} (\|u(t)\|_{H^1(\mathbb{T})}) \sup_{0 \leq t < T} (\|v(t)\|_{H^1(\mathbb{T})}) \int_0^t \frac{1}{(t-s)^{\frac{1}{2}}} ds. \end{aligned}$$

Then, for all $t \in (0,T)$, we have

$$\begin{aligned} \|B(u,v)(t)\|_{H^1(\mathbb{T})} &\leq Ct^{\frac{1}{2}} \|u\|_{L^\infty((0,T);H^1(\mathbb{T}))^2} \|v\|_{L^\infty((0,T);H^1(\mathbb{T}))^2} \\ &\leq CT^{\frac{1}{2}} \|u\|_{L^\infty((0,T);H^1(\mathbb{T}))^2} \|v\|_{L^\infty((0,T);H^1(\mathbb{T}))^2}. \end{aligned} \tag{13}$$

□

PROPOSITION 2.8 (linear operator). *Let $F_T = \left(L^\infty((0,T); H^1(\mathbb{T}))\right)^2$ and $a(\cdot)$ satisfy* (H3). *Then for all L_0 $T \geq 0$, and $u = (u_1, u_2) \in F_T$, the linear operator L defined in* (12) *is continuous from F_T to F_T. Moreover, there exists a positive constant $C = C(\alpha, \varepsilon, \|a\|_{L^\infty(0,T)}, L_0)$ such that*

$$\|L(u)\|_{F_T} \leq CT^{\frac{1}{2}} \|u\|_{F_T}.$$

The proof of Proposition 2.8 is similar to the one used in Proposition 2.7.

LEMMA 2.9. *For all L_0, $T \geq 0$, and $a(\cdot)$ satisfying* (H3), *if*

$$X_{a^\varepsilon}(t) = L_0 \vec{i} \int_0^t a^\varepsilon(s) ds, \quad t \in (0,T),$$

then

$$\|X_{a^\varepsilon}\|_{(L^\infty(0,T))^2} \leq L_0 T \|a\|_{L^\infty(0,T)}.$$

The proof of Lemma 2.9 is trivial (from Lemma 2.3(2)).

LEMMA 2.10 (continuity of the semigroup). *For all $f \in W^{2,2}(\mathbb{T})$ and $0 \le \theta < t$, we have the following estimates:*

(i)

$$\|(S_\varepsilon(t-\theta) - Id)f\|_{L^2(\mathbb{T})} \le C(t-\theta) \left\| \frac{\partial^2 f}{\partial x^2} \right\|_{L^2(\mathbb{T})}.$$

(ii)

$$\|(S_\varepsilon(t-\theta) - Id)f\|_{L^2(\mathbb{T})} \le 2\|f\|_{L^2(\mathbb{T})},$$

where $C = C(\varepsilon)$ is a positive constant depending on ε.

We refer the reader to Pazy [16, Lem. 6.2, p. 151] for the proof of this lemma.

LEMMA 2.11 (time continuity). *Assume* (H3). *If $\rho_{0,vec} = (\rho_0^{+,per}, \rho_0^{-,per}) \in (H^1(\mathbb{T}))^2$, then for all $T \ge 0$ and $u = (u_1, u_2) \in \big(L^\infty((0,T); H^1(\mathbb{T}))\big)^2$, we have the following applications:*

(A1) $t \to X_{a^\varepsilon}(t)$;

(A2) $t \to S_\varepsilon(t)\rho^\varepsilon_{0,vec}$, *where* $\rho^\varepsilon_{0,vec} = (\rho_0^{+,\varepsilon,per}, \rho_0^{-,\varepsilon,per})$;

(A3) $t \to B(u,u)(t)$;

(A4) $t \to L(u)(t)$ *are* $\big(C([0,T); H^1(\mathbb{T}))\big)^2$, *where X_{a^ε}, B, and L are defined in Lemma* 2.9, (11), *and* (12), *respectively.*

Proof of Lemma 2.11. The continuity of (A1) is trivial since $a \in L^\infty(0,T)$. From the fact that the semigroup $S_\varepsilon(\cdot)$ is continuous from $[0,T)$ to $(H^1(\mathbb{T}))^2$, we deduce the continuity of (A2).

It remains to prove the continuity of (A3) and (A4). Indeed, the continuity of (A3) at 0 is a consequence of inequality (13). Now we are going to prove the continuity of (A3) for all $\theta \in (0,T)$. For all t, such that $\theta < t \le \min(T, \frac{3\theta}{2})$, we write $t = (1+\gamma)\theta$ and denote $\tau = (1-\gamma)\theta$ (where $0 < \gamma \le \frac{1}{2}$), and we write

$$\begin{aligned} B(u,u)(t) - B(u,u)(\theta) &= \int_0^\tau (S(t-s) - S(\theta-s)) \left(C_\alpha[u_1(s) - u_2(s)] \frac{\partial u}{\partial x}(s) \right) ds \\ &\quad + \int_\tau^\theta (S(t-s) - S(\theta-s)) \left(C_\alpha[u_1(s) - u_2(s)] \frac{\partial u}{\partial x}(s) \right) ds \\ &\quad + \int_\theta^t S(t-s) \left(C_\alpha[u_1(s) - u_2(s)] \frac{\partial u}{\partial x}(s) \right) ds \end{aligned}$$

$$\begin{aligned} = & \overbrace{\int_0^\tau ((S(t-\theta) - Id)S(\theta-s)) \left(C_\alpha[u_1(s) - u_2(s)] \frac{\partial u}{\partial x}(s) \right) ds}^{I_1} \\ & + \overbrace{\int_\tau^\theta ((S(t-\theta) - Id)S(\theta-s)) \left(C_\alpha[u_1(s) - u_2(s)] \frac{\partial u}{\partial x}(s) \right) ds}^{I_2} \\ & + \int_\theta^t S(t-s) \left(C_\alpha[u_1(s) - u_2(s)] \frac{\partial u}{\partial x}(s) \right) ds. \end{aligned}$$

We apply Lemma 2.10(i) and Lemma 2.6(ii) to find an upper bound to I_1. We then apply Lemma 2.10(ii) to find an upper bound to I_2. After that, we follow the same steps of the proof of Proposition 2.7 to conclude that

$$\begin{aligned}\|B(u,u)(t)-B(u,u)(\theta)\|_{H^1}\leq\quad & C(t-\theta)\|u\|^2_{(L^\infty((0,T);H^1(\mathbb{T})))^2}\int_0^\tau\frac{1}{(\theta-s)^{\frac{3}{2}}}ds\\ & +C\|u\|^2_{(L^\infty((0,T);H^1(\mathbb{T})))^2}\int_\tau^\theta\frac{1}{(\theta-s)^{\frac{1}{2}}}ds\\ & +C\|u\|^2_{(L^\infty((0,T);H^1(\mathbb{T})))^2}\int_\theta^t\frac{1}{(t-s)^{\frac{1}{2}}}ds.\end{aligned}$$

After the computation of each integral we deduce that

$$\begin{aligned}\|B(u,u)(t)-B(u,u)(\theta)\|_{H^1}\leq\quad & C(t-\theta)\left(\tfrac{1}{(\theta-\tau)^{\frac{1}{2}}}-\tfrac{1}{\theta^{\frac{1}{2}}}\right)\|u\|^2_{(L^\infty((0,T);H^1(\mathbb{T})))^2}\\ & +C\left((\theta-\tau)^{\frac{1}{2}}+(t-\theta)^{\frac{1}{2}}\right)\|u\|^2_{(L^\infty((0,T);H^1(\mathbb{T})))^2}.\end{aligned}$$

Observing that $t-\theta=\theta-\tau=\gamma\theta$ we finally obtain the following inequality:

$$\|B(u,u)(t)-B(u,u)(\theta)\|_{H^1}\leq C(\theta,\gamma)\left((t-\theta)^{\frac{1}{2}}+(t-\theta)\right)\|u\|^2_{(L^\infty((0,T);H^1(\mathbb{T})))^2},$$

and hence we get the continuity of (A3). In the same way we get the continuity in time of (A4). ∎

2.3. Proof of Theorem 2.1. We rewrite the system (9) in the following vectorial form:

$$\begin{aligned}\rho^\varepsilon_{vec}(\cdot,t)=\quad & S_\varepsilon(t)\rho^\varepsilon_{0,vec}+L_0\vec{i}\int_0^t a^\varepsilon(s)ds+\bar{I}_1\int_0^t S_\varepsilon(t-s)\left(C_\alpha[\rho^\varepsilon(s)]\frac{\partial\rho^\varepsilon_{vec}}{\partial x}(s)\right)ds\\ & +L_0\vec{i}\int_0^t S_\varepsilon(t-s)C_\alpha[\rho^\varepsilon(s)]ds+\bar{I}_1\int_0^t S_\varepsilon(t-s)\left(a^\varepsilon(s)\frac{\partial\rho^\varepsilon_{vec}}{\partial x}(s)\right)ds\end{aligned}$$

such that ρ^ε_{vec} is the vector $(\rho^{+,\varepsilon,per},\rho^{-,\varepsilon,per})$ and $\rho^\varepsilon_{0,vec}$ is the vector $(\rho_0^{+,\varepsilon,per},\rho_0^{-,\varepsilon,per})$. $\vec{i}$ and $\bar{I}_1$ are defined in (10) and (11), respectively.

This altogether leads to the following equation:

$$\rho^\varepsilon_{vec}(\cdot,t)=S_\varepsilon(t)\rho^\varepsilon_{0,vec}+X_{a^\varepsilon}(t)+B(\rho^\varepsilon_{vec},\rho^\varepsilon_{vec})(t)+L(\rho^\varepsilon_{vec})(t),\tag{14}$$

where B is the bilinear application and L is the linear application defined in (11) and (12), respectively, and X_{a^ε} is defined in Lemma 2.9. Moreover, according to Lemmas 2.9 and 2.3 we know that

$$\begin{aligned}\|S(t)\rho^\varepsilon_{0,vec}+X_{a^\varepsilon}(t)\|_{(L^\infty((0,T);H^1(\mathbb{T})))^2}\quad & \leq\|\rho^\varepsilon_{0,vec}\|_{H^1(\mathbb{T})}+L_0T\|a^\varepsilon\|_{L^\infty(\mathbb{R})}\\ & \leq C_0\|\rho_{0,vec}\|_{H^1(\mathbb{T})}+L_0T\|a\|_{L^\infty(0,T)}.\end{aligned}$$

In order to apply Lemma 2.5, we want, for a well-chosen time T, that the following inequality holds:

$$C_0\|\rho^0_{vec}\|_{H^1(\mathbb{T})}+L_0T\|a\|_{L^\infty(0,T)}<\frac{1}{4CT^{\frac{1}{2}}}(CT^{\frac{1}{2}}-1)^2,\quad\text{and } CT^{\frac{1}{2}}<1,\tag{15}$$

where C is the largest constant between the two constants computed in Propositions 2.8 and 2.7. For

$$(T^\star)^{\frac{1}{2}}(\|\rho_{0,vec}\|_{H^1(\mathbb{T})}, \|a\|_{L^\infty(0,T)}, L_0, \varepsilon) = \min\left(1, \frac{1}{2C}, \frac{1}{16C(C_0\|\rho^0_{vec}\|_{H^1(\mathbb{T})} + L_0\|a\|_{L^\infty(0,T)})}\right), \tag{16}$$

we can easily verify that $T^\star$ satisfies the inequality (15). We apply Lemma 2.5 over the space $F_{T^\star} = \left(L^\infty((0,T^\star); H^1(\mathbb{T}))\right)^2$ to prove the existence of a solution for the system (14) in $F_{T^\star}$.

Then, according to Lemma 2.11, we deduce that the obtained solution is $(C([0,T^\star); H^1(\mathbb{T})))^2$.

This proves, by Lemma 2.4, the existence of a solution in the sense of distributions for the system (5)–(6) in $C([0,T^\star); H^1_{loc}(\mathbb{R}))$ that verifies (H1). ∎

3. Properties of the solution to the approximated system. In this section we show that the solutions of system (5)–(6) obtained in the previous section are regular and verify (H2), provided that initial conditions verify (H2).

Lemma 3.1 (regularity of the solution). *Assume* (H1), (H3), *and* $\rho_0^\pm \in H^1_{loc}(\mathbb{R})$; *if* $\rho^{\pm,\varepsilon} \in C([0,T); H^1_{loc}(\mathbb{R}))$ *are solutions of the system* (5)–(6), *then* $\rho^{\pm,\varepsilon} \in C^\infty(\mathbb{R} \times [0,T))$.

Proof of Lemma 3.1. If we denote the second term of the system (7) by

$$f^\pm_{a^\varepsilon,\alpha}[\rho^\varepsilon(t)] = \mp a^\varepsilon(t)\left(L_0 + \frac{\partial \rho^{\pm,\varepsilon,per}}{\partial x}\right) \mp C_\alpha[\rho^\varepsilon(t)]\left(\frac{\partial \rho^{\pm,\varepsilon,per}}{\partial x} + L_0\right),$$

we know that $f^\pm_{a^\varepsilon,\alpha}[\rho^\varepsilon] \in L^2(\mathbb{T}\times(0,T))$. Moreover, we know that the initial conditions $\rho_0^{\pm,\varepsilon,per} \in C^\infty(\mathbb{T})$, which allows us to apply the L^2 regularity of the heat equation over the system (7)–(8) (see Lions and Magenes [15, Thm. 8.2]). Then we deduce by induction that the solution is $C^\infty(\mathbb{T}\times[0,T))$. ∎

Lemma 3.2 (monotonicity of the solution in space). *Assume* (H1), (H2), (H3), *and* $\rho_0^\pm \in H^1_{loc}(\mathbb{R})$; *if* $\rho^{\pm,\varepsilon} \in C^\infty(\mathbb{R}\times[0,T))$ *are solutions of the system* (5)–(6), *then* $\rho^{\pm,\varepsilon}(.,t)$ *verifies* (H2) *for all* $t \in (0,T)$.

Proof of Lemma 3.2. First, we remark that if $\frac{\partial \rho_0^\pm}{\partial x} \geq 0$, then $\frac{\partial \rho_0^{\pm,\varepsilon}}{\partial x} \geq 0$. Indeed, we have

$$\begin{aligned}\frac{\partial \rho_0^{\pm,\varepsilon}}{\partial x} = \frac{\partial \rho_0^{\pm,per}}{\partial x} * \eta_\varepsilon + L_0 \quad &= \left(\frac{\partial \rho_0^{\pm,per}}{\partial x} + L_0\right) * \eta_\varepsilon \\ &= \left(\frac{\partial \rho_0^{\pm}}{\partial x}\right) * \eta_\varepsilon \geq 0, \quad \text{because } \eta \text{ is positive.}\end{aligned}$$

We apply the maximum principle over the derived system of (5)–(6):

$$\begin{cases} \dfrac{\partial \theta^{\pm,\varepsilon}}{\partial t} - \varepsilon \dfrac{\partial^2 \theta^{\pm,\varepsilon}}{\partial x^2} \pm (C_\alpha[\rho^\varepsilon(t)] + a^\varepsilon(t))\dfrac{\partial \theta^{\pm,\varepsilon}}{\partial x} \pm (\theta^{+,\varepsilon} - \theta^{-,\varepsilon})\theta^{\pm,\varepsilon} = 0 \\ \qquad\qquad \text{in} \quad \mathbb{T}\times(0,T), \\ \theta^{\pm,\varepsilon}(x,0) = \dfrac{\partial \rho_0^{\pm,\varepsilon}}{\partial x}, \end{cases}$$

where $\theta^{\pm,\varepsilon} = \frac{\partial \rho^{\pm,\varepsilon}}{\partial x}$ (see Gilbarg and Trudinger [10, Thm. 8.1]). Since $\rho^{\pm,\varepsilon} \in C^\infty(\mathbb{R}\times[0,T))$, we deduce that $\theta^{\pm,\varepsilon} \geq 0$ belongs to $\mathbb{T}\times(0,T)$. $\square$

Corollary 3.3 (short time existence of nondecreasing regular solutions). *For all $\alpha \in \mathbb{R}$ and $\rho_0^\pm \in H^1_{loc}(\mathbb{R})$, under the assumptions* (H1), (H2), *and* (H3), *there exists*

$$T^\star(\|\rho_0^{\pm,per}\|_{H^1(\mathbb{T})}, \|a\|_{L^\infty(0,T)}, L_0, \alpha, \varepsilon) > 0$$

such that the system (5)–(6) *admits a solution $\rho^{\pm,\varepsilon} \in C^\infty(\mathbb{R}\times[0,T^\star))$ with $\rho^{\pm,\varepsilon}(.,t)$ verifying* (H1) *and* (H2).

The proof of Corollary 3.3 is a consequence of Theorem 2.1 and Lemmas 3.1 and 3.2 (with $T = T^\star$).

Remark 3.4. Here we remark that the case of nondecreasing solutions corresponds to a nonshock case in the Burgers equation. On the other hand, the decreasing solutions represent the shock case.

4. A priori estimates and long time existence for the approximated system. In this section, we are going to show some ε-uniform estimates on the solutions of the system (5)–(6). These estimates will be used in section 4 for the passage to the limit as ε tends to zero.

Lemma 4.1 (L^2 estimates over the space derivatives of the solutions). *Assume* (H1), (H2), (H3), *and $\rho_0^\pm \in H^1_{loc}(\mathbb{R})$; if $\rho^{\pm,\varepsilon} \in C^\infty(\mathbb{R}\times[0,T))$ is a solution of the system* (5)–(6) *for all $T \geq 0$, then*

$$\left\|\frac{\partial \rho^{+,\varepsilon}}{\partial x}\right\|^2_{L^\infty((0,T);L^2(\mathbb{T}))} + \left\|\frac{\partial \rho^{-,\varepsilon}}{\partial x}\right\|^2_{L^\infty((0,T);L^2(\mathbb{T}))} \leq CB_0,$$

with $B_0 = (\|\frac{\partial \rho_0^+}{\partial x}\|^2_{L^2(\mathbb{T})} + \|\frac{\partial \rho_0^-}{\partial x}\|^2_{L^2(\mathbb{T})})$.

Proof of Lemma 4.1. If we denote $\rho^\varepsilon = \rho^{+,\varepsilon} - \rho^{-,\varepsilon}$ and $k^\varepsilon = \rho^{+,\varepsilon} + \rho^{-,\varepsilon}$, then, according to (H1), it is clear that ρ^ε, $\frac{\partial \rho^\varepsilon}{\partial x}$, and $\frac{\partial k^\varepsilon}{\partial x}$ are 1-periodic functions. Moreover, by Lemma 3.2, we know that $\frac{\partial k^\varepsilon}{\partial x} \geq 0$.

If we take into consideration the equations of the system (5), we can conclude that ρ^ε and k^ε verify the following system:

$$(17)\quad \begin{cases} \dfrac{\partial \rho^\varepsilon}{\partial t} - \varepsilon \dfrac{\partial^2 \rho^\varepsilon}{\partial x^2} = -\left(\rho^\varepsilon + \alpha \displaystyle\int_0^1 \rho^\varepsilon dx + a^\varepsilon(t)\right)\dfrac{\partial k^\varepsilon}{\partial x} & \text{in} \quad \mathcal{D}'(\mathbb{R}\times(0,T)), \\ \dfrac{\partial k^\varepsilon}{\partial t} - \varepsilon \dfrac{\partial^2 k^\varepsilon}{\partial x^2} = -\left(\rho^\varepsilon + \alpha \displaystyle\int_0^1 \rho^\varepsilon dx + a^\varepsilon(t)\right)\dfrac{\partial \rho^\varepsilon}{\partial x} & \text{in} \quad \mathcal{D}'(\mathbb{R}\times(0,T)). \end{cases}$$

We derive the first equation of the system (17) with respect to x, then we multiply the result by $\frac{\partial \rho^\varepsilon}{\partial x}$, and, finally, we integrate in space. For all $t \in (0,T)$, we then obtain

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\left\|\frac{\partial \rho^\varepsilon}{\partial x}(t)\right\|^2_{L^2(\mathbb{T})} + \varepsilon\left\|\frac{\partial^2 \rho^\varepsilon}{\partial x^2}(t)\right\|^2_{L^2(\mathbb{T})} = & -\int_0^1 \left(\frac{\partial \rho^\varepsilon}{\partial x}\right)^2 \frac{\partial k^\varepsilon}{\partial x} - \int_0^1 \rho^\varepsilon \frac{\partial \rho^\varepsilon}{\partial x}\frac{\partial^2 k^\varepsilon}{\partial x^2} \\ & -\left(\alpha\int_0^1 \rho^\varepsilon + a^\varepsilon(t)\right)\int_0^1 \frac{\partial^2 k^\varepsilon}{\partial x^2}\frac{\partial \rho^\varepsilon}{\partial x}. \end{aligned}$$

Now we proceed in the same way as for the previous equation, but we multiply the second equation of the system (17) by $\frac{\partial k^\varepsilon}{\partial x}$. For every $t \in (0,T)$, we obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\left\|\frac{\partial k^\varepsilon}{\partial x}(t)\right\|^2_{L^2(\mathbb{T})} + \varepsilon\left\|\frac{\partial^2 k^\varepsilon}{\partial x^2}(t)\right\|^2_{L^2(\mathbb{T})} = \; & -\int_0^1\left(\frac{\partial\rho^\varepsilon}{\partial x}\right)^2\frac{\partial k^\varepsilon}{\partial x} - \int_0^1 \rho^\varepsilon\frac{\partial k^\varepsilon}{\partial x}\frac{\partial^2\rho^\varepsilon}{\partial x^2}\\ & -\left(\alpha\int_0^1\rho^\varepsilon + a^\varepsilon(t)\right)\int_0^1\frac{\partial^2\rho^\varepsilon}{\partial x^2}\frac{\partial k^\varepsilon}{\partial x}.\end{aligned}$$

Adding the two previous equations, thanks to the periodicity of ρ^ε and $\frac{\partial k^\varepsilon}{\partial x}$, we infer that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\left\|\frac{\partial\rho^\varepsilon}{\partial x}(t)\right\|^2_{L^2(\mathbb{T})} + \frac{1}{2}\frac{d}{dt}\left\|\frac{\partial k^\varepsilon}{\partial x}(t)\right\|^2_{L^2(\mathbb{T})} \le & -\int_0^1\left(\frac{\partial\rho^\varepsilon}{\partial x}\right)^2\frac{\partial k^\varepsilon}{\partial x} - \int_0^1\frac{\partial}{\partial x}\left(\rho^\varepsilon\frac{\partial\rho^\varepsilon}{\partial x}\frac{\partial k^\varepsilon}{\partial x}\right)\\ & -\left(\alpha\int_0^1\rho^\varepsilon + a^\varepsilon(t)\right)\int_0^1\frac{\partial}{\partial x}\left(\frac{\partial\rho^\varepsilon}{\partial x}\frac{\partial k^\varepsilon}{\partial x}\right)\\ \le & -\int_0^1\left(\frac{\partial\rho^\varepsilon}{\partial x}\right)^2\frac{\partial k^\varepsilon}{\partial x} \le 0.\end{aligned}$$

We integrate in time and use the fact that $\rho^{\pm,\varepsilon} \in C^\infty(\mathbb{R}\times[0,T))$ and Lemma 2.3. We obtain, in particular,

$$\begin{aligned}&\sup_{t\in(0,T)}\left\|\frac{\partial\rho^\varepsilon}{\partial x}(t)\right\|^2_{L^2(\mathbb{T})} + \sup_{t\in(0,T)}\left\|\frac{\partial k^\varepsilon}{\partial x}(t)\right\|^2_{L^2(\mathbb{T})}\\ &\quad\le C\left(\left\|\frac{\partial(\rho_0^+ - \rho_0^-)}{\partial x}\right\|^2_{L^2(\mathbb{T})} + \left\|\frac{\partial(\rho_0^+ + \rho_0^-)}{\partial x}\right\|^2_{L^2(\mathbb{T})}\right).\end{aligned}$$

That leads to the desired result. ∎

Lemma 4.2 (L^2 estimates of the solutions). *Assume* (H1), (H2), (H3), *and* $\rho_0^\pm \in H^1_{loc}(\mathbb{R})$; *if* $\rho^{\pm,\varepsilon} \in C^\infty(\mathbb{R}\times[0,T))$ *are solutions of the system* (5)–(6) *for every* $T \ge 0$, *then*

$$\begin{aligned}&\|\rho^{+,\varepsilon}\|^2_{L^\infty((0,T);L^2(0,1))} + \|\rho^{-,\varepsilon}\|^2_{L^\infty((0,T);L^2(0,1))}\\ &\quad\le C\left(M_0 + (B_0 + \|a\|^2_{L^\infty(0,T)})\right)e^{4L_0(1+\alpha^2)T},\end{aligned}$$

where B_0 *is defined in Lemma* 4.1, *and* $M_0 = (\|\rho_0^+\|^2_{L^2(0,1)} + \|\rho_0^-\|^2_{L^2(0,1)})$.

Proof of Lemma 4.2. We will use the same procedure of the proof of Lemma 4.1. We multiply the first equation of the system (17) by ρ^ε; then we integrate in space. For every $t \in (0,T)$, we obtain

$$\frac{1}{2}\frac{d}{dt}\|\rho^\varepsilon(t)\|^2_{L^2(\mathbb{T})} + \varepsilon\left\|\frac{\partial\rho^\varepsilon}{\partial x}(t)\right\|^2_{L^2(\mathbb{T})} = -\int_0^1(\rho^\varepsilon)^2\frac{\partial k^\varepsilon}{\partial x} - \left(\alpha\int_0^1\rho^\varepsilon + a^\varepsilon(t)\right)\int_0^1\rho^\varepsilon\frac{\partial k^\varepsilon}{\partial x}.$$

Similarly, we multiply the second equation of the system (17) by k^ε and integrate in space. For every $t \in (0,T)$, we obtain

$$\frac{1}{2}\frac{d}{dt}\|k^\varepsilon(t)\|^2_{L^2(0,1)} + \varepsilon\left\|\frac{\partial k^\varepsilon}{\partial x}(t)\right\|^2_{L^2(\mathbb{T})} = -\int_0^1\rho^\varepsilon\frac{\partial\rho^\varepsilon}{\partial x}k^\varepsilon - \left(\alpha\int_0^1\rho^\varepsilon + a^\varepsilon(t)\right)\int_0^1 k^\varepsilon\frac{\partial\rho^\varepsilon}{\partial x}.$$

Now we add the two previous equations and get

$$
\begin{aligned}
&\frac{1}{2}\left(\frac{d}{dt}\|\rho^\varepsilon(t)\|^2_{L^2(\mathbb{T})}+\frac{d}{dt}\|k^\varepsilon(t)\|^2_{L^2(0,1)}\right)\\
&\quad\le -\int_0^1\left((\rho^\varepsilon)^2\frac{\partial k^\varepsilon}{\partial x}+\frac{1}{2}k^\varepsilon\frac{\partial(\rho^\varepsilon)^2}{\partial x}\right)\\
&\qquad -\left(\alpha\int_0^1\rho^\varepsilon+a^\varepsilon(t)\right)\left(\int_0^1 k^\varepsilon\frac{\partial\rho^\varepsilon}{\partial x}+\int_0^1\rho^\varepsilon\frac{\partial k^\varepsilon}{\partial x}\right)\\
&\quad\le -\frac{1}{2}\int_0^1(\rho^\varepsilon)^2\frac{\partial k^\varepsilon}{\partial x}-\frac{1}{2}\int_0^1\frac{\partial((\rho^\varepsilon)^2k^\varepsilon)}{\partial x}\\
&\qquad -\left(\alpha\int_0^1\rho^\varepsilon+a^\varepsilon(t)\right)\int_0^1\frac{\partial(k^\varepsilon\rho^\varepsilon)}{\partial x}.
\end{aligned}
$$

Recalling that ρ^ε is periodic and k^ε is nondecreasing, we see that

$$
\frac{1}{2}\left(\frac{d}{dt}\|\rho^\varepsilon(t)\|^2_{L^2(\mathbb{T})}+\frac{d}{dt}\|k^\varepsilon(t)\|^2_{L^2(0,1)}\right)\le -\left(\alpha\int_0^1\rho^\varepsilon+a^\varepsilon(t)\right)\int_0^1\frac{\partial(k^\varepsilon\rho^\varepsilon)}{\partial x}.
$$

But we know from (H1) that ρ^ε and $(k^\varepsilon-2L_0x)$ are 1-periodic functions, which implies that

$$
\int_0^1\frac{\partial(k^\varepsilon\rho^\varepsilon)}{\partial x}=\int_0^1\frac{\partial((k^\varepsilon-2L_0x)\rho^\varepsilon)}{\partial x}+2L_0\int_0^1\frac{\partial(x\rho^\varepsilon)}{\partial x}=2L_0\int_0^1 x\frac{\partial\rho^\varepsilon}{\partial x}+2L_0\int_0^1\rho^\varepsilon.
$$

We use Lemmas 4.1 and 2.3 and the fact that ($ab\le\frac{1}{2}(a^2+b^2)$ and $(a+b)^2\le 2(a^2+b^2)$) to deduce that

$$
\begin{aligned}
&\frac{d}{dt}\left(\|k^\varepsilon(t)\|^2_{L^2(0,1)}+\|\rho^\varepsilon(t)\|^2_{L^2(\mathbb{T})}\right)\\
&\quad\le 4L_0\left(|\alpha|\,\|\rho^\varepsilon(t)\|_{L^2(\mathbb{T})}+\|a\|_{L^\infty(0,T)}\right)\left(\left\|\frac{\partial\rho^\varepsilon}{\partial x}(t)\right\|_{L^2(\mathbb{T})}+\|\rho^\varepsilon(t)\|_{L^2(\mathbb{T})}\right)\\
&\quad\le 4L_0\left(\|a\|^2_{L^\infty(0,T)}+(1+\alpha^2)\|\rho^\varepsilon(t)\|^2_{L^2(\mathbb{T})}+\left\|\frac{\partial\rho^\varepsilon}{\partial x}(t)\right\|^2_{L^2(\mathbb{T})}\right)\\
&\quad\le 4L_0\left(CB_0+\|a\|^2_{L^\infty(0,T)}\right)+4L_0(1+\alpha^2)\left(\|\rho^\varepsilon(t)\|^2_{L^2(\mathbb{T})}+\|k^\varepsilon(t)\|^2_{L^2(0,1)}\right).
\end{aligned}
$$

Using the previous estimate and the fact that $\rho^{\pm,\varepsilon}\in C^\infty(\mathbb{R}\times[0,T))$, we finally obtain

$$
\|\rho^\varepsilon\|^2_{L^\infty((0,T);L^2(\mathbb{T}))}+\|k^\varepsilon\|^2_{L^\infty((0,T)L^2(0,1))}\le C\left(M_0+B_0+\|a\|^2_{L^\infty(0,T)}\right)e^{4L_0(1+\alpha^2)T}.
$$

This leads to the desired result. $\square$

LEMMA 4.3 (L^2 estimate on the time derivatives of the solutions). *Assume* (H1), (H2), (H3), *and* $\rho_0^\pm \in H^1_{loc}(\mathbb{R})$; *if* $\rho^{\pm,\varepsilon} \in C^\infty(\mathbb{R}\times[0,T))$ *is a solution of the system* (5)–(6) *for every* $T \geq 0$, *then there exists a constant* $C\left(T, L_0, \alpha, \|a\|_{L^\infty(0,T)}, M_0, B_0\right)$ *independent of* ε *such that*

$$\left\|\frac{\partial \rho^{\pm,\varepsilon}}{\partial t}\right\|_{L^2(\mathbb{T}\times(0,T))} \leq C.$$

Proof of Lemma 4.3. For the proof of Lemma 4.3, it is sufficient to show that the second term of the system (5),

$$f^\pm_{a^\varepsilon,\alpha}[\rho^\varepsilon(t)] = \mp\left(a^\varepsilon(t) + \rho^\varepsilon + \alpha\int_0^1 \rho^\varepsilon dx\right)\frac{\partial \rho^{\pm,\varepsilon}}{\partial x},$$

is bounded in $L^\infty((0,T);L^2(\mathbb{T}))$ uniformly in ε. Indeed,

$$\begin{aligned}
&\left\|f^\pm_{a^\varepsilon,\alpha}[\rho^\varepsilon]\right\|_{L^\infty((0,T^\star);L^2(\mathbb{T}))}\\
&\quad\leq \left\|\left(a^\varepsilon(\cdot) + \rho^\varepsilon + \alpha\int_0^1 \rho^\varepsilon dx\right)\frac{\partial \rho^{\pm,\varepsilon}}{\partial x}\right\|_{L^\infty((0,T);L^2(\mathbb{T}))}\\
&\quad\leq C\left(\|\rho^\varepsilon\|_{L^\infty(\mathbb{T}\times(0,T))} + \|a\|_{L^\infty(0,T)}\right)\left\|\frac{\partial \rho^{\pm,\varepsilon}}{\partial x}\right\|_{L^\infty((0,T);L^2(\mathbb{T}))}.
\end{aligned}$$

We use Lemmas 4.1 and 4.2 and the Sobolev injections to deduce that there exists a constant $C\left(T, L_0, \alpha, \|a\|_{L^\infty(0,T)}, M_0, B_0\right)$ such that

$$\left\|f^\pm_{a^\varepsilon,\alpha}[\rho^\varepsilon]\right\|_{L^\infty((0,T);L^2(\mathbb{T}))} \leq C.$$

To conclude, we multiply the first and the second equations of the system (5) by $\frac{\partial \rho^{+,\varepsilon}}{\partial t}$ and $\frac{\partial \rho^{-,\varepsilon}}{\partial t}$, respectively, and we integrate in space. We deduce that for every $t \in (0,T)$ we have

$$\left\|\frac{\partial \rho^{\pm,\varepsilon}}{\partial t}(t)\right\|^2_{L^2(\mathbb{T})} + \frac{\varepsilon}{2}\frac{d}{dt}\left\|\frac{\partial \rho^{\pm,\varepsilon}}{\partial x}(t)\right\|^2_{L^2(\mathbb{T})} = \int_0^1 f^\pm_{a^\varepsilon,\alpha}[\rho^\varepsilon(t)]\frac{\partial \rho^{\pm,\varepsilon}}{\partial t}.$$

Integrating in time and using the fact that $\rho^{\pm,\varepsilon} \in C(\mathbb{R}\times[0,T))$ for all $T \geq 0$, we get

$$\begin{aligned}
&\left\|\frac{\partial \rho^{\pm,\varepsilon}}{\partial t}\right\|^2_{L^2(\mathbb{T}\times(0,T))} + \frac{\varepsilon}{2}\left\|\frac{\partial \rho^{\pm,\varepsilon}}{\partial x}(T)\right\|^2_{L^2(\mathbb{T})}\\
&\qquad = \int_0^T\int_0^1 f^\pm_{a^\varepsilon,\alpha}[\rho^\varepsilon(t)]\frac{\partial \rho^{\pm,\varepsilon}}{\partial t} + \frac{\varepsilon}{2}\left\|\frac{\partial {\rho_0}^{\pm,\varepsilon}}{\partial x}\right\|^2_{L^2(\mathbb{T})}.
\end{aligned}$$

We apply Hölder's inequality and the fact that $\varepsilon < 1$ and $ab \leq \frac{1}{2}(a^2+b^2)$ to obtain that

$$\begin{aligned}
&\left\| \frac{\partial \rho^{\pm,\varepsilon}}{\partial t} \right\|^2_{L^2(\mathbb{T}\times(0,T))} \\
&\quad \le \left\| f^{\pm}_{a^\varepsilon,\alpha}[\rho^\varepsilon] \right\|_{L^2(\mathbb{T}\times(0,T))} \left\| \frac{\partial \rho^{\pm,\varepsilon}}{\partial t} \right\|_{L^2(\mathbb{T}\times(0,T))} + \frac{1}{2} \left\| \frac{\partial {\rho_0}^{\pm,\varepsilon}}{\partial x} \right\|^2_{L^2(\mathbb{T})} \\
&\quad \le \frac{C}{2} \left(\left\| f^{\pm}_{a^\varepsilon,\alpha}[\rho^\varepsilon] \right\|^2_{L^2(\mathbb{T}\times(0,T))} + \left\| \frac{\partial \rho^{\pm,\varepsilon}}{\partial t} \right\|^2_{L^2(\mathbb{T}\times(0,T))} + \left\| \frac{\partial {\rho_0}^{\pm}}{\partial x} \right\|^2_{L^2(\mathbb{T})} \right),
\end{aligned}$$

which leads to

$$\begin{aligned}
\left\| \frac{\partial \rho^{\pm,\varepsilon}}{\partial t} \right\|^2_{L^2(\mathbb{T}\times(0,T))} &\le C \left(\left\| f^{\pm}_{a^\varepsilon,\alpha}[\rho^\varepsilon] \right\|^2_{L^2(\mathbb{T}\times(0,T))} + \left\| \frac{\partial {\rho_0}^{\pm}}{\partial x} \right\|^2_{L^2(\mathbb{T})} \right) \\
&\le C \left(T \left\| f^{\pm}_{a^\varepsilon,\alpha}[\rho^\varepsilon] \right\|^2_{L^\infty((0,T);L^2(\mathbb{T}))} + \left\| \frac{\partial {\rho_0}^{\pm}}{\partial x} \right\|^2_{L^2(\mathbb{T})} \right) \le C,
\end{aligned}$$

where $C\left(T, L_0, \alpha, \|a\|_{L^\infty(0,T)}, M_0, B_0\right)$. ∎

Remark 4.4 (the sense of the initial conditions). According to Lemma 4.3, we have $\rho^{\pm,\varepsilon,per} \in C([0,T), L^2(\mathbb{T}))$ uniformly in ε. This will give a sense to the limit of the initial conditions.

THEOREM 4.5 (long time existence). *Assume* (H1), (H2), *and* (H3); *for all* L_0, $T \ge 0$, $\alpha \in \mathbb{R}$, *and* $\rho_0^{\pm} \in H^1_{loc}(\mathbb{R})$, *the system* (5)–(6) *admits the solutions* $\rho^{\pm,\varepsilon} \in C^\infty(\mathbb{R}\times[0,T))$, *with* $\rho^{\pm,\varepsilon}(.,t)$ *verifying* (H1) *and* (H2). *Moreover, there exists a constant* $C\left(T, L_0, \alpha, \|a\|_{L^\infty(0,T)}, M_0, B_0\right)$ *independent of* ε, *with* B_0 *and* M_0 *defined in Lemmas* 4.1 *and* 4.2, *respectively, such that*

$$\left\|\rho^{\pm,\varepsilon,per}\right\|_{L^\infty((0,T);L^2(\mathbb{T}))} + \left\| \frac{\partial \rho^{\pm,\varepsilon}}{\partial x} \right\|_{L^\infty((0,T);L^2(\mathbb{T}))} + \left\| \frac{\partial \rho^{\pm,\varepsilon}}{\partial t} \right\|_{L^2(\mathbb{T}\times(0,T))} \le C, \tag{18}$$

where $\rho^{\pm,\varepsilon,per} = \rho^{\pm,\varepsilon} - L_0 x$.

Proof of Theorem 4.5. We are going to prove that local time solutions obtained by Corollary 3.3 can be extended to global time solutions for the same system.

We argue by contradiction: Assume that there exists a maximum time T_{max} such that we have the existence of solutions of the system (5)–(6) in the function space $C^\infty(\mathbb{R}\times[0,T_{max}))$.

For every $\delta > 0$, we consider the system (5) with the initial conditions

$$\rho^{\pm,\varepsilon}_{\delta,max} = \rho^{\pm,\varepsilon}(x, T_{max} - \delta).$$

We apply for the second time the same technique of Corollary 3.3 to deduce that there exists a time

$$T^\star_{\delta,max}(\|\rho^{\pm,\varepsilon,per}_{\delta,max}\|_{H^1(\mathbb{T})}, \|a\|_{L^\infty(0,T)}, L_0, \alpha, \varepsilon) > 0, \qquad \text{where } \rho^{\pm,\varepsilon,per}_{\delta,max} = \rho^{\pm,\varepsilon}_{\delta,max} - L_0 x,$$

such that the system (5)–(6) admits a solution defined until the time

$$T_0 = (T_{max} - \delta) + T^\star_{\delta,max}.$$

Moreover, according to Lemmas 4.1 and 4.2, we know that $\rho^{\pm,\varepsilon,per}_{\delta,max}$ are δ-uniformly bounded in $H^1(\mathbb{T})$. We use (16) to deduce that there exists a constant $C(\varepsilon, T_{max}, \alpha, \|a\|_{L^\infty(0,T)}, L_0) > 0$ independent of δ such that $T^\star_{\delta,max} \ge C > 0$; then $\lim_{\delta\to 0} T^\star_{\delta,max} \ge C > 0$, which implies that $T_0 > T_{max}$, and so we have a contradiction.

The estimation (18) is a consequence of Lemmas 4.1, 4.2, and 4.3. ∎

5. Existence and uniqueness of the solution of (1)–(2). In this section, we are going to prove that the system (1)–(2) admits a unique solution $\rho^{\pm}$ (in the distribution sense) which is the limit as $\varepsilon \to 0$ of $\rho^{\pm,\varepsilon}$ given by Theorem 4.5. In order to do that, we pass to the limit when ε tends to 0 in the system (7)–(8), and we use (18) in order to ensure the compactness. The proof of the uniqueness uses direct arguments.

Proof of Theorem 1.1. We first prove the existence and then establish the uniqueness.

Step 1 *(existence).* Let $\rho^{\pm,\varepsilon}$ be the solution of the system (5) given by Theorem 4.5. According to (18) we know that $\rho^{\pm,\varepsilon,per}$ are ε-uniformly bounded in $H^1(\mathbb{T}\times(0,T))$; then we can extract a subsequence that converges weakly in $H^1(\mathbb{T}\times(0,T))$. Knowing that $H^1(\mathbb{T}\times(0,T))$ is compact in $L^2(\mathbb{T}\times(0,T))$, this subsequence strongly converges in $L^2(\mathbb{T}\times(0,T))$. If we denote by $\rho^{\pm,per}$ the limit of this subsequence, we have to prove that $\rho^{\pm,per}+L_0x$ is a solution of the system (1)–(2) in the sense of distribution. Indeed, by Lemma 2.3, the term $\mp(L_0a^{\varepsilon})$ of (7) converges strongly to $(\mp L_0a)$ in $L^2(0,T)$.

The linear term

$$\mp\left(L_0C_\alpha[\rho^\varepsilon]+a^\varepsilon(t)\frac{\partial\rho}{\partial x}^{\pm,\varepsilon,per}\right)$$

of (7) weakly converges in $L^1(\mathbb{T}\times(0,T))$, and the reason is, on the one hand, that $\frac{\partial\rho}{\partial x}^{\pm,\varepsilon,per}$ are ε-uniformly bounded in $L^2(\mathbb{T}\times(0,T))$, which gives us the weak convergence in $L^2(\mathbb{T}\times(0,T))$, and, on the other hand, that a^ε strongly converges in $L^2(0,T)$. Then the linear term converges in the sense of distributions (i.e., in $\mathcal{D}'(\mathbb{T}\times(0,T))$). It remains to prove that the bilinear term

$$C_\alpha[\rho^\varepsilon]\frac{\partial\rho}{\partial x}^{\pm,\varepsilon,per}$$

of (7) also converges in the sense of distributions. We have the following:

1. The sequence $C_\alpha[\rho^\varepsilon]$ is compact in $L^2(\mathbb{T}\times(0,T))$.
2. The functions $\frac{\partial\rho}{\partial x}^{\pm,\varepsilon,per}$ are ε-uniformly bounded in $L^2(\mathbb{T}\times(0,T))$.

This gives us a strong convergence in $L^2(\mathbb{T}\times(0,T))$ times a weak convergence in $L^2(\mathbb{T}\times(0,T))$ and hence a weak convergence of the product in $L^1(\mathbb{T}\times(0,T))$. This leads, as a consequence, to the convergence in the distribution sense. This altogether shows that $\rho^{\pm,per}+L_0x$ is a solution in the sense of distribution of the system (1)–(2) and $\rho^{\pm,per}$ verifies estimate (18).

It remains to prove that the initial condition is satisfied by the limit function $\rho^{\pm,per}$. In fact, according to the estimate (18) on $\rho^{\pm,\varepsilon,per}$, $\frac{\partial\rho}{\partial t}^{\pm,\varepsilon}$, and $\frac{\partial\rho}{\partial x}^{\pm,\varepsilon}$, we see that $\rho^{\pm,\varepsilon,per}$ is ε-uniformly bounded in $H^1(\mathbb{T}\times(0,T))$.

From the fact that the injection of $H^1(\mathbb{T}\times(0,T))$ in $C([0,T];L^2(\mathbb{T}))$ is continuous and compact by classical arguments, we see that, for all $v\in L^2(\mathbb{T})$, the application $\gamma: U\longmapsto\int_0^1 U(0)v$ is a continuous linear form for $U\in C([0,T];L^2(\mathbb{T}))$ and hence $\gamma(\rho^{\pm,\varepsilon,per})\to\gamma(\rho^{\pm,per})$ as $\varepsilon\to 0$, because up to a subsequence $\rho^{\pm,\varepsilon,per}$ converges strongly in $C([0,T];L^2(\mathbb{T}))$. This altogether proves that the solution verifies the initial conditions (2).

Step 2 *(uniqueness).* Let $\rho_1^\pm$ and $\rho_2^\pm$ be two solutions of the system (1) such that $\rho_1^\pm(\cdot,0)=\rho_2^\pm(\cdot,0)=\rho_0^\pm$ and $\rho_i^\pm(\cdot,t)$ verify (H1), (H2), and estimate (18) for $i=1,2$, $t\in(0,T)$.

If we denote $\rho_i = \rho_i^+ - \rho_i^-$, $k_i = \rho_i^+ + \rho_i^-$ for $i = 1, 2$, then it is clear that $(\rho_1 - \rho_2)$ and $(k_1 - k_2)$ are 1-periodic functions in space and ρ_i, k_i verify the following system for $i = 1, 2$:

$$\left\{ \begin{aligned} \frac{\partial \rho_i}{\partial t} &= -\left(\rho_i + \alpha \int_0^1 \rho_i dx + a(t)\right) \frac{\partial k_i}{\partial x} \quad \text{in} \quad \mathcal{D}'(\mathbb{R} \times (0,T)), \\ \frac{\partial k_i}{\partial t} &= -\left(\rho_i + \alpha \int_0^1 \rho_i dx + a(t)\right) \frac{\partial \rho_i}{\partial x} \quad \text{in} \quad \mathcal{D}'(\mathbb{R} \times (0,T)). \end{aligned} \right. \tag{19}$$

We substract the two systems to obtain that

$$\left\{ \begin{aligned} \frac{\partial(\rho_1 - \rho_2)}{\partial t} &= -\left(\rho_1 + \alpha \int_0^1 \rho_1 dx\right) \frac{\partial k_1}{\partial x} + \left(\rho_2 + \alpha \int_0^1 \rho_2 dx\right) \frac{\partial k_2}{\partial x} \\ &\quad - a(t) \frac{\partial(k_1 - k_2)}{\partial x}, \\ \frac{\partial(k_1 - k_2)}{\partial t} &= -\left(\rho_1 + \alpha \int_0^1 \rho_1 dx\right) \frac{\partial \rho_1}{\partial x} + \left(\rho_2 + \alpha \int_0^1 \rho_2 dx\right) \frac{\partial \rho_2}{\partial x} \\ &\quad - a(t) \frac{\partial(\rho_1 - \rho_2)}{\partial x}. \end{aligned} \right.$$

The previous system is equivalent to

$$\left\{ \begin{aligned} \frac{\partial(\rho_1 - \rho_2)}{\partial t} &= -\left((\rho_1 - \rho_2) + \alpha \int_0^1 (\rho_1 - \rho_2) dx\right) \frac{\partial k_1}{\partial x} \\ &\quad - \left(\rho_2 + \alpha \int_0^1 \rho_2 dx\right) \frac{\partial(k_1 - k_2)}{\partial x} - a(t) \frac{\partial(k_1 - k_2)}{\partial x}, \\ \frac{\partial(k_1 - k_2)}{\partial t} &= -\left((\rho_1 - \rho_2) + \alpha \int_0^1 (\rho_1 - \rho_2) dx\right) \frac{\partial \rho_1}{\partial x} \\ &\quad - \left(\rho_2 + \alpha \int_0^1 \rho_2 dx\right) \frac{\partial(\rho_1 - \rho_2)}{\partial x} - a(t) \frac{\partial(\rho_1 - \rho_2)}{\partial x}. \end{aligned} \right.$$

We multiply the first equation of this system by $(\rho_1 - \rho_2)$ and integrate in space to obtain, for almost every t, that

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|(\rho_1 - \rho_2)(t)\|^2_{L^2(\mathbb{T})} = \; & -\int_0^1 \left((\rho_1 - \rho_2)^2 \frac{\partial k_1}{\partial x}\right) \\ & -\alpha \left(\int_0^1 (\rho_1 - \rho_2)\right) \int_0^1 \left((\rho_1 - \rho_2) \frac{\partial k_1}{\partial x}\right) \\ & -\int_0^1 \left((\rho_1 - \rho_2) \left(\rho_2 + \alpha \int_0^1 \rho_2\right) \frac{\partial(k_1 - k_2)}{\partial x}\right) \\ & -a(t) \int_0^1 \left((\rho_1 - \rho_2) \frac{\partial(k_1 - k_2)}{\partial x}\right). \end{aligned}$$

Similarly, we multiply the second equation by $(k_1 - k_2)$ and integrate in space to get, for almost every time t,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|(k_1-k_2)(t)\|^2_{L^2(\mathbb{T})} = \; & -\int_0^1\left((\rho_1-\rho_2)(k_1-k_2)\frac{\partial\rho_1}{\partial x}\right)\\ & -\alpha\left(\int_0^1(\rho_1-\rho_2)\right)\int_0^1(k_1-k_2)\frac{\partial\rho_1}{\partial x}\\ & -\int_0^1\left((k_1-k_2)\left(\rho_2+\alpha\int_0^1\rho_2\right)\frac{\partial(\rho_1-\rho_2)}{\partial x}\right)\\ & -a(t)\int_0^1\left((k_1-k_2)\frac{\partial(\rho_1-\rho_2)}{\partial x}\right).\end{aligned}$$

We add the two previous equations to obtain, for almost every time t,

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\left(\|(\rho_1-\rho_2)(t)\|^2_{L^2(\mathbb{T})}+\|(k_1-k_2)(t)\|^2_{L^2(\mathbb{T})}\right)\\ &\quad=-\int_0^1\left((\rho_1-\rho_2)^2\frac{\partial k_1}{\partial x}\right)-\alpha\left(\int_0^1(\rho_1-\rho_2)\right)\int_0^1\left((\rho_1-\rho_2)\frac{\partial k_1}{\partial x}\right)\\ &\qquad-\alpha\left(\int_0^1(\rho_1-\rho_2)\right)\int_0^1\left((k_1-k_2)\frac{\partial\rho_1}{\partial x}\right)\\ &\qquad-\int_0^1\left(\frac{\partial}{\partial x}\left((\rho_1-\rho_2)(k_1-k_2)\left(\rho_2+\alpha\int_0^1\rho_2\right)\right)\right)\\ &\qquad-\int_0^1\left((\rho_1-\rho_2)(k_1-k_2)\frac{\partial(\rho_1-\rho_2)}{\partial x}\right)-a(t)\int_0^1\left(\frac{\partial}{\partial x}\left((\rho_1-\rho_2)(k_1-k_2)\right)\right).\end{aligned}$$

From the fact that ρ_i, $i=1,2$, and (k_1-k_2) are 1-periodic functions in space, the previous equation becomes

$$\begin{aligned}&=\overbrace{-\int_0^1\left((\rho_1-\rho_2)^2\frac{\partial k_1}{\partial x}\right)}^{I_1}\;\overbrace{-\alpha\left(\int_0^1(\rho_1-\rho_2)\right)\int_0^1\left((\rho_1-\rho_2)\frac{\partial k_1}{\partial x}\right)}^{I_2}\\ &\overbrace{-\alpha\left(\int_0^1(\rho_1-\rho_2)\right)\int_0^1\left((k_1-k_2)\frac{\partial\rho_1}{\partial x}\right)}^{I_3}\;\overbrace{-\int_0^1\left((\rho_1-\rho_2)(k_1-k_2)\frac{\partial(\rho_1-\rho_2)}{\partial x}\right)}^{I_4}.\end{aligned}$$

And since $\frac{\partial k_i}{\partial x}\geq 0$ for $i=1,2$, we know that

$$\begin{aligned}I_1+I_4 &= -\int_0^1\left((\rho_1-\rho_2)^2\frac{\partial k_1}{\partial x}\right)-\frac{1}{2}\int_0^1\left((k_1-k_2)\frac{\partial}{\partial x}\left((\rho_1-\rho_2)^2\right)\right)\\ &= -\int_0^1\left((\rho_1-\rho_2)^2\frac{\partial k_1}{\partial x}\right)+\frac{1}{2}\int_0^1\left((\rho_1-\rho_2)^2\frac{\partial(k_1-k_2)}{\partial x}\right)\\ &= -\frac{1}{2}\int_0^1\left((\rho_1-\rho_2)^2\frac{\partial(k_1+k_2)}{\partial x}\right)\leq 0.\end{aligned}$$

Moreover, from (18), we have, for almost every t,

$$\begin{aligned} I_2 &\leq |\alpha| \|(\rho_1 - \rho_2)(t)\|_{L^2(\mathbb{T})} \|(\rho_1 - \rho_2)(t)\|_{L^2(\mathbb{T})} \left\| \frac{\partial k_1}{\partial x}(t) \right\|_{L^2(\mathbb{T})} \\ &\leq C \|(\rho_1 - \rho_2)(t)\|^2_{L^2(\mathbb{T})}. \end{aligned}$$

Similarly, from (18), we have, for almost every t,

$$\begin{aligned} I_3 &\leq |\alpha| \|(\rho_1 - \rho_2)(t)\|_{L^2(\mathbb{T})} \|(k_1 - k_2)(t)\|_{L^2(\mathbb{T})} \left\| \frac{\partial \rho_1}{\partial x}(t) \right\|_{L^2(\mathbb{T})} \\ &\leq C \left(\|(\rho_1 - \rho_2)(t)\|^2_{L^2(\mathbb{T})} + \|(k_1 - k_2)(t)\|^2_{L^2(\mathbb{T})} \right). \end{aligned}$$

Then

$$\begin{aligned} &\frac{d}{dt} \left(\|(\rho_1 - \rho_2)(t)\|^2_{L^2(\mathbb{T})} + \|(k_1 - k_2)(t)\|^2_{L^2(\mathbb{T})} \right) \\ &\qquad \leq C \left(\|(\rho_1 - \rho_2)(t)\|^2_{L^2(\mathbb{T})} + \|(k_1 - k_2)(t)\|^2_{L^2(\mathbb{T})} \right). \end{aligned}$$

Now we integrate in time and use the fact that $\rho_i,\, k_i \in C([0,T), L^2_{loc}(\mathbb{R}))$, $\rho_1(\cdot,0) = \rho_2(\cdot,0)$, and $k_1(\cdot,0) = k_2(\cdot,0)$ to obtain that

$$\sup_{t\in(0,T)} \|(\rho_1 - \rho_2)(t)\|^2_{L^2(\mathbb{T})} + \sup_{t\in(0,T)} \|(k_1 - k_2)(t)\|^2_{L^2(\mathbb{T})} \leq 0.$$

This achieves the proof of uniqueness. $\square$

Remark 5.1. In Theorem 1.1, we have proved a result of existence and uniqueness in $H^1_{loc}(\mathbb{R}\times[0,T))$ depending on some uniform estimates in this space. These estimates give a sufficient compactness in order to ensure the passage to the limit as ε tends to 0 in the bilinear term. However, the space $W^{1,1}_{loc}(\mathbb{R}\times[0,T))$ does not give enough compactness. On the other hand, the space of functions $L^2_{loc}(\mathbb{R}\times[0,T))$ having their derivatives in $L^\infty((0,T); \left(L^1 \log L^1\right)_{loc}(\mathbb{R}))$ requires the minimal properties to ensure the passage to the limit in the bilinear term. The result of existence in this space will be the core of a paper in preparation.

6. Further properties: Comparison principle with case $\boldsymbol{\alpha = 0}$. In this section, we are going to prove a comparison principle result of the system (1) in the case $\alpha = 0$ (i.e., Theorem 1.2). In order to do this, first we prove in the following subsection the same result for the approximate system (5). Then we give the proof of Theorem 1.2.

6.1. Comparison principle for the regularized system with case $\boldsymbol{\alpha = 0}$.

LEMMA 6.1 (comparison principle). *Let $a(\cdot)$ satisfy* (H3) *and $\rho_1^{\pm,\varepsilon}$, $\rho_2^{\pm,\varepsilon} \in C^\infty(\mathbb{R}\times[0,T))$ be two solutions of the system* (5) *with $\alpha = 0$. Moreover, let $\rho_1^{\pm,\varepsilon}(.,t)$, $\rho_2^{\pm,\varepsilon}(.,t)$ verify* (H1) *and* (H2) *for all $t \in [0,T)$. Then, if $\rho_1^{\pm,\varepsilon}(\cdot,0) \leq \rho_2^{\pm,\varepsilon}(\cdot,0)$ in $\mathbb{R}$, we have $\rho_1^{\pm,\varepsilon} \leq \rho_2^{\pm,\varepsilon}$ on $\mathbb{R}\times[0,T)$.*

Proof of Lemma 6.1. We know that $\rho_1^{\pm,\varepsilon}$ and $\rho_2^{\pm,\varepsilon}$ verify the following systems:

$$\begin{cases} \dfrac{\partial \rho_1^{+,\varepsilon}}{\partial t} - \varepsilon \dfrac{\partial^2 \rho_1^{+,\varepsilon}}{\partial x^2} = & -\left(\rho_1^{+,\varepsilon} - \rho_1^{-,\varepsilon} + a^\varepsilon(t)\right) \dfrac{\partial \rho_1^{+,\varepsilon}}{\partial x} \quad \text{in } \mathcal{D}'(\mathbb{R}\times(0,T)),\\ \dfrac{\partial \rho_1^{-,\varepsilon}}{\partial t} - \varepsilon \dfrac{\partial^2 \rho_1^{-,\varepsilon}}{\partial x^2} = & \left(\rho_1^{+,\varepsilon} - \rho_1^{-,\varepsilon} + a^\varepsilon(t)\right) \dfrac{\partial \rho_1^{-,\varepsilon}}{\partial x} \quad \text{in } \mathcal{D}'(\mathbb{R}\times(0,T)),\end{cases}$$

$$\begin{cases} \dfrac{\partial \rho_2^{+,\varepsilon}}{\partial t} - \varepsilon \dfrac{\partial^2 \rho_2^{+,\varepsilon}}{\partial x^2} = & -\left(\rho_2^{+,\varepsilon} - \rho_2^{-,\varepsilon} + a^\varepsilon(t)\right) \dfrac{\partial \rho_2^{+,\varepsilon}}{\partial x} \quad \text{in } \mathcal{D}'(\mathbb{R}\times(0,T)),\\ \dfrac{\partial \rho_2^{-,\varepsilon}}{\partial t} - \varepsilon \dfrac{\partial^2 \rho_2^{-,\varepsilon}}{\partial x^2} = & \left(\rho_2^{+,\varepsilon} - \rho_2^{-,\varepsilon} + a^\varepsilon(t)\right) \dfrac{\partial \rho_2^{-,\varepsilon}}{\partial x} \quad \text{in } \mathcal{D}'(\mathbb{R}\times(0,T)),\end{cases}$$

respectively.

If we denote $w^{\pm,\varepsilon}$ by $\tilde{\rho}_2^{\pm,\varepsilon} - \tilde{\rho}_1^{\pm,\varepsilon}$, where

$$\tilde{\rho}_2^{\pm,\varepsilon} = \rho_2^{\pm,\varepsilon} e^{-\gamma t} \quad \text{and} \quad \tilde{\rho}_1^{\pm,\varepsilon} = \rho_1^{\pm,\varepsilon} e^{-\gamma t} \quad \text{with} \quad \gamma > 0,$$

we can easily check that $w^{\pm,\varepsilon}$ are solutions of the following system:

$$\begin{cases} \dfrac{\partial w^{+,\varepsilon}}{\partial t} - \varepsilon \dfrac{\partial^2 w^{+,\varepsilon}}{\partial x^2} + \gamma w^{+,\varepsilon} & = -\, e^{\gamma t}\left(w^{+,\varepsilon} - w^{-,\varepsilon}\right) \dfrac{\partial \tilde{\rho}_2^{+,\varepsilon}}{\partial x} \\ & \quad - e^{\gamma t}\left(\tilde{\rho}_1^{+,\varepsilon} - \tilde{\rho}_1^{-,\varepsilon} + e^{-\gamma t} a^\varepsilon(t)\right) \dfrac{\partial w^{+,\varepsilon}}{\partial x}, \\ \dfrac{\partial w^{-,\varepsilon}}{\partial t} - \varepsilon \dfrac{\partial^2 w^{-,\varepsilon}}{\partial x^2} + \gamma w^{-,\varepsilon} & = \quad e^{\gamma t}\left(w^{+,\varepsilon} - w^{-,\varepsilon}\right) \dfrac{\partial \tilde{\rho}_2^{-,\varepsilon}}{\partial x} \\ & \quad + e^{\gamma t}\left(\tilde{\rho}_1^{+,\varepsilon} - \tilde{\rho}_1^{-,\varepsilon} + e^{-\gamma t} a^\varepsilon(t)\right) \dfrac{\partial w^{-,\varepsilon}}{\partial x}. \end{cases} \tag{20}$$

We are interested in the $\min_{(k,x,t)\in\{+,-\}\times\mathbb{T}\times(0,T)}(w^{k,\varepsilon}(x,t))$. Our result follows if we can prove that this minimum is positive. However, this minimum is attained at a point $(k_0, x_0, t_0) \in \{+,-\}\times\mathbb{T}\times[0,T]$ (because $w^{+,\varepsilon}$ and $w^{-,\varepsilon}$ are $C^\infty(\mathbb{T}\times(0,T))$). Two cases may occur:

1. Case $t_0 = 0$. We have

$$\min_{(k,x,t)\in\{+,-\}\times\mathbb{T}\times(0,T)}(w^{k,\varepsilon}(x,t))$$
$$= w^{k_0,\varepsilon}(x_0,t_0) = \left(\rho_2^{k_0,\varepsilon}(x_0,0) - \rho_1^{k_0,\varepsilon}(x_0,0)\right) e^{-\gamma t_0} \geq 0;$$

   and we are done.

2. Case $t_0 \in (0,T]$. We have that (k_0,x_0,t_0) is a minimum point; then

$$\frac{\partial^2 w^{k_0,\varepsilon}}{\partial x^2}(x_0,t_0) \geq 0, \tag{21}$$

$$\frac{\partial w^{k_0,\varepsilon}}{\partial t}(x_0,t_0) \leq 0, \tag{22}$$

$$\frac{\partial w^{k_0,\varepsilon}}{\partial x}(x_0,t_0) = 0. \tag{23}$$

Combining (21), (22), (23) and taking into consideration that $w^{\pm,\varepsilon}$ verifies the system (20), we obtain that

$$\begin{aligned}\gamma w^{k_0,\varepsilon}(x_0,t_0) &\geq e^{\gamma t_0} sign(w^{+,\varepsilon}(x_0,t_0) - w^{-,\varepsilon}(x_0,t_0))(w^{+,\varepsilon}(x_0,t_0) \\ &\quad - w^{-,\varepsilon}(x_0,t_0))\frac{\partial \tilde{\rho}_2^{k_0,\varepsilon}}{\partial x} \\ &\geq e^{\gamma t_0}|w^{+,\varepsilon}(x_0,t_0) - w^{-,\varepsilon}(x_0,t_0)|\frac{\partial \tilde{\rho}_2^{k_0,\varepsilon}}{\partial x} \geq 0.\end{aligned}$$

Then $\tilde{\rho}_1^{\pm,\varepsilon} \leq \tilde{\rho}_2^{\pm,\varepsilon}$ in $\mathbb{R} \times (0,T)$, which gives $\rho_1^{\pm,\varepsilon} \leq \rho_2^{\pm,\varepsilon}$. ◻

We now give the proof of Theorem 1.2.

6.2. Proof of Theorem 1.2. Let

$$\rho_1^{\pm}(x,0) = \rho_{1,0}^{\pm}(x) = \rho_{1,0}^{\pm,per}(x) + L_0 x \quad \text{and} \quad \rho_2^{\pm}(x,0) = \rho_{2,0}^{\pm}(x) = \rho_{2,0}^{\pm,per}(x) + L_0 x.$$

If we denote

$$\rho_{1,0}^{\pm,\varepsilon}(x) = \rho_{1,0}^{\pm,per} * \eta_\varepsilon(x) + L_0 x \quad \text{and} \quad \rho_{2,0}^{\pm,\varepsilon}(x) = \rho_{2,0}^{\pm,per} * \eta_\varepsilon(x) + L_0 x,$$

where η_ε is a regularization sequence, we can easily check that $\rho_{1,0}^{\pm,\varepsilon} \leq \rho_{2,0}^{\pm,\varepsilon}$.

Moreover, according to the uniqueness of the solution, we know that there exist $\rho_1^{\pm,\varepsilon}$, $\rho_2^{\pm,\varepsilon} \in C^\infty(\mathbb{R} \times [0,T))$, verifying (H2) for all $t \in (0,T)$, which are solutions of the system (5), such that

$$\rho_1^{\pm} = \lim_{\varepsilon\to 0} \rho_1^{\pm,\varepsilon}, \quad \rho_2^{\pm} = \lim_{\varepsilon\to 0} \rho_2^{\pm,\varepsilon},$$

$$\rho_1^{\pm,\varepsilon}(x,0) = \rho_{1,0}^{\pm,\varepsilon}(x) \quad \text{and} \quad \rho_2^{\pm,\varepsilon}(x,0) = \rho_{2,0}^{\pm,\varepsilon}(x).$$

We apply Lemma 6.1 to obtain that $\rho_1^{\pm,\varepsilon} \leq \rho_2^{\pm,\varepsilon}$. We pass to the limit as $\varepsilon \to 0$ to deduce that $\rho_1^{\pm} \leq \rho_2^{\pm}$ a.e. in $\mathbb{R} \times (0,T)$. ◻

Remark 6.2. Thanks to this comparison result, we proved in a previous paper [8] the existence and the uniqueness of a solution (in the viscosity sense). Here this comparison result is an indirect explanation of our estimates obtained in Lemmas 4.1, 4.2, and 4.3 that have ensured our principal theorem, Theorem 1.1.

7. Application in the case of the classical Burgers equation. In this paragraph we are going to prove that this technique can be also applied to the classical Burgers equation, even in the frame of functions in $W^{1,p}_{loc}(\mathbb{R} \times (0,T))$ for all $1 \leq p < +\infty$, constituting the proof of Theorem 1.4.

Proof of Theorem 1.4. First, we remark that the existence of solution to the regularized problem can be done thanks to the continuous injection $W^{1,p}(\mathbb{T})$ in $L^\infty(\mathbb{T})$. Now, for the proof of this theorem, it suffices to show an estimation over the space derivatives of the solution (i.e., a result similar to that of Lemma 4.1).

First, we put ourselves in the hypothesis of Lemma 4.1. We derive the equation (4) with respect to x, then we multiply it by $(\frac{\partial u}{\partial x})^{p-1}$, and, finally, we integrate over

$(0,1)$; since u verifies (H2), we obtain that

$$\begin{aligned}\frac{1}{p}\frac{d}{dt}\left\|\frac{\partial u}{\partial x}(t)\right\|^p_{L^p(\mathbb{T})} &= -\int_0^1 f''(u)\frac{\partial u}{\partial x}\left(\frac{\partial u}{\partial x}\right)^p - \int_0^1 f'(u)\frac{\partial^2 u}{\partial x^2}\left(\frac{\partial u}{\partial x}\right)^{p-1}\\ &= -\int_0^1 \frac{\partial (f'(u))}{\partial x}\left(\frac{\partial u}{\partial x}\right)^p - \frac{1}{p}\int_0^1 f'(u)\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right)^p\\ &= -\frac{1}{p}\int_0^1 \frac{\partial}{\partial x}\left(f'(u)\left(\frac{\partial u}{\partial x}\right)^p\right) - \left(1-\frac{1}{p}\right)\int_0^1 f''(u)\frac{\partial u}{\partial x}\left(\frac{\partial u}{\partial x}\right)^p \le 0,\end{aligned}$$

because f is convex, u verifies (H2), and $p \ge 1$. To terminate the demonstration, we follow the same steps of the proof of Theorem 1.1. We remark that here we do not need the L^2 bound over the solution and also the compactness in the passage to the limit, because (4) is in the conservative form, which was not the case of our study.

Acknowledgments. We would like to thank M. Cannone, R. Monneau, and F. Ribaud for fruitful discussions in the preparation of this paper. We thank also N. Forcadel and H. Ibrahim for the second reading of this paper.

REFERENCES

[1] O. Alvarez, P. Cardaliaguet, and R. Monneau, *Existence and uniqueness for dislocation dynamics with nonnegative velocity*, Interfaces Free Bound., 7 (2005), pp. 415–434.
[2] O. Alvarez, P. Hoch, Y. Le Bouar, and R. Monneau, *Dislocation dynamics: Short-time existence and uniqueness of the solution*, Arch. Ration. Mech. Anal., 181 (2006), pp. 449–504.
[3] L. Ambrosio, *Transport equation and Cauchy problem for BV vector fields*, Invent. Math., 158 (2004), pp. 227–260.
[4] G. Barles and O. Ley, *Nonlocal first-order Hamilton-Jacobi equations modelling dislocations dynamics*, Comm. Partial Differential Equations, 31 (2006), pp. 1191–1208.
[5] M. Cannone, *Ondelettes, paraproduits et Navier-Stokes*, Diderot Editeur, Paris, 1995.
[6] G.-Q. Chen and D. Wang, *The Cauchy problem for the Euler equations for compressible fluids*, in Handbook of Mathematical Fluid Dynamics, Vol. I, S. Friedlander and D. Serre, eds., North–Holland, Amsterdam, 2002, pp. 421–543.
[7] R. J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.
[8] A. El Hajj and N. Forcadel, *A convergent scheme for a non-local coupled system modelling dislocations densities dynamics*, Math. Comp., to appear.
[9] N. Forcadel, *Dislocations Dynamics with a Mean Curvature Term: Short Time Existence and Uniqueness*, preprint, 2005.
[10] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Classics Math., Springer-Verlag, Berlin, 2001.
[11] I. Groma and P. Balogh, *Investigation of dislocation pattern formation in a two-dimensional self-consistent field approximation*, Acta Mater., 47 (1999), pp. 3647–3654.
[12] J. P. Hirth and J. Lothe, *Theory of Dislocations*, 2nd ed., Krieger, Malabar, FL, 1992.
[13] P. LeFloch, *Entropy weak solutions to nonlinear hyperbolic systems under nonconservative form*, Comm. Partial Differential Equations, 13 (1988), pp. 669–727.
[14] P. LeFloch and T.-P. Liu, *Existence theory for nonlinear hyperbolic systems in nonconservative form*, Forum Math., 5 (1993), pp. 261–280.
[15] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Travaux et Recherches Mathématiques 17, Dunod, Paris, 1968.
[16] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.
[17] D. Serre, *Systems of Conservation Laws.* I, Cambridge University Press, Cambridge, UK, 1999.
[18] D. Serre, *Systems of Conservation Laws.* II, Cambridge University Press, Cambridge, UK, 2000.

SIAM J. MATH. ANAL.
Vol. 39, No. 3, pp. 987–1017

HOMOGENIZATION OF A DOUBLY NONLINEAR STEFAN-TYPE PROBLEM*

AUGUSTO VISINTIN†

Abstract. Temperature and phase evolution in phase transitions are represented here by coupling the energy balance equation with a multivalued constitutive relation between the density of internal energy and the temperature, and with a nonlinear conduction law. This doubly nonlinear problem generalizes the classical Stefan model. Existence of a weak solution is proved via time discretization, a priori estimates, and passage to the limit. A medium exhibiting periodic oscillations in space is then considered; as the oscillation period vanishes, *two-scale convergence* (in the sense of Nguetseng) to a corresponding two-scale homogenized problem is proved. The latter is shown to be equivalent to a coarse-scale model. The cases of Fourier's law with either temperature- or phase-dependent conductivity are also treated.

Key words. phase transitions, Stefan problem, homogenization, two-scale convergence

AMS subject classifications. 35K60, 35R35, 78M40, 80A22

DOI. 10.1137/060676647

1. Introduction.

This paper deals with the two- and single-scale homogenization of a class of quasi-linear parabolic problems that includes Stefan-type models of phase transitions.

The model. Let us consider a (possibly inhomogeneous and anisotropic) incompressible material capable of attaining two phases, say solid and liquid, that occupies a three-dimensional domain Ω. Let us denote the temperature by u, the density of internal energy (or enthalpy at fixed pressure) by w, the heat flux by $\vec{q}$, and the intensity of a heat source or sink by f.

Let $\varphi : \mathbf{R} \times \Omega \to \mathbf{R} \cup \{+\infty\}$ and $\vec{\alpha} : \mathbf{R}^3 \times \mathbf{R} \times \Omega \times]0, T[\to \mathbf{R}^3$, and assume that φ is convex w.r.t. the first variable and that $\vec{\alpha}(\vec{\xi}, u, x, t)$ is continuous w.r.t. the pair $(\vec{\xi}, u)$ and nondecreasing w.r.t. $\vec{\xi}$. We fix any $T > 0$, set $\Omega_T := \Omega \times]0, T[$, and provide a weak formulation for an initial- and boundary-value problem for the following nonlinear system:

$$\frac{\partial w}{\partial t} + \nabla \cdot \vec{q} = f(u, \nabla u) \qquad \text{in } \mathcal{D}'(\Omega_T)\ (\nabla \cdot := \mathrm{div}), \tag{1.1}$$

$$w \in \partial\varphi(u, x) \qquad \text{a.e. in } \Omega_T, \tag{1.2}$$

$$\vec{q} = \vec{\alpha}(-\nabla u, u, x, t) \qquad \text{a.e. in } \Omega_T \tag{1.3}$$

(by $\partial\varphi(\cdot, x)$ we denote the subdifferential of $\varphi(\cdot, x)$) or, more synthetically,

$$\frac{\partial}{\partial t}\partial\varphi(u, x) + \nabla \cdot \vec{\alpha}(-\nabla u, u, x, t) \ni f(u, \nabla u) \qquad \text{in } \mathcal{D}'(\Omega_T). \tag{1.4}$$

The occurrence of a double nonlinearity in the principal part of the equation may be noticed. An initial condition for w and boundary conditions for either u or the normal

*Received by the editors December 4, 2006; accepted for publication (in revised form) April 13, 2007; published electronically October 19, 2007. This research was partially supported by the project "Free boundary problems, phase transitions and models of hysteresis" of Italian M.U.R.
http://www.siam.org/journals/sima/39-3/67664.html

†Università degli Studi di Trento, Dipartimento di Matematica, via Sommarive 14, 38050 Povo di Trento, Italy (Visintin@science.unitn.it).

component of $\vec{q}$ may be appended to this system. Equation (1.1) represents the energy balance, whereas (1.2) and (1.3) are constitutive relations. The explicit dependence on x of the constitutive functions φ and $\vec{\alpha}$ may account for either a mixture of different materials or a single material whose behavior varies in space; in the first case these functions will be discontinuous w.r.t. x, whereas in the latter case they need not be so.

A free boundary problem. The discontinuity of the dependence of w on u may account for a simplified model of phase transitions, e.g., in a solid-liquid system, in which density variations and convection are neglected. For instance let us assume that (1.2) is of the form $w \in \int_0^u C(\xi, x)\, d\xi + L(x)H(u - u_*(x))$ a.e. in Ω_T, with $C = C(u,x)$ being the temperature- and space-dependent specific heat, L the space-dependent density of latent heat of phase transition, H the Heaviside function (i.e., $H(v) = 0$ for $v \le 0$, $H(v) = 1$ for $v > 0$), and u_* the space-dependent temperature of phase equilibrium. Equation (1.1) may then account for the energy balance in a solid-liquid system, and the space-time domains

$$\Omega_T^+ := \{(x,t) \in \Omega_T : u > u_*(x)\}, \qquad \Omega_T^- := \{(x,t) \in \Omega_T : u < u_*(x)\}$$

represent, respectively, the evolution of the liquid and solid phases. The set $\Omega_{T*} := \{(x,t) \in \Omega_T : u(x,t) = u_*(x)\}$ may consist of one or more moving surfaces that represent the space-time interface $\mathcal{S}$ between the phases; but Ω_{T*} may also comprise a component with nonempty interior that would represent a fine mixture of liquid and solid and is known as a *mushy region*.

In general for an inhomogeneous material the interface $\mathcal{S}$ need not be smooth even in the absence of a mushy region. For instance a connected component of $\mathcal{S}_t := \mathcal{S} \cap (\Omega \times \{t\})$ may end at the boundary between two grains of materials that are characterized by different temperatures of phase transition. If $\partial\varphi$ is as above, in the neighborhood U of any regular point $(x,t) \in \mathcal{S}$, (1.1) is equivalent to the following bulk and interface conditions:

$$C(u,x)\frac{\partial u}{\partial t} + \nabla \cdot \vec{q} = f \qquad \text{in } U \setminus \mathcal{S}, \tag{1.5}$$

$$u = u_*(x) \qquad \text{on } U \cap \mathcal{S}, \tag{1.6}$$

$$\vec{q}_1 \cdot \vec{\nu} - \vec{q}_2 \cdot \vec{\nu} = L(x)\vec{v} \cdot \vec{\nu} \qquad \text{on } U \cap \mathcal{S}. \tag{1.7}$$

Here by $\vec{q}_1$ we denote the heat flux contributed by the liquid phase, by $\vec{q}_2$ that absorbed by the solid phase, by $\vec{v}$ the interface velocity, and by $\vec{\nu}$ a normal vector field to $\mathcal{S}_t$.

Equation (1.3) includes the nonlinear Fourier conduction law

(1.8)
$$\vec{q} = -K(u,x) \cdot \nabla u \quad \text{in } \Omega_T \ \Big(\text{i.e., } \ q_i = -\textstyle\sum_{j=1,2,3} K_{ij}(u,x)\partial u/\partial x_j \ \text{ for } i = 1,2,3\Big);$$

here K is a symmetric and positive-definite tensor function and represents the heat conductivity. For an isotropic material $K(u,x) = k(u,x)I$ (I being the identity tensor), and (1.7) is reduced to the classical *Stefan condition*

$$k(u,x)\frac{\partial u_+}{\partial \nu} - k(u,x)\frac{\partial u_-}{\partial \nu} = -L(x)\vec{v} \cdot \vec{\nu} \qquad \text{on } U \cap \mathcal{S}, \tag{1.9}$$

where by $\partial u_\pm/\partial \nu$ we denote the limit of the normal derivative of u taken from the set $\Omega_T^\pm$. The system (1.5)–(1.8) is a (local) strong formulation of the classical two-phase Stefan problem in several space dimensions. This is a *free boundary problem*, for the evolution of the interface between the phases is unknown. Here we shall be concerned with the more general system (1.1)–(1.3). From the physical point of view, the fact that the conduction law (1.3) allows for the dependence of the heat flux on the temperature seems more relevant than the nonlinearity of this equation.

Existence of a solution. We formulate an initial- and boundary-value problem for the system (1.1)–(1.3) in the framework of Sobolev spaces and prove existence of a weak solution via approximation by time discretization, derivation of a priori estimates, and passage to the limit.

Some remarks about this argument are in order. Here the standard estimate procedure based on multiplying the approximate equation by the time derivative of the approximate temperature seems to hardly be applicable, because of the dependence of the conductivity on the temperature. Therefore it is not evident that a uniform L^p-estimate might be derived for the time derivative of the approximating temperature if we exclude the special cases in which the Kirchhoff transformation can be applied (see section 6). However, here we prove the strong convergence of the approximate temperature in $L^2(\Omega_T)$ via *compactness by strict convexity*—namely the property that (freely speaking) weak convergence in L^p_{loc} ($p > 1$) joint with convergence of a strictly convex potential entail strong convergence in L^q_{loc} for any $q < p$; see, e.g., [77], [78, Chap. X]. Along the lines of [4], one might also derive strong L^2-convergence by multiplying the approximate equation by the time increment (rather than the time-incremental ratio) of the approximate temperature.

We deal with a heat conductivity that depends continuously on the temperature and (possibly nonstrictly) monotonically on the temperature gradient. In case of a degenerate parabolic part (e.g., for φ identically constant), we would then miss the strong convergence of the temperature; however, if (1.3) were independent of the temperature, our results might easily be extended to this case, too. Via the classical Kirchhoff transformation we also treat the case of phase-dependent conductivity under a structure hypothesis on the conductivity itself.

Homogenization. The main concern of this work is the homogenization of composite materials in which the constitutive functions φ and $\vec{\alpha}$ are discontinuous functions of the space variable x. By a classical procedure, we assume that φ and $\vec{\alpha}$ depend not only on the coarse-scale variable x but also on a fine-scale variable x/ε (ε being a *small* scalar parameter), and we assume that the latter dependence if periodic. We then let ε vanish and show that the solution of the ε-dependent problem *two-scale converges* (in the sense of Nguetseng [65] and Allaire [2]) to a solution of a two-scale problem in which two further fields $u_\sharp$ and $\vec{q}_\sharp$ occur besides the coarse-scale fields u and $\vec{q}$. Our argument is based on two-scale techniques somehow analogous to those that we used for the existence result. (The analogy between these methods seems to confirm the potentialities of an approach based on two-scale convergence.)

We then retrieve a purely coarse-scale problem of the form (1.1)–(1.3) (so-called *upscaling*) with different constitutive functions φ and $\vec{\alpha}$. These functions are determined via the solution of a family of nonlinear *cell problems*; in the case of the linear Fourier law, we also construct a solution of these problems along the lines of a well-known procedure.

We also show the inverse statement: any solution of the coarse-scale problem may be retrieved from the two-scale model (this procedure might be referred to as *downscaling*). The two- and single-scale formulations are thus equivalent, although the latter provides a somehow more synthetic picture of the phenomenon. This entails that no spurious solution may be introduced by dealing with the coarse-scale model and that inspection of the macroscopic behavior does not allow one to distinguish a composite from a mesoscopically homogeneous material. This part is the main issue of this paper and is based on techniques that are also studied in [83]. The restrictions that are required here are consistent with the Stefan model.

Plan of the paper. In section 2 we state some preliminary results that are used in the remainder of this paper. More specifically we review Nguetseng's notion of two-scale convergence and illustrate some properties of convexity and monotonicity that involve either single- or two-scale convergence, mainly referring to [81]. In section 3 we provide the weak formulation of an initial- and boundary-value problem for the system (1.1)–(1.3) and show existence of a weak solution.

In section 4 we introduce the dependence on the fine-scale variable x/ε, pass to the limit as the space period ε vanishes, and prove convergence to a solution of a two-scale problem. In section 5 we complete the homogenization procedure by upscaling the latter problem to a purely coarse-scale formulation and show that the two problems are equivalent. In section 6 we assume a linear dependence of the heat flux on the temperature gradient and retrieve the homogenized conductivity tensor via the two-scale formulation of a standard procedure; see [2]. (If the function K is scalar and independent of the temperature, we thus retrieve the results of [38].) In this section we also illustrate the use of the classical Kirchhoff transformation.

Literature. The Stefan problem and its various generalizations were studied in an impressive number of works; see, e.g., the monographs [1, 21, 36, 46, 48, 50, 62, 70, 71, 78], the references therein, and the extensive bibliography [75]. Physical and engineering aspects of phase transitions were dealt with, e.g., in [28, 47, 56, 85].

Homogenization, namely the search for effective models representing the macroscopic behavior of mesoscopically inhomogeneous materials, has also been and still is the object of intense research; see, e.g., [3, 7, 9, 11, 19, 33, 37, 39, 55, 61, 63, 72, 74]. In particular the homogenization of integral functionals was studied in [24, 60]; see also [19, 23, 30, 31, 37]. A new approach to periodic homogenization based on two-scale convergence was proposed by Nguetseng [65] and then developed by Allaire [2] and others; see also, e.g., [6, 18, 32]. This notion is receiving increasing attention; see, e.g., [59] for a recent review. Some results of [81, 82, 83] are applied in the present paper.

Apparently so far just a few works have been devoted to the homogenization of models of phase transitions. In [17, 38, 69] this was accomplished assuming the linear Fourier law with a space-oscillating conductivity $\vec{q} = -k(x/\varepsilon)\nabla u$, with k being a positive definite tensor function. In [17] these results were also applied to a univariate magnetic medium that under the eddy-current approximation was represented by a similar model. [17, 38] dealt with the homogenization of the weak formulation of the two-phase Stefan problem, whereas [69] addressed the analogous question for the single-phase problem via a well-known integral transformation due to C. Baiocchi and G. Duvaut, and also proved the convergence of the free boundary. [38] also raised the question of extending the homogenization result to a temperature-dependent conductivity (this is performed in the present work). The homogenization of the phase-field model of phase transition for binary mixtures was also addressed via two-scale convergence in [42, 43, 44]. Recently two-scale convergence was applied to the homogenization of stationary variational inequalities in [29].

Doubly nonlinear parabolic problems were studied in a number of works, e.g., [4, 10, 12, 13, 14, 25, 27, 40, 49, 53, 57, 58, 66, 76, 78]. Equations of this class may model not only phase transitions but also filtration of either gas or liquid through porous media; see, e.g., [4]; their homogenization was addressed, for instance, in [18, 52, 64]. In [64] oscillations w.r.t. the time variable were also accounted for in the elliptic term, and consequently space-time two-scale convergence was used. The uniqueness of the solution for doubly nonlinear elliptic-parabolic equations is less obvious than for the standard Stefan model; however, it was proved in [14, 25, 26, 53] by using the notions of entropy solutions, renormalization, and L^1-contractions. These results cover a fairly general setting that includes the problem addressed in the present work.

Our argument for the existence of a solution for the parabolic problem in the presence of a nonstrictly monotone elliptic part is based on classical techniques; in particular it rests on a result (i.e., Theorem 2.10 below) that, in turn, is based on Lemma 5 of [22, p. 27]; in this respect see also [15, 16, 57].

The methods of this paper may also be applied to other quasi-linear equations that represent different phenomena. For instance, a doubly nonlinear parabolic problem arises as a model of electromagnetic processes under the eddy-current approximation; its homogenization is studied in the parallel paper [84]. That setting differs from the present one under several respects: there the unknown fields are vectors, the energy balance and the Fourier law are replaced by the Maxwell and Ohm equations, the curl and the divergence operators occur in place of the gradient, compensated compactness plays a key role, and so on. Some results of [84] and of the present work were announced in the note [80].

2. Two-scale convergence, convexity, and monotonicity. In this section first we review some basic properties of two-scale convergence, along the lines of the fundamental works [2, 65]. We then state some results concerning convexity, monotonicity, and either single- or two-scale convergence that will be used afterwards; for these arguments we mainly refer to [81], with the exception of Theorem 2.10, which we prove here.

Let us set $Y := [0,1[^3$, denote by $\mathcal{Y}$ the same set equipped with the metric of the three-dimensional unit torus, and identify any Y-periodic function on $\mathbf{R}^3$ with a function on $\mathcal{Y}$. Let us denote by ε a parameter that we assume vanishes along a prescribed sequence. Let $p \in [1,+\infty[$ (resp., $p=\infty$), $\{u_\varepsilon\}$ be any bounded sequence in $L^p(\mathbf{R}^3)$, and $u \in L^p(\mathbf{R}^3 \times \mathcal{Y})$; by a natural extension of Nguetseng's definition of [65], we say that u_ε *weakly* (resp., *weakly star*) *two-scale converges* to u in $L^p(\mathbf{R}^3 \times \mathcal{Y})$, and write $u_\varepsilon \underset{2}{\rightharpoonup} u$ (resp., $u_\varepsilon \underset{2}{\overset{*}{\rightharpoonup}} u$), whenever

$$(2.1)\qquad \int_{\mathbf{R}^3} u_\varepsilon(x)\, v(x, x/\varepsilon)\, dx \to \iint_{\mathbf{R}^3\times\mathcal{Y}} u(x,y)\, v(x,y)\, dxdy \qquad \forall v \in \mathcal{D}(\mathbf{R}^3 \times \mathcal{Y}).$$

If $p \in]1,+\infty[$, following Allaire [2] we say that u_ε *strongly two-scale converges* to u in $L^p(\mathbf{R}^3 \times \mathcal{Y})$, and write $u_\varepsilon \underset{2}{\to} u$, whenever (2.1) holds and $\|u_\varepsilon\|_{L^p(\mathbf{R}^3)} \to \|u\|_{L^p(\mathbf{R}^3\times\mathcal{Y})}$. On the other hand, we denote the standard *(single-scale)* weak (resp., weak star, strong) convergence by $\rightharpoonup$ (resp., $\overset{*}{\rightharpoonup}$, $\to$). These definitions and the next two statements trivially take over to vector functions and also to functions defined on a domain Ω of $\mathbf{R}^3$ just by extending these functions with vanishing value outside Ω.

PROPOSITION 2.1 (see [2, 65]). *For any bounded sequence $\{u_\varepsilon\}$ of $L^p(\mathbf{R}^3)$ ($p \in]1,+\infty]$), there exists $u \in L^p(\mathbf{R}^3 \times \mathcal{Y})$ such that, possibly extracting a subsequence,*

$$(2.2)\qquad u_\varepsilon \underset{2}{\rightharpoonup} u \qquad in\ L^p(\mathbf{R}^3 \times \mathcal{Y}) \quad (u_\varepsilon \underset{2}{\overset{*}{\rightharpoonup}} u \quad if\ \ p=\infty).$$

Proposition 2.2 (see [2]). *If $\{u_\varepsilon\}$ is a sequence of $L^p(\mathbf{R}^3)$ ($p \in [1,+\infty[$) such that $u_\varepsilon \underset{2}{\rightharpoonup} u$ in $L^p(\mathbf{R}^3 \times \mathcal{Y})$, then*

$$u_\varepsilon \rightharpoonup \hat{u} := \int_{\mathcal{Y}} u(\cdot, y)\, dy \qquad \text{in } L^p(\mathbf{R}^3), \tag{2.3}$$

$$\liminf_{\varepsilon\to 0} \|u_\varepsilon\|_{L^p(\mathbf{R}^3)} \geq \|u\|_{L^p(\mathbf{R}^3\times\mathcal{Y})} \geq \|\hat{u}\|_{L^p(\mathbf{R}^3)}. \tag{2.4}$$

If $u_\varepsilon \to \hat{u}$ in $L^p(\mathbf{R}^3)$, then $u = \hat{u}$ (that is, u does not depend on y).

Dealing with functions of (x, y) we shall denote the gradient operator w.r.t. x (resp., y) by ∇_x (resp., ∇_y). For any $v \in L^1_{\text{loc}}(\mathbf{R}^3 \times \mathcal{Y})$ we define the average and fluctuating components:

$$\hat{v}(x) := \int_{\mathcal{Y}} v(x, y)\, dy, \qquad \tilde{v}(x) := v(x, y) - \hat{v}(x) \qquad \text{for a.e. } (x, y) \in \mathbf{R}^3 \times \mathcal{Y}, \tag{2.5}$$

and similarly for vectors. The next two statements deal with the two-scale limit of derivatives.

Proposition 2.3 (see [2, 32, 65]). *Let $p \in]1,+\infty[$, and let a sequence $\{u_\varepsilon\}$ of $W^{1,p}(\mathbf{R}^3)$ be such that $u_\varepsilon \rightharpoonup u$ in $W^{1,p}(\mathbf{R}^3)$. Then there exists $u_\sharp \in L^p\big(\mathbf{R}^3; W^{1,p}(\mathcal{Y})\big)$ such that $\widehat{u_\sharp} = 0$ a.e. in $\mathbf{R}^3$ and, as $\varepsilon \to 0$ along a suitable subsequence,*

$$\nabla u_\varepsilon \underset{2}{\rightharpoonup} \nabla_x u + \nabla_y u_\sharp \qquad \text{in } L^p(\mathbf{R}^3 \times \mathcal{Y})^3. \tag{2.6}$$

Henceforth we confine ourselves to $p = 2$. Let us set $L^2_{\text{div}}(\mathbf{R}^3)^3 := \big\{\vec{v} \in L^2(\mathbf{R}^3)^3 : \nabla \cdot \vec{v} \in L^2(\mathbf{R}^3)\big\}$, which is a Hilbert space equipped with the graph norm, and denote by $\nabla\times$ the curl operator.

Proposition 2.4 (see [79, 82]). *Let $\{\vec{u}_\varepsilon\}$ be a bounded sequence of $L^2_{\text{div}}(\mathbf{R}^3)^3$ such that $\vec{u}_\varepsilon \underset{2}{\rightharpoonup} \vec{u}$ in $L^2(\mathbf{R}^3\times\mathcal{Y})^3$. Then $\hat{\vec{u}} \in L^2_{\text{div}}(\mathbf{R}^3)^3$ and $\nabla_y \cdot \vec{u} = 0$ in $\mathcal{D}'(\mathbf{R}^3\times\mathcal{Y})$. Moreover, there exists $\vec{u}_\sharp \in L^2\big(\mathbf{R}^3; H^1(\mathcal{Y})^3\big)$ such that $\widehat{\vec{u}_\sharp} = \vec{0}$ a.e. in $\mathbf{R}^3$, $\nabla_y \times \vec{u}_\sharp = \vec{0}$ a.e. in $\mathbf{R}^3 \times \mathcal{Y}$, and, as $\varepsilon \to 0$ along a suitable subsequence,*

$$\nabla \cdot \vec{u}_\varepsilon \underset{2}{\rightharpoonup} \nabla_x \cdot \hat{\vec{u}} + \nabla_y \cdot \vec{u}_\sharp \qquad \text{in } L^2(\mathbf{R}^3 \times \mathcal{Y}). \tag{2.7}$$

(An analogous statement holds for the curl operator, with exchanged roles of curl and divergence [79, 82].) The above definitions and results are trivially extended to time-dependent functions, if time is regarded as a parameter, and to functions defined on a subdomain of $\mathbf{R}^3$.

Convexity and two-scale convergence. Here we review some results about convex integral functionals and maximal monotone graphs that will be applied in the next sections. For several statements we provide two versions: the first one is written for a fixed value of the parameter $\varepsilon > 0$, concerns the standard (single-scale) convergence, and stems from classical properties. We then extend it to two-scale convergence as $\varepsilon \to 0$; see [81].

First, we state some known properties of integral functionals. We shall denote by $\mathcal{L}(\Omega)$ (resp., $\mathcal{B}(\Omega)$) the σ-algebra of Lebesgue- (resp., Borel-) measurable subsets of Ω, define $\mathcal{L}(\mathcal{Y})$ and $\mathcal{B}(\mathcal{Y})$ similarly, and denote by $\mathcal{A}_1 \otimes \mathcal{A}_2$ the σ-algebra generated by any pair of σ-algebras $\mathcal{A}_1, \mathcal{A}_2$.

LEMMA 2.5 (see [81]). (i) *Assume that*

$$\begin{aligned}
&\varphi : \mathbf{R} \times \Omega \times \mathcal{Y} \to \mathbf{R} \cup \{+\infty\},\\
&\varphi(v,\cdot,\cdot) \text{ is measurable w.r.t. either } \mathcal{B}(\Omega) \otimes \mathcal{L}(\mathcal{Y}) \text{ or } \mathcal{L}(\Omega) \otimes \mathcal{B}(\mathcal{Y}) \qquad \forall v,\\
&\varphi(\cdot,x,y) \text{ is convex and lower semicontinuous} \qquad \text{for a.e. } (x,y),\\
&\{v \in \mathbf{R} : \varphi(v,x,y) < +\infty\} \text{ has nonempty interior} \qquad \text{for a.e. } (x,y).
\end{aligned} \tag{2.8}$$

Then

$$\begin{aligned}
&x \mapsto \varphi(v(x), x, x/\varepsilon) \text{ is measurable} \qquad \forall v \in L^1_{\mathrm{loc}}(\Omega),\ \forall \varepsilon > 0,\\
&(x,y) \mapsto \varphi(v(x,y), x, y) \text{ is measurable} \qquad \forall v \in L^1_{\mathrm{loc}}(\Omega \times \mathcal{Y}).
\end{aligned} \tag{2.9}$$

(ii) *Moreover, assume that*

$$\exists w \in L^2(\Omega),\ \exists h \in L^1(\Omega) : \quad \varphi(v,x,y) \ge w(x)v + h(x) \qquad \forall v \text{ for a.e. } (x,y), \tag{2.10}$$

and set

$$\begin{aligned}
&\varphi_\varepsilon(v,x) := \varphi(v,x,x/\varepsilon) \qquad \forall (v,x) \in \mathbf{R} \times \Omega,\ \forall \varepsilon > 0,\\
&\Phi_\varepsilon : L^2(\Omega) \to \mathbf{R} \cup \{+\infty\} : v \mapsto \int_\Omega \varphi(v,x,x/\varepsilon)\,dx \qquad \forall \varepsilon > 0,\\
&\Phi : L^2(\Omega \times \mathcal{Y}) \to \mathbf{R} \cup \{+\infty\} : v \mapsto \iint_{\Omega \times \mathcal{Y}} \varphi(v(x,y),x,y)\,dxdy.
\end{aligned} \tag{2.11}$$

The functionals Φ_ε and Φ are then convex and lower semicontinuous.

(iii) *Under the above hypotheses, for any sequence $\{v_\varepsilon\}$ in $L^2(\Omega)$,*

$$v_\varepsilon \underset{2}{\to} v \quad \text{in } L^2(\Omega \times \mathcal{Y}) \quad \Rightarrow \quad \Phi_\varepsilon(v_\varepsilon) \to \Phi(v), \tag{2.12}$$

$$v_\varepsilon \underset{2}{\rightharpoonup} v \quad \text{in } L^2(\Omega \times \mathcal{Y}) \quad \Rightarrow \quad \liminf_{\varepsilon \to 0} \Phi_\varepsilon(v_\varepsilon) \ge \Phi(v). \tag{2.13}$$

We shall denote by φ^* the *convex conjugate function* of φ w.r.t. the first variable; see, e.g., [45, 51, 68]. The convex conjugate functionals Φ^*_ε and Φ^* then coincide with the integral functionals of φ^*_ε and φ^*, respectively; see, e.g., [45, 67].

The next result is based on simple properties of convexity. We shall not display the proof and just refer the reader to the analogous argument of Theorem 2.1 of [81]. The weight function θ is here introduced in view of the application of this result that we shall perform in the next sections.

PROPOSITION 2.6. *Let φ fulfil* (2.8) *and* (2.10)*, fix any $\varepsilon > 0$, and define $\varphi_\varepsilon, \Phi_\varepsilon, \Phi$ as in* (2.11)*. Let θ be any nonnegative function of $\mathcal{D}(\bar{\Omega})$, and let $\{u_m\}$ and $\{w_m\}$ be two sequences of $L^2(\Omega_T)$. If*

$$w_m(x,t) \in \partial\varphi_\varepsilon(u_m(x,t),x) \qquad \text{for a.e. } (x,t) \in \Omega_T\ \forall m, \tag{2.14}$$

$$u_m \rightharpoonup u,\ \ w_m \rightharpoonup w \qquad \text{in } L^2(\Omega_T), \tag{2.15}$$

$$\liminf_{m \to \infty} \iint_{\Omega_T} u_m w_m \theta\,dxdt \le \iint_{\Omega_T} uw\,\theta\,dxdt, \tag{2.16}$$

then, denoting by $\tilde{\Omega}$ the support of θ,

$$w(x,t) \in \partial\varphi_\varepsilon(u(x,t),x) \qquad \textit{for a.e. } (x,t) \in \tilde{\Omega}_T, \tag{2.17}$$

$$\iint_{\Omega_T} \varphi_\varepsilon(u_m,x)\theta \, dxdt \to \iint_{\Omega_T} \varphi_\varepsilon(u,x)\theta \, dxdt, \tag{2.18}$$

$$\iint_{\Omega_T} \varphi^*_\varepsilon(w_m,x)\theta \, dxdt \to \iint_{\Omega_T} \varphi^*_\varepsilon(w,x)\theta \, dxdt, \tag{2.19}$$

$$\iint_{\Omega_T} u_m w_m \theta \, dxdt \to \iint_{\Omega_T} uw \, \theta \, dxdt. \tag{2.20}$$

Next we state an extension to two-scale convergence.

Proposition 2.7 (see [81]). *Let φ fulfil* (2.8) *and* (2.10), *and define $\varphi_\varepsilon, \Phi_\varepsilon, \Phi$ as in* (2.11) *for any $\varepsilon > 0$. Let $\{u_\varepsilon\}$ and $\{w_\varepsilon\}$ be two sequences of $L^2(\Omega_T)$, and let θ be any nonnegative function of $\mathcal{D}(\bar{\Omega})$. If*

$$w_\varepsilon(x,t) \in \partial\varphi_\varepsilon(u_\varepsilon(x,t),x) \qquad \textit{for a.e. } (x,t) \in \Omega_T \ \forall\varepsilon, \tag{2.21}$$

$$u_\varepsilon \underset{2}{\rightharpoonup} u, \ \ w_\varepsilon \underset{2}{\rightharpoonup} w \qquad \textit{in } L^2(\Omega_T \times \mathcal{Y}), \tag{2.22}$$

$$\liminf_{\varepsilon\to 0} \iint_{\Omega_T} u_\varepsilon w_\varepsilon \, \theta \, dxdt \le \iiint_{\Omega_T\times\mathcal{Y}} uw \, \theta \, dxdydt, \tag{2.23}$$

then, denoting by $\tilde{\Omega}$ the support of θ,

$$w(x,y,t) \in \partial\varphi(u(x,y,t),x,y) \qquad \textit{for a.e. } (x,y,t) \in \tilde{\Omega}_T \times \mathcal{Y}, \tag{2.24}$$

$$\iint_{\Omega_T} \varphi_\varepsilon(u_m,x)\theta \, dxdt \to \iiint_{\Omega_T\times\mathcal{Y}} \varphi(u,x,y)\theta \, dxdydt, \tag{2.25}$$

$$\iint_{\Omega_T} \varphi^*_\varepsilon(w_m,x)\theta \, dxdt \to \iiint_{\Omega_T\times\mathcal{Y}} \varphi^*(w,x,y)\theta \, dxdydt, \tag{2.26}$$

$$\iint_{\Omega_T} u_\varepsilon w_\varepsilon \theta \, dxdt \to \iiint_{\Omega_T\times\mathcal{Y}} uw \, \theta \, dxdydt. \tag{2.27}$$

Compactness by strict convexity. The following result will be used in the next section.

Proposition 2.8 (see [77], [78, Chap. X]). *Let φ fulfil* (2.8) *and be such that*

$$\begin{aligned} &\exists c > 0, \ \exists w \in L^2(\Omega), \ \exists h \in L^1(\Omega): \\ &\varphi(v,x,y) \ge w(x)v + cv^2 + h(x) \qquad \forall v \textit{ for a.e. } (x,y), \end{aligned} \tag{2.28}$$

$$v \mapsto \varphi(v,x,y) \textit{ is strictly convex} \qquad \textit{for a.e. } (x,y). \tag{2.29}$$

Let us then fix any $\varepsilon > 0$, and define φ_ε as in $(2.11)_1$. *If $\{u_m\}$ is a sequence of $L^2(\Omega_T)$ such that*

$$u_m \rightharpoonup u \qquad \textit{in } L^2(\Omega_T), \tag{2.30}$$

$$\iint_{\Omega_T} \varphi_\varepsilon(u_m,x) \, dxdt \to \iint_{\Omega_T} \varphi_\varepsilon(u,x) \, dxdt, \tag{2.31}$$

then

$$u_m \to u \qquad \textit{in } L^2(\Omega_T). \tag{2.32}$$

Here is an extension to two-scale convergence.

PROPOSITION 2.9 (see [81]). *Let φ fulfil* (2.8), (2.28), *and* (2.29), *and define φ_ε as in* $(2.11)_1$ *for any $\varepsilon > 0$. If $\{u_\varepsilon\}$ is a sequence of $L^2(\Omega_T)$ such that*

$$u_\varepsilon \underset{2}{\rightharpoonup} u \qquad \text{in } L^2(\Omega_T \times \mathcal{Y}), \tag{2.33}$$

$$\iint_{\Omega_T} \varphi_\varepsilon(u_\varepsilon, x)\, dxdt \to \iiint_{\Omega_T \times \mathcal{Y}} \varphi(u, x, y)\, dxdydt, \tag{2.34}$$

then

$$u_\varepsilon \underset{2}{\to} u \qquad \text{in } L^2(\Omega_T \times \mathcal{Y}). \tag{2.35}$$

Monotonicity and two-scale convergence. After dealing with the subdifferential of convex lower semicontinuous functions, namely maximal cyclically monotone functions, now we extend some of these results to the larger class of maximal monotone functions. Here we confine ourselves to single-valued functions, and in view of the developments of the next sections we also allow for continuous dependence on a further scalar argument. Let us assume that

$$\begin{aligned} &\vec{\alpha} : \mathbf{R}^3 \times \mathbf{R} \times \Omega \times \mathcal{Y} \times]0, T[\to \mathbf{R}^3, \\ &\vec{\alpha}(\cdot, u, x, y, t) \text{ is monotone} \qquad \forall u \text{ and for a.e. } (x, y, t), \\ &\vec{\alpha}(\cdot, \cdot, x, y, t) \text{ is continuous} \qquad \text{for a.e. } (x, y, t), \text{ and} \\ &\vec{\alpha}(\vec{z}, u, \cdot, \cdot, \cdot) \text{ is measurable w.r.t. either } \mathcal{B}(\Omega) \otimes \mathcal{L}(\mathcal{Y}) \otimes \mathcal{L}(]0, T[) \\ &\text{or } \mathcal{L}(\Omega) \otimes \mathcal{B}(\mathcal{Y}) \otimes \mathcal{L}(]0, T[) \qquad \forall (\vec{z}, u), \end{aligned} \tag{2.36}$$

$$\begin{aligned} &\exists L > 0,\, \exists \xi \in L^2(\Omega_T) : \forall (\vec{z}, u) \qquad \text{for a.e. } (x, y, t), \\ &|\vec{\alpha}(\vec{z}, u, x, y, t)| \le L|\vec{z}| + L|u| + \xi(x, t), \end{aligned} \tag{2.37}$$

and set

$$\vec{\alpha}_\varepsilon(\vec{z}, u, x, t) := \vec{\alpha}(\vec{z}, u, x, x/\varepsilon, t) \qquad \forall (\vec{z}, u) \in \mathbf{R}^3 \times \mathbf{R} \text{ for a.e. } (x, t) \in \Omega_T\ \forall \varepsilon > 0. \tag{2.38}$$

This function is measurable w.r.t. x and thus is of Caratheodory class.

The first part of the next statement is a simple extension of a standard result of the theory of maximal monotone graphs; see, e.g., [20]. The second part is based on Lemma 5 of [22, p. 27]; see also [16, 15], [57, p. 183]. (A proof is needed, however, for this statement is not a direct consequence of the known results.)

THEOREM 2.10. *Let $\vec{\alpha}$ fulfil* (2.36) *and* (2.37), *fix any $\varepsilon > 0$, and define $\vec{\alpha}_\varepsilon$ as in* (2.38). *Let $\{\vec{z}_m\}$ (resp., $\{u_m\}$) be a sequence of $L^2(\Omega_T)^3$ (resp., $L^2(\Omega_T)$).*

(i) *If*

$$\vec{r}_m(x, t) := \vec{\alpha}_\varepsilon(\vec{z}_m(x, t), u_m(x, t), x, t) \rightharpoonup \vec{r} \qquad \text{in } L^2(\Omega_T)^3, \tag{2.39}$$

$$\vec{z}_m \rightharpoonup \vec{z} \qquad \text{in } L^2(\Omega_T)^3, \tag{2.40}$$

$$u_m \to u \qquad \text{in } L^2(\Omega_T), \tag{2.41}$$

$$\liminf_{m \to \infty} \iint_{\Omega_T} \vec{r}_m \cdot \vec{z}_m\, dxdt \le \iint_{\Omega_T} \vec{r} \cdot \vec{z}\, dxdt, \tag{2.42}$$

then

$$\vec{r}(x, t) = \vec{\alpha}_\varepsilon(\vec{z}(x, t), u(x, t), x, t) \qquad \text{for a.e. } (x, t) \in \Omega_T. \tag{2.43}$$

(ii) *If, moreover,*

$$
\begin{aligned}
&\text{for a.e. } (x,y) \in \Omega \times \mathcal{Y} \qquad \forall \vec{s} \in \mathbf{R}^3,\ \forall \{\vec{s}_n\} \subset \mathbf{R}^3,\\
&\forall \text{ bounded sequence } \{v_n\} \subset \mathbf{R} \qquad \text{for a.e. } t \in\,]0,T[,\\
&[\vec{\alpha}(\vec{s}, v_n, x, y, t) - \vec{\alpha}(\vec{s}_n, v_n, x, y, t)] \cdot (\vec{s} - \vec{s}_n) \to 0 \quad \Rightarrow \quad \vec{s}_n \to \vec{s},
\end{aligned}
\tag{2.44}
$$

$$
\begin{aligned}
&\exists c_1 \in \mathbf{R},\ \exists c_2 > 0,\ \exists \ell \in L^1(\Omega_T):\\
&\text{for a.e. } (x,y,t) \in \Omega_T \times \mathcal{Y} \qquad \forall (\vec{z}, v) \in \mathbf{R}^3 \times \mathbf{R},\\
&\vec{\alpha}(\vec{z}, v, x, y, t) \cdot \vec{z} + c_1 |v|^2 \geq c_2 |\vec{z}|^2 + \ell(x,t),
\end{aligned}
\tag{2.45}
$$

then

$$
\vec{z}_m \to \vec{z}, \qquad \vec{\alpha}_\varepsilon(\vec{z}_m, u_m, x, t) \to \vec{\alpha}_\varepsilon(\vec{z}, u, x, t) \qquad \text{in } L^2(\Omega_T)^3. \tag{2.46}
$$

The condition (2.44) may be interpreted as an assumption of strict monotonicity of $\vec{w} \mapsto \vec{\alpha}_\varepsilon(\vec{w}, v, x, t)$ uniform w.r.t. v.

Proof. (i) The monotonicity of $\vec{\alpha}_\varepsilon$ w.r.t. its first argument yields

$$
\iint_{\Omega_T} [\vec{r}_m - \vec{\alpha}_\varepsilon(\vec{s}, u_m, x, t)] \cdot (\vec{z}_m - \vec{s})\, dxdt \geq 0 \qquad \forall \vec{s} \in L^2(\Omega_T)^3.
$$

By (2.36), (2.37), and (2.41),

$$
\vec{\alpha}_\varepsilon(\vec{s}, u_m, x, t) \to \vec{\alpha}_\varepsilon(\vec{s}, u, x, t) \qquad \text{in } L^2(\Omega_T)^3\ \forall \vec{s} \in L^2(\Omega_T)^3. \tag{2.47}
$$

By passing to the inferior limit as $m \to \infty$ in the latter inequality, by (2.39)–(2.42) we then get

$$
\iint_{\Omega_T} [\vec{r} - \vec{\alpha}_\varepsilon(\vec{s}, u, x, t)] \cdot (\vec{z} - \vec{s})\, dxdt \geq 0 \qquad \forall \vec{s} \in L^2(\Omega_T)^3.
$$

By the maximal monotonicity of $\vec{\alpha}_\varepsilon$ w.r.t. its first argument, this inequality is equivalent to (2.43).

(ii) Let us now set

$$
P_m(x,t) := [\vec{\alpha}_\varepsilon(\vec{z}_m, u_m, x, t) - \vec{\alpha}_\varepsilon(\vec{z}, u_m, x, t)] \cdot (\vec{z}_m - \vec{z}) \qquad \text{for a.e. } (x,t) \in \Omega_T.
$$

By (2.39), (2.40), (2.42), and (2.47), $\liminf_{m\to\infty} \iint_{\Omega_T} P_m(x,t)\, dxdt \leq 0$. As P_m is nonnegative this means that for a suitable sequence that we label by m':

$$
P_{m'} \to 0 \qquad \text{in } L^1(\Omega_T). \tag{2.48}
$$

Hence there exists a subsequence that we label by m'' such that $P_{m''}(\cdot,\cdot) \to 0$ a.e. in Ω_T. By (2.41), possibly extracting a further subsequence, $u_{m''} \to u$ a.e. in Ω_T. By (2.44) we then get

$$
\vec{z}_{m''} \to \vec{z} \qquad \text{a.e. in } \Omega_T, \tag{2.49}
$$

whence

$$
\vec{\alpha}_\varepsilon(\vec{z}_{m''}, u_{m''}, x, t) \to \vec{\alpha}_\varepsilon(\vec{z}, u, x, t) = \vec{r} \qquad \text{a.e. in } \Omega_T. \tag{2.50}
$$

By (2.37), (2.40), (2.41), and (2.50), we have

$$\iint_{\Omega_T} \vec{\alpha}_\varepsilon(\vec{z}, u_{m''}, x, t) \cdot (\vec{z}_{m''} - z)\, dxdt \to 0. \tag{2.51}$$

By the definition of P_m, (2.40), (2.48), and (2.51), we get

$$\begin{aligned}
&\iint_{\Omega_T} \vec{\alpha}_\varepsilon(\vec{z}_{m''}, u_{m''}, x, t) \cdot \vec{z}_{m''}\, dxdt \\
&= \iint_{\Omega_T} [P_{m''}(x,t) + \vec{\alpha}_\varepsilon(\vec{z}, u_{m''}, x, t) \cdot (\vec{z}_{m''} - z)]\, dxdt \\
&+ \iint_{\Omega_T} \vec{\alpha}_\varepsilon(\vec{z}_{m''}, u_{m''}, x, t) \cdot \vec{z}\, dxdt \to \iint_{\Omega_T} \vec{r} \cdot \vec{z}\, dxdt.
\end{aligned}$$

We claim that this entails that

$$\xi_{m''} := \vec{r}_{m''} \cdot \vec{z}_{m''} \to \vec{r} \cdot \vec{z} =: \xi \qquad \text{in } L^1(\Omega_T)^3. \tag{2.52}$$

In view of proving this statement, first notice that we already know that

$$\xi_{m''} \geq 0, \qquad \xi_{m''} \to \xi \qquad \text{a.e. in } \Omega_T, \qquad \iint_{\Omega_T} \xi_{m''}\, dxdt \to \iint_{\Omega_T} \xi\, dxdt.$$

Moreover, setting $A_{m''} := \{(x,t) \in \Omega_T : \xi_{m''}(x,t) \leq \xi(x,t)\}$,

$$\iint_{\Omega_T} |\xi_{m''} - \xi|\, dxdt = 2 \iint_{A_{m'}} (\xi - \xi_{m''})\, dxdt + \iint_{\Omega_T} (\xi_{m''} - \xi)\, dxdt.$$

The Lebesgue dominated convergence theorem then yields (2.52).

By (2.45) we have

$$c_2 |\vec{z}_{m''}|^2 \leq \vec{r}_{m''} \cdot \vec{z}_{m''} + c_1 u_{m''}^2 - \ell \qquad \text{a.e. in } \Omega;$$

(2.41) and (2.52) then entail that the sequence $\{|\vec{z}_{m''}|^2\}$ is equi-integrable in Ω. By (2.37) the same then applies to $\{|\vec{r}_{m''}|^2\}$, so that by Vitali's convergence theorem we infer that

$$\vec{z}_{m''} \to \vec{z}, \qquad \vec{r}_{m''} = \vec{\alpha}_\varepsilon(\vec{z}_{m''}, u_{m''}, x, t) \to \vec{\alpha}_\varepsilon(\vec{z}, u, x, t) \qquad \text{in } L^2(\Omega_T)^3. \tag{2.53}$$

As both limits are independent of the extracted subsequence, the whole sequence converges; (2.46) is thus established. ☐

Here is an extension to two-scale convergence that rests on a similar argument.

PROPOSITION 2.11 (see [81]). *Let $\vec{\alpha}$ fulfil* (2.36) *and* (2.37), *and define $\vec{\alpha}_\varepsilon$ as in* (2.38) *for any $\varepsilon > 0$. Let $\{\vec{r}_\varepsilon\}$ (resp., $\{u_\varepsilon\}$) be a sequence of $L^2(\Omega_T)^3$ (resp., $L^2(\Omega_T)$).*

(i) *If*

$$\vec{r}_\varepsilon(x,t) = \vec{\alpha}_\varepsilon(\vec{z}_\varepsilon(x,t), u_\varepsilon(x,t), x, t) \qquad \textit{for a.e. } (x,t) \in \Omega_T\ \forall \varepsilon, \tag{2.54}$$

$$\vec{z}_\varepsilon \underset{2}{\rightharpoonup} \vec{z}, \quad \vec{r}_\varepsilon \underset{2}{\rightharpoonup} \vec{r} \qquad \textit{in } L^2(\Omega_T \times \mathcal{Y})^3, \tag{2.55}$$

$$u_\varepsilon \underset{2}{\to} u \qquad \textit{in } L^2(\Omega_T \times \mathcal{Y}), \tag{2.56}$$

$$\liminf_{\varepsilon \to 0} \iint_{\Omega_T} \vec{r}_\varepsilon(x,t) \cdot \vec{z}_\varepsilon(x,t)\, dxdt \leq \iiint_{\Omega_T \times \mathcal{Y}} \vec{r}(x,y,t) \cdot \vec{z}(x,y,t)\, dxdydt, \tag{2.57}$$

then

$$\vec{r}(x,y,t) = \vec{\alpha}(\vec{z}(x,y,t), u(x,y,t), x, y, t) \qquad \textit{for a.e. } (x,y,t) \in \Omega_T \times \mathcal{Y}. \tag{2.58}$$

(ii) *If, moreover,* (2.44) *and* (2.45) *are satisfied, then*

$$\vec{z}_\varepsilon \underset{2}{\to} \vec{z}, \qquad \vec{\alpha}_\varepsilon(\vec{z}_\varepsilon, u_\varepsilon, x, t) \underset{2}{\to} \vec{\alpha}(\vec{z}, u, x, y, t) \qquad \textit{in } L^2(\Omega_T)^3. \tag{2.59}$$

3. A quasi-linear parabolic problem. In this section we deal with an initial- and boundary-value problem for the system (1.1)–(1.3): we provide a weak formulation in Sobolev spaces and prove existence of a solution.

We denote by Ω a domain of $\mathbf{R}^3$, fix any $T > 0$, and set $A_t := A \times]0,t[$ for any set A and any $t > 0$. We fix a partition $\{\Gamma_0, \Gamma_1\}$ of the boundary of Ω and prescribe the evolution of u on Γ_0 and of the normal component of ∇u on Γ_1. We also assume that we are given three constitutive functions

$$\varphi : \mathbf{R} \times \Omega \times \mathcal{Y} \to \mathbf{R} \cup \{+\infty\}, \qquad \vec{\alpha} : \mathbf{R}^3 \times \mathbf{R} \times \Omega \times \mathcal{Y} \times]0,T[\to \mathbf{R}^3,$$
$$\beta : \mathbf{R} \times \mathbf{R}^3 \times \Omega \times \mathcal{Y} \times]0,T[\to \mathbf{R}$$

that satisfy (2.8) and (2.36) and such that

$$\begin{aligned} &\exists c_1, \tilde{c}_1 > 0,\ \exists h_1, \tilde{h}_1 \in L^1(\Omega) : \text{ for a.e. } (x,y),\\ &\forall v \in \mathbf{R},\ \forall w \in \partial\varphi(v,x,y), \qquad c_1|v| + h_1(x) \le |w| \le \tilde{c}_1|v| + \tilde{h}_1(x), \end{aligned} \tag{3.1}$$

$$\begin{aligned} &\exists L > 0,\ \exists \ell_1 \in L^2(\Omega_T) : \forall(\vec{z}, v) \qquad \text{for a.e. } (x,y,t),\\ &|\vec{\alpha}(\vec{z}, v, x, y, t)| \le L|\vec{z}| + L|v| + \ell_1(x,t), \end{aligned} \tag{3.2}$$

$$\begin{aligned} &\text{for a.e. } (x,y,t) \in \Omega_T \times \mathcal{Y} \qquad \forall \vec{s} \in \mathbf{R}^3,\ \forall\{\vec{s}_n\} \subset \mathbf{R}^3,\ \forall\{v_n\} \subset \mathbf{R},\\ &[\vec{\alpha}(\vec{s}, v_n, x, y, t) - \vec{\alpha}(\vec{s}_n, v_n, x, y, t)] \cdot (\vec{s} - \vec{s}_n) \to 0 \quad \Rightarrow \quad \vec{s}_n \to \vec{s}, \end{aligned} \tag{3.3}$$

$$\begin{aligned} &\exists c_2 > 0,\ \exists \ell_2 \in L^1(\Omega_T) : \forall(\vec{z}, v) \qquad \text{for a.e. } (x,y,t),\\ &\vec{\alpha}(\vec{z}, v, x, y, t) \cdot \vec{z} \ge c_2|\vec{z}|^2 + \ell_2(x,t), \end{aligned} \tag{3.4}$$

$$\begin{aligned} &\beta(\cdot,\cdot,x,y,t) \text{ is continuous} \qquad \text{for a.e. } (x,y,t),\\ &\beta(v,\vec{z},\cdot,\cdot,\cdot) \text{ is measurable w.r.t. either } \mathcal{B}(\Omega) \otimes \mathcal{L}(\mathcal{Y}) \otimes \mathcal{L}(]0,T[)\\ &\text{or } \mathcal{L}(\Omega) \otimes \mathcal{B}(\mathcal{Y}) \otimes \mathcal{L}(]0,T[) \qquad \forall v, \vec{z}, \end{aligned} \tag{3.5}$$

$$\begin{aligned} &\exists c_3 > 0,\ \exists \ell_3 \in L^2(\Omega_T) : \forall v \in \mathbf{R},\ \forall \vec{z} \in \mathbf{R}^3,\\ &|\beta(v, \vec{z}, x, y, t)| \le c_3(|v| + |\vec{z}|) + \ell_3(x,t) \qquad \text{for a.e. } (x,y,t). \end{aligned} \tag{3.6}$$

Let us fix any $\varepsilon > 0$, define φ_ε, $\vec{\alpha}_\varepsilon$, and β_ε as in $(2.11)_1$ and (2.38), and set

$$\beta_\varepsilon(v, \vec{z}, x, t) := \beta(v, \vec{z}, x, x/\varepsilon, t) \qquad \forall v \in \mathbf{R},\ \forall \vec{z} \in \mathbf{R}^3 \text{ for a.e. } (x,t) \in \Omega_T\ \forall \varepsilon > 0. \tag{3.7}$$

In view of specifying the functional framework, we assume that the domain Ω is bounded and of Lipschitz class and that Γ_0 is measurable and has positive bidimensional Hausdorff measure. We denote by γ_0 the trace operator and set

$$V := \big\{ v \in H^1(\Omega) : \gamma_0 v = 0 \ \text{ a.e. on } \Gamma_0 \big\}, \qquad \|v\|_V := \|\nabla v\|_{L^2(\Omega)^3}.$$

By the Poincaré inequality, $\|\cdot\|_V$ is equivalent to the usual Sobolev norm, so that V is a closed Banach subspace of $H^1(\Omega)$. Identifying $H := L^2(\Omega)$ with its dual H' and the latter with a closed subspace of the dual space V' of V, we get the Hilbert triplet

$$V \subset H = H' \subset V' \qquad \text{with compact, continuous, and dense injections.}$$

We denote by $\langle \cdot, \cdot \rangle$ the duality pairing between V' and V and define the linear and continuous operator

$$\nabla^* \cdot : L^2(\Omega)^3 \to V', \qquad \langle \nabla^* \cdot \vec{z}, v \rangle := -\int_\Omega \vec{z} \cdot \nabla v \, dx \qquad \forall \vec{z} \in L^2(\Omega)^3, \, \forall v \in V$$

(thus $\nabla^* \cdot \vec{z} = \nabla \cdot \vec{z}$ in $\mathcal{D}'(\Omega)$). Finally, we assume that

$$w^0 \in L^2(\Omega), \qquad z \in H^1(\Omega_T), \qquad g \in L^2(0, T; V'), \tag{3.8}$$

and for any $\varepsilon > 0$ we introduce the weak formulation of an initial- and boundary-value problem for the system (1.1)–(1.3) with φ_ε and $\vec{\alpha}_\varepsilon$ in place of φ and $\vec{\alpha}$.

Problem 3.1_ε. *Find $u_\varepsilon \in L^2(0, T; V) + z$ and $w_\varepsilon \in L^2(\Omega_T)$ such that, setting*

$$\vec{q}_\varepsilon(x, t) = \vec{\alpha}_\varepsilon(-\nabla u_\varepsilon(x, t), u_\varepsilon(x, t), x, t) \qquad \textit{for a.e. } (x, t) \in \Omega_T, \tag{3.9}$$

$$f_\varepsilon(x, t) = \beta_\varepsilon(u_\varepsilon(x, t), \nabla u_\varepsilon(x, t), x, t) \qquad \textit{for a.e. } (x, t) \in \Omega_T, \tag{3.10}$$

one has

$$w_\varepsilon(x, t) \in \partial \varphi_\varepsilon(u_\varepsilon(x, t), x) \qquad \textit{for a.e. } (x, t) \in \Omega_T, \tag{3.11}$$

$$\begin{gathered} \iint_{\Omega_T} \left((w^0 - w_\varepsilon) \frac{\partial v}{\partial t} - \vec{q}_\varepsilon \cdot \nabla v - f_\varepsilon v \right) dxdt = \int_0^T \langle g, v \rangle \, dt \\ \forall v \in H^1(0, T; H) \cap L^2(0, T; V), \, v(\cdot, T) = 0 \qquad \textit{a.e. in } \Omega. \end{gathered} \tag{3.12}$$

Interpretation. Equation (3.12) yields the equation

$$\frac{\partial w_\varepsilon}{\partial t} + \nabla^* \cdot \vec{q}_\varepsilon = f_\varepsilon + g \qquad \text{in } V' \text{ a.e. in }]0, T[. \tag{3.13}$$

By comparing these terms, we see that $w_\varepsilon \in H^1(0, T; V')$. Integrating by parts in time in (3.12) we then get the initial condition

$$w_\varepsilon(\cdot, 0) = w_\varepsilon^0 \qquad \text{in } V'. \tag{3.14}$$

If

$$h \in L^2(\Gamma_{1T}), \qquad \langle g, v \rangle = \int_{\Gamma_1} h \, \gamma_0 v \, d\sigma \qquad \forall v \in V \text{ a.e. in }]0, T[, \tag{3.15}$$

then (3.13) accounts for the energy balance equation

$$\frac{\partial w_\varepsilon}{\partial t} + \nabla \cdot \vec{q}_\varepsilon = f_\varepsilon \qquad \text{in } \mathcal{D}'(\Omega_T), \tag{3.16}$$

coupled with the boundary condition

$$\vec{q}_\varepsilon \cdot \vec{\nu} = -h \qquad \text{on } \Gamma_{1T} \text{ (with } \vec{\nu} = \text{outward-oriented unit normal vector)} \tag{3.17}$$

in weak form.

THEOREM 3.1. *For any fixed $\varepsilon > 0$, let (2.8), (2.36), (3.1)–(3.6), and (3.8) be fulfilled, and define φ_ε, $\vec{\alpha}_\varepsilon$, and β_ε as in $(2.11)_1$, (2.38), and (3.7). Assume also that*

$$v \mapsto \varphi(v,x,y) \text{ is strictly convex} \qquad \text{for a.e. } (x,y). \tag{3.18}$$

Then there exists a solution of Problem 3.1_ε such that $w_\varepsilon \in L^\infty(0,T;H)$.

It is easily seen that, under the hypothesis (2.8), the assumption (3.18) is equivalent to the continuity of $(\partial\varphi)^{-1}(\cdot,x,y) = \partial\varphi^*(\cdot,x,y)$.

Proof. Throughout this argument we drop the index ε, in order to make formulas more readable.

(i) *Approximation.* Let us fix any $m \in \mathbf{N}$, set $k := T/m$, and

$$\left.\begin{aligned} z_m^n &:= \frac{1}{k}\int_{(n-1)k}^{nk} z(\cdot,t)\,dt, & g_m^n &:= \frac{1}{k}\int_{(n-1)k}^{nk} g(\cdot,t)\,dt,\\ \vec{\alpha}_m^n &:= \frac{1}{k}\int_{(n-1)k}^{nk} \vec{\alpha}(\cdot,\cdot,\cdot,t)\,dt, & \beta_m^n &:= \frac{1}{k}\int_{(n-1)k}^{nk} \beta(\cdot,\cdot,\cdot,t)\,dt \end{aligned}\right. \qquad (n = 1,\dots,m).$$

We then introduce an implicit time-discretization scheme of our problem.

PROBLEM 3.1_m. *Find $u_m^n \in V + z_m^n$ and $w_m^n \in L^2(\Omega)$ for $n = 1,\dots,m$ such that, setting $u_m^0 := 0$, $w_m^0 := w^0$, and*

$$\vec{q}_m^{\,n}(x) = \vec{\alpha}_m^n(-\nabla u_m^n(x), u_m^{n-1}(x), x) \qquad \text{for a.e. } x \in \Omega, \tag{3.19}$$

$$f_m^n(x) = \beta_m^n(u_m^{n-1}(x), \nabla u_m^{n-1}(x), x) \qquad \text{for a.e. } x \in \Omega, \tag{3.20}$$

one has

$$w_m^n(x) \in \partial\varphi(u_m^n(x), x) \qquad \text{for a.e. } x \in \Omega, \tag{3.21}$$

$$\frac{w_m^n - w_m^{n-1}}{k} + \nabla^* \cdot \vec{q}_m^{\,n} = f_m^n + g_m^n \qquad \text{in } V'. \tag{3.22}$$

In view of proving the existence of a solution of this problem, let us fix any $n \in \{1,\dots,m\}$, set

$$\mathcal{A}_m^n(v) := \partial\varphi(v,\cdot) + k\nabla^* \cdot \vec{\alpha}_m^n(-\nabla v, u_m^{n-1}, \cdot)\ (\subset V') \qquad \forall v \in H^1(\Omega),$$

and notice that for any n the system (3.21), (3.22) also reads

$$\mathcal{A}_m^n(u_m^n) \ni w_m^{n-1} + kf_m^n + kg_m^n \qquad \text{in } V'. \tag{3.23}$$

At the nth step the right-hand side of this equation is known. The operator $\mathcal{A}_m^n$ is maximal monotone, for it is the sum of two maximal monotone operators, and $v \mapsto \nabla^* \cdot \vec{\alpha}_m^n(-\nabla v, u_m^{n-1}, \cdot)$ is defined on the whole $H^1(\Omega)$. Because of (2.8) and (3.4), $\mathcal{A}_m^n$ is also coercive. Therefore (3.23) has a solution u_m^n, and this determines a solution of Problem 3.1_m.

(ii) *A priori estimates.* Let us now define time-interpolate functions as follows. For any family $\{v_m^n\}_{n=0,\dots,m}$ of numbers let us denote by v_m the piecewise linear time interpolate of $v_m^0 := v^0$, v_m^1, $\dots$, v_m^m a.e. in Ω. Let us denote by $\bar{v}_m$ the corresponding piecewise constant interpolate function; that is, $\bar{v}_m(t) := v_m^n$ if $(n-1)k < t \le nk$

for $n = 1, \dots, m$. Let us also set $\tau_k v := v(\cdot - k)$ for any function v of time. Setting $\bar{u}_m(x,t) := 0$ for any $t < 0$, the system (3.19)–(3.22) then reads

$$\bar{\vec{q}}_m(x,t) = \vec{\alpha}_m(-\nabla \bar{u}_m(x,t), \tau_k \bar{u}_m(x,t), x, t) \qquad \text{for a.e. } (x,t) \in \Omega_T, \tag{3.24}$$

$$\bar{f}_m(x,t) = \beta_m(\tau_k \bar{u}_m(x,t), \tau_k \nabla \bar{u}_m(x,t), x, t) \qquad \text{for a.e. } (x,t) \in \Omega_T, \tag{3.25}$$

$$\bar{w}_m(x,t) \in \partial\varphi(\bar{u}_m(x,t), x) \qquad \text{for a.e. } (x,t) \in \Omega_T, \tag{3.26}$$

$$\frac{\partial w_m}{\partial t} + \nabla^* \cdot \bar{\vec{q}}_m = \bar{f}_m + \bar{g}_m \qquad \text{in } V' \text{ a.e. in }]0,T[. \tag{3.27}$$

Multiplying the latter equation by $\bar{u}_m - \bar{z}_m$ and integrating in time, we get

$$\iint_{\Omega_t} \left(\frac{\partial w_m}{\partial t} \bar{u}_m - \bar{\vec{q}}_m \cdot \nabla(\bar{u}_m - \bar{z}_m) \right) dx d\tau = \iint_{\Omega_t} \frac{\partial w_m}{\partial t} \bar{z}_m \, dx d\tau$$
$$+ \iint_{\Omega_t} \bar{f}_m (\bar{u}_m - \bar{z}_m) \, dx d\tau + \int_0^t \langle \bar{g}_m, \bar{u}_m - \bar{z}_m \rangle \, d\tau \qquad \forall t \in]0,T].$$

By (3.26) $\bar{u}_m(x,t) \in \partial\varphi^*(\bar{w}_m(x,t), x)$ for a.e. $(x,t) \in \Omega_T$, whence

$$\iint_{\Omega_t} \frac{\partial w_m}{\partial t} \bar{u}_m \, dx d\tau \geq \int_\Omega \left[\varphi^*(\bar{w}_m(\cdot,t), x) - \varphi^*(w^0, x) \right] dx.$$

Moreover, denoting by $C_1, C_2, \dots$ suitable constants that may depend on ε but not on m, by (3.1) and (3.6) we have

$$\iint_{\Omega_t} \frac{\partial w_m}{\partial t} \bar{z}_m \, dx d\tau = - \iint_{\Omega_t} (\tau_k \bar{w}_m) \frac{\partial z_m}{\partial t} \, dx d\tau + \int_\Omega [\bar{w}_m(x,t) \bar{z}_m(x,t) - w^0(x) z_m(x,0)] \, dx$$
$$\leq C_1 \big(\|\bar{w}_m\|_{L^2(\Omega_t)} + \|\bar{w}_m(\cdot,t)\|_H + \|w^0\|_H \big) \|z_m\|_{H^1(0,T:H)},$$

$$\int_\Omega \bar{f}_m (\bar{u}_m - \bar{z}_m) \, dx \leq \big(c_3 \|\bar{u}_m\|_{H^1(\Omega)} + \|\ell_3\|_{L^2(\Omega)} \big) \left(\|\bar{u}_m\|_H + \|\bar{z}_m\|_H \right)$$
$$\leq \big(c_3 \|\bar{u}_m\|_{H^1(\Omega)} + \|\ell_3\|_{L^2(\Omega)} \big) \left(\frac{1}{c_1} \|\bar{w}_m\|_H + \frac{1}{c_1} \|h_1\|_H + \|\bar{z}_m\|_H \right) \qquad \text{a.e. in }]0,T[.$$

We thus get

$$\int_\Omega \left[\varphi^*(\bar{w}_m(\cdot,t), x) - \varphi^*(w^0, x) \right] dx - \iint_{\Omega_t} \bar{\vec{q}}_m \cdot \nabla(\bar{u}_m - \bar{z}_m) \, dx d\tau$$
$$\leq C_1 \big(\|\bar{w}_m\|_{L^2(\Omega_t)} + \|\bar{w}_m(\cdot,t)\|_H + \|w^0\|_H \big) \|z_m\|_{H^1(0,T:H)} + \int_0^t \langle \bar{g}_m, \bar{u}_m - \bar{z}_m \rangle \, d\tau$$
$$+ \big(c_3 \|\bar{u}_m\|_{H^1(\Omega)} + \|\ell_3\|_{L^2(\Omega)} \big) \left(\frac{1}{c_1} \|\bar{w}_m\|_H + \frac{1}{c_1} \|h_1\|_H + \|\bar{z}_m\|_H \right) \qquad \forall t \in]0,T]. \tag{3.28}$$

Notice that by (3.1)

$$\frac{c_1}{2} |\bar{w}_m|^2 \leq \varphi(\bar{w}_m, \cdot) + |h_1 \bar{w}_m| \qquad \text{a.e. in } \Omega. \tag{3.29}$$

Recalling (3.4), (3.6), and (3.8) a standard calculation then yields

$$\|w_m\|_{L^\infty(0,T;H)}, \ \|u_m\|_{L^2(0,T;H^1(\Omega))}, \ \|f_m\|_{L^2(\Omega_T)} \leq C_2. \tag{3.30}$$

By applying (3.2) and comparing the terms of (3.27), we also infer that

$$\|\vec{q}_m\|_{L^2(\Omega_T)},\ \|w_m\|_{L^\infty(0,T;H)\cap H^1(0,T;V')} \leq C_3. \tag{3.31}$$

(iii) *Limit procedure.* By these estimates there exist $u, w, \vec{q}, f$ such that, as $m \to \infty$ along a suitable sequence (still omitting the fixed index ε),

$$u_m \rightharpoonup u \qquad \text{in } L^2\big(0,T;H^1(\Omega)\big), \tag{3.32}$$

$$w_m \overset{*}{\rightharpoonup} w \qquad \text{in } L^\infty(0,T;H) \cap H^1(0,T;V'), \tag{3.33}$$

$$\vec{q}_m \rightharpoonup \vec{q} \qquad \text{in } L^2(\Omega_T)^3, \tag{3.34}$$

$$f_m \rightharpoonup f \qquad \text{in } L^2(\Omega_T). \tag{3.35}$$

Passing to the limit in (3.27) we then get (3.13), whence (3.12) follows. Let us now fix any nonnegative function $\theta \in \mathcal{D}(\bar{\Omega})$. By a compactness result of Aubin (see [8, 57, 73]), (3.32) and (3.33) yield

$$\iint_{\Omega_t} \bar{w}_m\, \bar{u}_m\, \theta\, dxd\tau \to \iint_{\Omega_t} w\, u\, \theta\, dxd\tau \qquad \forall t \in\,]0,T]. \tag{3.36}$$

By Proposition 2.6 we then get (3.11) and

$$\iint_{\Omega_t} \varphi(\bar{u}_m(x,\tau),x)\,\theta(x)\, dxd\tau \to \iint_{\Omega_t} \varphi(u(x,\tau),x)\,\theta(x)\, dxd\tau \qquad \forall t \in\,]0,T], \tag{3.37}$$

$$\iint_{\Omega_t} \varphi^*(\bar{w}_m(x,\tau),x)\,\theta(x)\, dxd\tau \to \iint_{\Omega_t} \varphi^*(w(x,\tau),x)\,\theta(x)\, dxd\tau \qquad \forall t \in\,]0,T]. \tag{3.38}$$

Of course this also applies for $\theta \equiv 1$; by (3.29), (3.18), and (3.37), Proposition 2.8 yields

$$u_m \underset{2}{\to} u \qquad \text{in } L^2(\Omega_T), \tag{3.39}$$

whence

$$\iint_{\Omega_t} \bar{u}_m \bar{\vec{q}}_m \cdot \nabla\theta\, dxd\tau \to \iint_{\Omega_t} u\vec{q} \cdot \nabla\theta\, dxd\tau \qquad \forall t \in\,]0,T]. \tag{3.40}$$

In view of proving (3.9), let us now assume that θ is a nonnegative function of $\mathcal{D}(\Omega)$, multiply (3.13) by $u\theta$, and integrate in time. (We still imply the index ε.) By (3.11) $u\theta \in \partial\varphi^*(w)\theta$ a.e. in Ω_T, whence

$$\int_0^t \left\langle \frac{\partial w}{\partial t}, u\theta \right\rangle = \int_\Omega \big[\varphi^*(w(x,t),x) - \varphi^*(w^0(x),x)\big]\theta\, dx \qquad \text{for a.e. } t \in\,]0,T[. \tag{3.41}$$

We then get

$$\begin{aligned} &\int_\Omega \big[\varphi^*(w(x,t),x) - \varphi^*(w^0(x),x)\big]\theta\, dx - \iint_{\Omega_t} \vec{q} \cdot (\nabla u)\,\theta\, dxd\tau \\ &= \iint_{\Omega_t} (u\vec{q} \cdot \nabla\theta + fu\theta)\, dxd\tau + \int_0^t \langle g, u\theta\rangle\, d\tau \qquad \text{for a.e. } t \in\,]0,T[. \end{aligned} \tag{3.42}$$

Similarly, multiplying (3.27) by $\bar{u}_m\theta$ we have

$$\begin{aligned}&\int_\Omega \big(\varphi^*(\bar{w}_m(\cdot,t)) - \varphi^*(w^0)\big)\theta\, dx - \iint_{\Omega_t} \bar{\vec{q}}_m \cdot (\nabla \bar{u}_m)\theta\, dxd\tau \\ &\le \iint_{\Omega_t} (\bar{u}_m\bar{\vec{q}}_m \cdot \nabla\theta + \bar{f}_m\bar{u}_m\theta)\, dxd\tau + \int_0^t \langle \bar{g}_m, \bar{u}_m\theta\rangle\, d\tau \qquad \forall t \in\,]0,T].\end{aligned} \tag{3.43}$$

By (3.38), (3.40), (3.42), and (3.43), after a further integration in time we then obtain

$$\liminf_{m\to\infty} \int_0^T dt \iint_{\Omega_t} \bar{\vec{q}}_m \cdot (-\nabla \bar{u}_m)\theta\, dxd\tau \le \int_0^T dt \iint_{\Omega_t} \vec{q} \cdot (-\nabla u)\theta\, dxd\tau.$$

The constitutive relation (3.9) then follows from Theorem 2.10. Finally, by part (ii) of the same theorem we infer that

$$\nabla u_m \to \nabla u \qquad \text{in } L^2(\Omega_T)^3. \tag{3.44}$$

This and (3.39) yield (3.10). □

Remarks. (i) In view of the developments of the next section, notice that (3.30) and (3.31) yield the following uniform estimates for the solutions of Problem 3.1_ε (here written displaying the index ε):

$$\begin{aligned}&\|w_\varepsilon\|_{L^\infty(0,T;H)\cap H^1(0,T;V')}, \|u_\varepsilon\|_{L^2(0,T;H^1(\Omega))}, \|\vec{q}_\varepsilon\|_{L^2(\Omega_T)^3}, \|f_\varepsilon\|_{L^2(\Omega_T)} \\ &\le \text{Constant (independent of } \varepsilon).\end{aligned} \tag{3.45}$$

(ii) If f_ε is assumed to be independent of ∇u_ε, the assumption (3.3) may be dropped ($\vec{\alpha}$ is thus assumed only to be nondecreasing w.r.t. its first argument). □

4. Two-scale homogenized problem. In this section we show that as ε vanishes the solutions of Problem 3.1_ε converge to a solution of a two-scale homogenized problem.

Two-scale formulation. Let us assume that (2.8), (2.36), and (3.5) are fulfilled and that

$$w_\varepsilon^0 \underset{2}{\to} w^0 \qquad \text{in } L^2(\Omega \times \mathcal{Y}), \qquad z \in H^1(\Omega_T), \qquad g \in L^2(0,T;V'). \tag{4.1}$$

We shall deal with functions that depend on two (vector) space variables x and y and denote the respective gradient operators by ∇_x (resp., ∇_y). We shall use the notation (2.5), set

$$Z := \big\{v \in L^2\big(\Omega; H^1(\mathcal{Y})\big) : \hat{v} \in V\big\}, \tag{4.2}$$

and equip this space with the graph norm. Now we introduce the weak formulation of a two-scale problem that we then retrieve by letting ε vanish in Problem 3.1_ε.

PROBLEM 4.1. *Find*

$$\begin{aligned}&u \in L^2(0,T;V) + z, \qquad w \in L^2(\Omega_T \times \mathcal{Y}), \qquad \vec{q} \in L^2(\Omega_T \times \mathcal{Y})^3, \\ &u_\sharp \in L^2\big(\Omega_T; H^1(\mathcal{Y})\big), \qquad \vec{q}_\sharp \in L^2(\Omega_T \times \mathcal{Y})^3\end{aligned} \tag{4.3}$$

such that $\hat{u}_\sharp \equiv 0$, $\hat{\vec{q}}_\sharp \equiv \vec{0}$ *a.e. in* Ω_T *and, setting*

$$f = \beta(u, \nabla_x u + \nabla_y u_\sharp, x, y, t) \qquad \text{a.e. in } \Omega_T \times \mathcal{Y}, \tag{4.4}$$

one has

$$\vec{q} = \vec{\alpha}(-\nabla_x u - \nabla_y u_\sharp, u, x, y, t) \qquad \text{a.e. in } \Omega_T \times \mathcal{Y}, \tag{4.5}$$

$$w \in \partial\varphi(u, x, y) \qquad \text{a.e. in } \Omega_T \times \mathcal{Y}, \tag{4.6}$$

$$\begin{gathered}\iiint_{\Omega_T \times \mathcal{Y}} \left((w^0 - w)\,\frac{\partial v}{\partial t} - \hat{\vec{q}} \cdot \nabla_x \hat{v} - \vec{q}_\sharp \cdot \nabla_y v - fv \right) dxdydt = \int_0^T \langle g, v\rangle \, dt \\ \forall v \in H^1\big(0,T; L^2(\Omega \times \mathcal{Y})\big) \cap L^2(0,T;Z),\ v(\cdot,\cdot,T) = 0 \qquad \text{a.e. in } \Omega \times \mathcal{Y},\end{gathered} \tag{4.7}$$

$$\int_{\mathcal{Y}} \vec{q} \cdot \nabla_y v \, dy = 0 \qquad \forall v \in H^1(\mathcal{Y}) \text{ a.e. in } \Omega_T. \tag{4.8}$$

(It should be noticed that the function u is independent of y.)

Interpretation. Equation (4.7) yields the two-scale equation

$$\frac{\partial w}{\partial t} + \nabla_x \cdot \hat{\vec{q}} + \nabla_y \cdot \vec{q}_\sharp = f + g \qquad \text{in } Z' \text{ a.e. in }]0,T[\tag{4.9}$$

and the initial condition

$$w(\cdot,\cdot,0) = w^0 \qquad \text{in } Z', \tag{4.10}$$

for by comparing the terms of (4.9) we infer that $w \in H^1(0,T;Z')$. Conversely (4.9) and (4.10) entail (4.7). By selecting v independent of y in (4.7), we get the following system for the coarse-scale fields $\hat{w}, \hat{\vec{q}}$:

$$\begin{aligned}&\frac{\partial \hat{w}}{\partial t} + \nabla_x \cdot \hat{\vec{q}} = \hat{f} + g \qquad \text{in } V' \text{ a.e. in }]0,T[,\\ &\hat{w}(\cdot,0) = \widehat{w^0} \qquad \text{in } V'.\end{aligned} \tag{4.11}$$

Equations (4.9)–(4.11) then yield the fine-scale system

$$\begin{aligned}&\frac{\partial \tilde{w}}{\partial t} + \nabla_y \cdot \vec{q}_\sharp = \tilde{f} \qquad \text{in } L^2\big(\Omega; H^1(\mathcal{Y})'\big) \text{ a.e. in }]0,T[,\\ &\tilde{w}(\cdot,\cdot,0) = \widetilde{w^0} \qquad \text{in } V' \text{ for a.e. } y \in \mathcal{Y}.\end{aligned} \tag{4.12}$$

The two latter systems are coupled via the nonlinear constitutive equations (4.4)–(4.6). Equation (4.8) also reads $\nabla_y \cdot \vec{q} = 0$ in $\mathcal{D}'(\mathcal{Y})$ a.e. in Ω_T.

We now retrieve Problem 4.1 from Problem 3.1_ε by letting ε vanish.

THEOREM 4.1. *For any fixed $\varepsilon > 0$, let* (4.1) *and the hypotheses of Theorem* 3.1 *be satisfied. Let $\{(u_\varepsilon, w_\varepsilon) : \varepsilon > 0\}$ be a family of solutions of Problem 3.1_ε that also fulfils the uniform estimates* (3.45) *(we saw that such a family exists). Then there exist $u, w, \vec{q}, u_\sharp, \vec{q}_\sharp$ such that, as $\varepsilon \to 0$ along a suitable sequence,*

$$u_\varepsilon \rightharpoonup u \qquad \text{in } L^2\big(0,T;H^1(\Omega)\big), \tag{4.13}$$

$$\nabla u_\varepsilon \underset{2}{\rightharpoonup} \nabla_x u + \nabla_y u_\sharp \qquad \text{in } L^2(\Omega_T \times \mathcal{Y})^3, \tag{4.14}$$

$$w_\varepsilon \underset{2}{\overset{*}{\rightharpoonup}} w \qquad \text{in } L^\infty\big(0,T;L^2(\Omega \times \mathcal{Y})\big), \tag{4.15}$$

$$\vec{q}_\varepsilon \underset{2}{\rightharpoonup} \vec{q} \qquad \text{in } L^2(\Omega_T \times \mathcal{Y})^3, \tag{4.16}$$

$$\nabla \cdot \vec{q}_\varepsilon \underset{2}{\rightharpoonup} \nabla_x \cdot \hat{\vec{q}} + \nabla_y \cdot \vec{q}_\sharp \qquad \text{in } L^2(\Omega_T \times \mathcal{Y}). \tag{4.17}$$

This entails that $(u, w, \vec{q}, u_\sharp, \vec{q}_\sharp)$ *is a solution of Problem* 4.1.

Proof. This argument follows the lines of part (iii) of the proof of Theorem 3.1, obviously with two-scale in place of single-scale convergence. By the estimates (3.45) and Propositions 2.1, 2.3, and 2.4 there exist $u, w, \vec{q}, u_\sharp, \vec{q}_\sharp$ such that, as $\varepsilon \to 0$ along a suitable subsequence, (4.8) and (4.13)–(4.17) hold. By passing to the limit in (3.12), we then get (4.7).

In view of deriving the relations (4.4)–(4.6), let us now fix any nonnegative $\theta \in \mathcal{D}(\bar{\Omega})$, and notice that by (3.45), (4.15), and Proposition 2.2

$$w_\varepsilon \overset{*}{\rightharpoonup} \hat{w} \qquad \text{in } L^\infty(0,T;H) \cap H^1(0,T;V').$$

By the aforementioned compactness result of Aubin (see [8, 57, 73]), we then have

$$\begin{aligned}(4.18)\qquad &\iint_{\Omega_T} w_\varepsilon(x,t) u_\varepsilon(x,t)\,\theta(x)\,dxdt \\ &\to \iint_{\Omega_T} \hat{w}(x,t)u(x,t)\,\theta(x)\,dxdt = \iiint_{\Omega_T\times\mathcal{Y}} w(x,y,t)u(x,t)\,\theta(x)\,dxdydt.\end{aligned}$$

By (3.11) and Proposition 2.7, we then get (4.6) and, setting $\bar{\varphi}(v,x) := \int_{\mathcal{Y}} \varphi(v,x,y)\,dy$ for any v and a.e. x,

$$\begin{aligned}(4.19)\qquad &\iint_{\Omega_T} \varphi_\varepsilon(u_\varepsilon(x,t),x)\theta(x)\,dxdt \\ &\to \iiint_{\Omega_T\times\mathcal{Y}} \varphi(u(x,t),x,y)\theta(x)\,dxdydt = \iint_{\Omega_T} \bar{\varphi}(u(x,t),x)\theta(x)\,dxdt,\end{aligned}$$

$$(4.20)\qquad \iint_{\Omega_T} \varphi_\varepsilon^*(w_\varepsilon(x,t),x)\theta(x)\,dxdt \to \iiint_{\Omega_T\times\mathcal{Y}} \varphi^*(w(x,y,t),x,y)\theta(x)\,dxdydt.$$

This also applies for $\theta \equiv 1$. As (3.18) entails that the mapping $v \mapsto \bar{\varphi}(v,x)$ is strictly convex for a.e. x, Proposition 2.9 then yields $u_\varepsilon \underset{2}{\to} u$ in $L^2(\Omega_T \times \mathcal{Y})$; that is, as u is independent of y,

$$(4.21)\qquad u_\varepsilon \to u \qquad \text{in } L^2(\Omega_T).$$

In view of the proof of (4.5) let us now assume that θ is a nonnegative function of $\mathcal{D}(\Omega)$; notice that $\vec{q}_\varepsilon \rightharpoonup \hat{\vec{q}}$ in $L^2(\Omega_T)^3$ and by (4.21)

$$(4.22)\qquad \iint_{\Omega_t} u_\varepsilon \vec{q}_\varepsilon \cdot \nabla_x \theta\,dxd\tau \to \iint_{\Omega_t} u\hat{\vec{q}} \cdot \nabla_x\theta\,dxd\tau \qquad \forall t \in\,]0,T].$$

Let us now select $v = u\theta$ in (4.7) and integrate in time. As by (4.6)

(4.23)

$$\iint_{\mathcal{Y}_t} \left\langle \frac{\partial w}{\partial t}, u\theta \right\rangle dyd\tau = \iint_{\Omega\times\mathcal{Y}} \left[\varphi^*(w(x,y,t),x,y) - \varphi^*(w^0(x,y),x,y)\right]\theta\,dxdy \qquad \text{for a.e. } t \in\,]0,T[,$$

we thus get

(4.24)

$$\begin{aligned}&\iint_{\Omega\times\mathcal{Y}} \left[\varphi^*(w(x,y,t),x,y) - \varphi^*(w^0(x,y),x,y)\right]\theta\,dxdy - \iiint_{\Omega_t\times\mathcal{Y}} \vec{q}\cdot(\nabla_x u)\theta\,dxdyd\tau \\ &= \iint_{\Omega_t} \left(u\hat{\vec{q}}\cdot\nabla_x\theta + fu\theta\right) dxd\tau + \int_0^t \langle g, u\theta\rangle\,d\tau \qquad \text{for a.e. } t \in\,]0,T[.\end{aligned}$$

Let us also rewrite (3.42) displaying the index ε,

$$
\begin{aligned}
&\int_\Omega \big[\varphi^*_\varepsilon(w_\varepsilon(x,t),x) - \varphi^*(w^0_\varepsilon(x),x)\big]\theta(x)\,dx - \iint_{\Omega_t} \vec{q}_\varepsilon \cdot (\nabla u_\varepsilon)\,\theta\,dxd\tau \\
&= \iint_{\Omega_t} (u_\varepsilon \vec{q}_\varepsilon \cdot \nabla\theta + f_\varepsilon u_\varepsilon \theta)\,dxd\tau + \int_0^t \langle g, u_\varepsilon\theta\rangle\,d\tau \qquad \text{for a.e. } t\in\,]0,T[,
\end{aligned}
\tag{4.25}
$$

and notice that by (4.1) $\int_\Omega \varphi^*_\varepsilon(w^0_\varepsilon)\theta\,dx \to \iint_{\Omega\times\mathcal{Y}} \varphi^*(w^0)\theta\,dxdy$. After a further integration in time (4.20)–(4.25) yield

$$
\begin{aligned}
&\int_0^T dt \iint_{\Omega_t} \vec{q}_\varepsilon \cdot (\nabla u_\varepsilon)\theta\,dxd\tau \\
&= \int_0^T dt \int_\Omega \big[\varphi^*(w(x,y,t),x,y) - \varphi^*(w^0(x,y),x,y)\big]\theta(x)\,dxdy \\
&- \int_0^T dt \iint_{\Omega_t} (u_\varepsilon \vec{q}_\varepsilon \cdot \nabla\theta + f_\varepsilon u_\varepsilon\theta)\,dxd\tau - \int_0^T dt \int_0^t \langle g, u_\varepsilon\theta\rangle\,d\tau \\
&\to \int_0^T dt \int_\Omega \big[\varphi^*(w(x,t),x) - \varphi^*(w^0(x),x)\big]\theta(x)\,dx - \int_0^T dt \iint_{\Omega_t} (u\vec{q}\cdot\nabla\theta + fu\theta)\,dxd\tau \\
&- \int_0^T dt \int_0^t \langle g, u\theta\rangle\,d\tau = \int_0^T dt \iiint_{\Omega_t\times\mathcal{Y}} \vec{q}\cdot(\nabla_x u)\theta\,dxdyd\tau;
\end{aligned}
$$

that is, as by (4.8) $\int_{\mathcal{Y}} \vec{q}\cdot\nabla_y u_\sharp\,dy = 0$ a.e. in Ω_T,

$$
\int_0^T dt \iint_{\Omega_t} \vec{q}_\varepsilon \cdot (\nabla u_\varepsilon)\theta\,dxd\tau \to \int_0^T dt \iiint_{\Omega_t\times\mathcal{Y}} \vec{q}\cdot(\nabla_x u + \nabla_y u_\sharp)\theta\,dxdyd\tau.
$$

Proposition 2.11 then yields (4.5). Finally, by part (ii) of Proposition 2.11 we infer that

$$
\nabla u_\varepsilon \underset{2}{\to} \nabla_x u + \nabla_y u\sharp \qquad \text{in } L^2(\Omega_T\times\mathcal{Y})^3. \tag{4.26}
$$

This convergence and (4.21) yield (4.4). $\square$

Remark. Similarly to what we saw for Theorem 3.1, if f_ε were assumed to be independent of ∇u_ε, the assumption (3.3) might be dropped. $\square$

5. Coarse-scale homogenized problem. In this section we reduce the two-scale problem of the last section to an equivalent coarse-scale formulation under some minor restrictions. This will complete the homogenization procedure.

Here we assume that the source term f is independent of u and ∇u and that the maximal monotone function $\vec{\alpha}$ is the subdifferential $\partial\psi(\cdot,v,x,y,t)$ of a function $\psi : \mathbf{R}^3\times\mathbf{R}\times\Omega\times\mathcal{Y}\times\,]0,T[\,\to\mathbf{R}$ such that

(5.1)

$$
\begin{aligned}
&\psi \text{ is measurable w.r.t. either } \mathcal{B}(\Omega)\otimes\mathcal{L}(\mathbf{R}^3\times\mathbf{R}\times\mathcal{Y}_T) \text{ or } \mathcal{L}(\mathbf{R}^3\times\mathbf{R}\times\Omega_T)\otimes\mathcal{B}(\mathcal{Y}),\\
&\psi(\vec{\xi},\cdot,x,y,t) \text{ is continuous} \qquad \forall\vec{\xi} \text{ and for a.e. } (x,y,t),\\
&\psi(\cdot,v,x,y,t) \text{ is lower semicontinuous and convex} \qquad \forall v \text{ and for a.e. } (x,y,t),
\end{aligned}
$$

$$
\begin{aligned}
&\exists L>0,\ \exists \ell_4\in L^2(\Omega_T) : \forall(\vec{\xi},v) \qquad \text{for a.e. } (x,y,t),\\
&\forall\vec{\eta}\in\partial\psi(\vec{\xi},v,x,y,t), \qquad |\vec{\eta}| \le L|\vec{\xi}| + L|v| + \ell_4(x,t),
\end{aligned}
\tag{5.2}
$$

$$\begin{aligned}&\exists c_4 > 0,\, \exists h_4 \in L^1(\Omega_T) : \forall(\vec{\xi}, v) \qquad \text{for a.e. } (x,y,t),\\ &\psi(\vec{\xi}, v, x, y, t) \geq c_4|\vec{\xi}|^2 + h_4(x,t).\end{aligned} \tag{5.3}$$

We define the function φ as in section 3, assume (3.1)–(3.6), and define the homogenized potentials φ_0 and ψ_0:

$$\varphi_0(v,x) := \int_{\mathcal{Y}} \varphi(v,x,y)\,dy \qquad \forall v \in \mathbf{R} \text{ for a.e. } x \in \Omega, \tag{5.4}$$

$$\begin{aligned}\psi_0(\vec{\xi}, v, x, t) := \inf\Bigg\{ &\int_{\mathcal{Y}} \psi(\vec{\xi} + \vec{\xi}_1(y), v, x, y, t)\,dy :\\ &\vec{\xi}_1 \in L^2(\mathcal{Y})^3,\, \hat{\vec{\xi}}_1 = \vec{0},\, \nabla_y \times \vec{\xi}_1 = \vec{0} \text{ in } \mathcal{D}'(\mathcal{Y})^3 \Bigg\}\\ &\forall \vec{\xi} \in \mathbf{R}^3,\, \forall v \in \mathbf{R} \qquad \text{for a.e. } (x,t) \in \Omega_T.\end{aligned} \tag{5.5}$$

This minimization problem is equivalent to the corresponding Euler equation

$$\begin{cases}\theta \in H^1(\mathcal{Y})^3,\\ \nabla_y \cdot \partial\psi(\vec{\xi} + \nabla_y\theta(y), v, x, y, t) \ni 0 \quad \text{in } H^{-1}(\mathcal{Y})^3\end{cases} \qquad \text{for a.e. } (x,t) \in \Omega_T. \tag{5.6}$$

In the case of a quadratic function ψ this family of *cell problems* will be studied in the next section.

Here we assume that

$$w^0 \in L^2(\Omega \times \mathcal{Y}), \qquad z \in H^1(\Omega_T), \qquad f \in L^2(\Omega_T \times \mathcal{Y}), \qquad g \in L^2(0,T;V') \tag{5.7}$$

and introduce a coarse-scale problem using the notation (2.5).

PROBLEM 5.1. *Find $u \in L^2(0,T;V) + z$, $\bar{w} \in L^2(\Omega_T)$, and $\bar{\vec{q}} \in L^2(\Omega_T)^3$ such that*

$$\bar{w}(x,t) \in \partial\varphi_0(u(x,t), x) \qquad \textit{for a.e. } (x,t) \in \Omega_T, \tag{5.8}$$

$$\bar{\vec{q}}(x,t) \in \partial\psi_0(-\nabla u(x,t), u(x,t), x, t) \qquad \textit{for a.e. } (x,t) \in \Omega_T, \tag{5.9}$$

$$\begin{aligned}&\iint_{\Omega_T} \left(\left(\bar{w} - \widehat{w^0}\right) \frac{\partial v}{\partial t} - \bar{\vec{q}} \cdot \nabla v - \hat{f} v \right) dxdt = \int_0^T \langle g, v \rangle\, dt\\ &\forall v \in H^1(0,T;H) \cap L^2(0,T;V),\ v(\cdot,T) = 0 \qquad \textit{a.e. in } \Omega.\end{aligned} \tag{5.10}$$

This formulation may be compared with Problem 3.1_ε. In particular (5.10) yields the energy balance equation in the form

$$\frac{\partial \bar{w}}{\partial t} + \nabla \cdot \bar{\vec{q}} = \hat{f} + g \qquad \text{in } \mathcal{D}'(\Omega_T). \tag{5.11}$$

In view of relating Problems 4.1 and 5.1 to each other, first we review some basic properties of subdifferentials and convex conjugate functions that will be used in what follows.

LEMMA 5.1. *Let B be a real Banach space and $\langle \cdot, \cdot \rangle$ be the duality pairing between B and its dual B'. If $F : B \to \mathbf{R} \cup \{+\infty\}$ is such that $F \not\equiv +\infty$, then for any $u \in B$ and any $w \in B'$,*

$$F(u) + F^*(w) \geq \langle w, u \rangle, \tag{5.12}$$

$$w \in \partial F(u) \quad \Leftrightarrow \quad F(u) + F^*(w) \leq \langle w, u \rangle, \tag{5.13}$$

$$w \in \partial F(u) \quad \Leftrightarrow \quad F(u) + F^*(w) = \langle w, u \rangle. \tag{5.14}$$

The inequality (5.12) directly follows from the definition of the convex conjugate function F^*. The proof of (5.13) may be found, e.g., in [45, 51, 54, 69]. The statement (5.14) is an obvious consequence of (5.12) and (5.13).

The next lemma will allow us to upscale from Problem 4.1 to 5.1 and may be compared with classical results about the homogenization of integral functionals of [24, 60].

LEMMA 5.2. *Let the function* ψ *fulfil* (5.1)–(5.3), *define* ψ_0 *as in* (5.5), *and let* $\nu \in L^2(\Omega_T)$. *If* $\vec{\xi}, \vec{\eta} \in L^2(\Omega_T \times \mathcal{Y})^3$ *are such that*

$$\vec{\eta}(x,y,t) \in \partial\psi(\vec{\xi}(x,y,t), \nu(x,t), x, y, t) \qquad \text{for a.e. } (x,y,t) \in \Omega_T \times \mathcal{Y}, \tag{5.15}$$

$$\nabla_y \times \vec{\xi} = \vec{0} \qquad \text{in } \mathcal{D}'(\Omega_T \times \mathcal{Y})^3, \qquad \nabla_y \cdot \vec{\eta} = 0 \qquad \text{in } \mathcal{D}'(\Omega_T \times \mathcal{Y}), \tag{5.16}$$

then

$$\hat{\vec{\eta}}(x,t) \in \partial\psi_0(\hat{\vec{\xi}}(x,t), \nu(x,t), x, t) \qquad \text{for a.e. } (x,t) \in \Omega_T. \tag{5.17}$$

Proof. As here the independent variables $(x,t) \in \Omega_T$ just occur as parameters, for the sake of simplicity we shall omit them as well as $\nu(x,t)$ throughout this argument. For instance, we thus replace (5.15) by

$$\vec{\eta}(y) \in \partial\psi(\vec{\xi}(y), y) \qquad \text{for a.e. } y \in \mathcal{Y}. \tag{5.18}$$

Using the notation (2.5), let us first set

$$\begin{aligned} W &:= \{\vec{w} \in L^2(\mathcal{Y})^3 : \nabla \cdot \vec{w} = 0 \quad \text{in } \mathcal{D}'(\mathcal{Y})\}, \qquad W_* := \{\vec{v} \in W : \hat{\vec{v}} = \vec{0}\}, \\ Z &:= \{\vec{v} \in L^2(\mathcal{Y})^3 : \nabla \times \vec{v} = \vec{0} \quad \text{in } \mathcal{D}'(\mathcal{Y})^3\}, \qquad Z_* := \{\vec{v} \in Z : \hat{\vec{v}} = \vec{0}\}; \end{aligned}$$

these are Hilbert subspaces of $L^2(\mathcal{Y})^3$. Notice that $Z_* = \nabla H^1_*(\mathcal{Y})$, and this is the orthogonal space of W_* in $L^2(\mathcal{Y})^3$. Hence

$$\begin{aligned} &\int_{\mathcal{Y}} \vec{v}(y) \cdot \vec{w}(y)\, dy - \hat{\vec{v}} \cdot \hat{\vec{w}} = \int_{\mathcal{Y}} [\hat{\vec{v}} \cdot \tilde{\vec{w}}(y) + \tilde{\vec{v}}(y) \cdot \hat{\vec{w}} + \tilde{\vec{v}}(y) \cdot \tilde{\vec{w}}(y)]\, dy \\ &= \hat{\vec{v}} \cdot \int_{\mathcal{Y}} \tilde{\vec{w}}(y)\, dy + \left(\int_{\mathcal{Y}} \tilde{\vec{v}}(y)\, dy\right) \cdot \hat{\vec{w}} + \int_{\mathcal{Y}} \tilde{\vec{v}}(y) \cdot \tilde{\vec{w}}(y)\, dy = 0 \qquad \forall (\vec{v}, \vec{w}) \in Z \times W. \end{aligned} \tag{5.19}$$

By (5.15)

$$\int_{\mathcal{Y}} \vec{\eta}(y) \cdot [\vec{\xi}(y) - \vec{v}(y)]\, dy \geq \int_{\mathcal{Y}} \psi(\vec{\xi}(y), y)\, dy - \int_{\mathcal{Y}} \psi(\vec{v}(y), y)\, dy \qquad \forall \vec{v} \in L^2(\mathcal{Y})^3;$$

hence, applying (5.19) and noticing that $\int_{\mathcal{Y}} \psi(\vec{\xi}(y), y)\, dy \geq \psi_0(\hat{\vec{\xi}})$,

$$\hat{\vec{\eta}} \cdot (\hat{\vec{\xi}} - \hat{\vec{v}}) = \int_{\mathcal{Y}} \vec{\eta}(y) \cdot [\vec{\xi}(y) - \vec{v}(y)]\, dy \geq \psi_0(\hat{\vec{\xi}}) - \int_{\mathcal{Y}} \psi(\vec{v}(y), y)\, dy \qquad \forall \vec{v} \in Z.$$

By the arbitrariness of $\tilde{\vec{v}}$, we then get $\hat{\vec{\eta}} \cdot (\hat{\vec{\xi}} - \hat{\vec{v}}) \geq \psi_0(\hat{\vec{\xi}}) - \psi_0(\hat{\vec{v}})$ for any $\hat{\vec{v}} \in \mathbf{R}^3$, that is, $\hat{\vec{\eta}} \in \partial\psi_0(\hat{\vec{\xi}})$. ∎

The next statement inverts Lemma 5.2 and will be used to downscale from Problem 5.1 to 4.1.

LEMMA 5.3. *Let the function ψ fulfil* (5.1)–(5.3), *define ψ_0 as in* (5.5), *and let $\nu \in L^2(\Omega_T)$. If*

(5.20)
$$\vec r, \vec s \in L^2(\Omega_T)^3, \qquad \vec s(x,t) \in \partial\psi_0(\vec r(x,t), \nu(x,t), x, t) \qquad \textit{for a.e. } (x,t) \in \Omega_T,$$

then there exist $\vec\xi, \vec\eta \in L^2(\Omega_T \times \mathcal{Y})^3$ such that

$$\hat{\vec\xi}(x,t) = \vec r(x,t), \qquad \hat{\vec\eta}(x,t) = \vec s(x,t) \qquad \textit{for a.e. } (x,t) \in \Omega_T,$$
$$\vec\eta(x,y,t) \in \partial\psi(\vec\xi(x,y,t), \nu(x,t), x, y, t) \qquad \textit{for a.e. } (x,y,t) \in \Omega_T \times \mathcal{Y}. \tag{5.21}$$

If, moreover, $\psi(\cdot, v, x, y, t)$ and $\psi^*(\cdot, v, x, y, t)$ are strictly convex for any $v \in \mathbf{R}$ and a.e. $(x,y,t) \in \Omega_T \times \mathcal{Y}$, then the pair $(\vec\xi, \vec\eta)$ is unique.

Proof. Here we imply the dependence on (x,t) and $\nu(x,t)$, as we did in the proof of Lemma 5.2. Let us first set

$$\rho_0(\vec\eta) := \inf\left\{\int_{\mathcal{Y}} \psi^*(\vec\eta + \vec z(y), y)\, dy : \vec z \in Z\right\} \qquad \forall \vec\eta \in \mathbf{R}^3. \tag{5.22}$$

By (5.3) the conjugate function ψ^* also has quadratic growth at infinity, so that

$$\psi(\vec\lambda, y) \to +\infty, \qquad \psi^*(\vec\lambda, y) \to +\infty \qquad \text{as } |\vec\lambda| \to +\infty, \text{ uniformly for } y \in \mathcal{Y}. \tag{5.23}$$

By the convexity and the lower semicontinuity of ψ and ψ^*, the infima $\psi_0(\hat{\vec\xi})$ and $\rho_0(\hat{\vec\eta})$ (cf. (5.4) and (5.22)) are then both attained, that is,

$$\exists \vec\xi \in \mathbf{R}^N + W: \quad \hat{\vec\xi} = r, \qquad \psi_0(\hat{\vec\xi}) = \int_{\mathcal{Y}} \psi(\vec\xi(y), y)\, dy,$$
$$\exists \vec\eta \in \mathbf{R}^N + Z: \quad \hat{\vec\eta} = s, \qquad \rho_0(\hat{\vec\eta}) = \int_{\mathcal{Y}} \psi^*(\vec\eta(y), y)\, dy. \tag{5.24}$$

By (5.13) the inclusion $\vec s \in \partial\psi_0(\vec r)$ is thus equivalent to

$$\hat{\vec\xi} \cdot \hat{\vec\eta} \geq \psi_0(\hat{\vec\xi}) + \psi_0^*(\hat{\vec\eta}). \tag{5.25}$$

The definitions of ψ_0 and ρ_0, respectively, yield $\psi_0 \leq \psi$ and $\rho_0 \leq \psi^*$ in $\mathbf{R}^3 \times \mathcal{Y}$; the former inequality entails that $\psi^* \leq \psi_0^*$. Thus $\rho_0 \leq \psi^* \leq \psi_0^*$, namely, $\rho_0 \leq \psi_0^*$ in $\mathbf{R}^3$. The inequality (5.25) then yields $\hat{\vec\xi} \cdot \hat{\vec\eta} \geq \psi_0(\hat{\vec\xi}) + \rho_0(\hat{\vec\eta})$, which by (5.19) and (5.24) reads

$$\int_{\mathcal{Y}} \vec\xi(y) \cdot \vec\eta(y)\, dy \geq \int_{\mathcal{Y}} [\psi(\vec\xi(y), y) + \psi^*(\vec\eta(y), y)]\, dy.$$

By (5.12), on the other hand, $\vec\xi(y) \cdot \vec\eta(y) \leq \psi(\vec\xi(y), y) + \psi^*(\vec\eta(y), y)$ for a.e. $y \in \mathcal{Y}$. By the two latter inequalities, we infer that

$$\vec\xi(y) \cdot \vec\eta(y) = \psi(\vec\xi(y), y) + \psi^*(\vec\eta(y), y) \qquad \text{for a.e. } y \in \mathcal{Y},$$

and by (5.14) this is tantamount to $\vec\eta(y) \in \partial\psi(\vec\xi(y), y)$ for a.e. $y \in \mathcal{Y}$. The final statement about uniqueness is obvious. ∎

Next by Lemmata 5.2 and 5.3 we show that Problems 4.1 and 5.1 are equivalent.

THEOREM 5.4. *Let φ and $\vec{\alpha} := \partial\psi$ fulfil* (2.8) *and* (2.36). *Let* (3.1)–(3.4), (3.6), (5.1)–(5.5), *and* (5.7) *also be satisfied (with β independent of v and $\vec{z}$). Then we have the following:*

(i) *If $(u, w, \vec{q}, u_\sharp, \vec{q}_\sharp)$ is a solution of Problem* 4.1 *(by Theorem* 4.1 *such a solution exists), then $(u, \hat{w}, \hat{\vec{q}})$ is a solution of Problem* 5.1.

(ii) *Conversely, for any solution $(u, \bar{w}, \bar{\vec{q}})$ of Problem* 5.1 *there exists a solution $(u, w, \vec{q}, u_\sharp, \vec{q}_\sharp)$ of Problem* 4.1 *such that $\hat{w} = \bar{w}$ and $\hat{\vec{q}} = \bar{\vec{q}}$ a.e. in Ω_T.*

Proof. (i) The inclusion (5.9) follows from (4.5) and Lemma 5.2, because of (4.8). Equation (5.8) is an even simpler consequence of (4.6), for u is independent of y. Indeed (4.6) is tantamount to

$$\iiint_{\Omega_T\times\mathcal{Y}} [\varphi(u(x,t),x,y) - u(x,t)w(x,y,t)]\,dxdydt$$
$$\le \iiint_{\Omega_T\times\mathcal{Y}} [\varphi(v(x,y,t),x,y) - v(x,y,t)w(x,y,t)]\,dxdydt \qquad \forall v \in L^2(\Omega_T\times\mathcal{Y});$$

selecting v independent of y we then get

$$\iint_{\Omega_T} [\varphi_0(u(x,t),x) - u\hat{w}]\,dxdt \le \iint_{\Omega_T} [\varphi_0(v(x,t),x) - v\hat{w}]\,dxdt \qquad \forall v \in L^2(\Omega_T),$$

that is, (5.8).

(ii) This part is a direct consequence of Lemma 5.3. ◻

We can now state our single-scale homogenization result.

COROLLARY 5.5. *Under the hypotheses of Theorems* 4.1 *and* 5.4 *there exist $u, \bar{w}, \bar{\vec{q}}$ such that, as $\varepsilon \to 0$ along a suitable sequence,*

$$u_\varepsilon \rightharpoonup u \qquad \text{in } L^2\big(0,T;H^1(\Omega)\big), \tag{5.26}$$

$$w_\varepsilon \overset{*}{\underset{2}{\rightharpoonup}} \bar{w} \qquad \text{in } L^\infty(0,T;H), \tag{5.27}$$

$$\vec{q}_\varepsilon \underset{2}{\rightharpoonup} \bar{\vec{q}} \qquad \text{in } L^2(\Omega_T)^3. \tag{5.28}$$

This entails that $(u, \bar{w}, \bar{\vec{q}})$ is a solution of Problem 5.1.

Proof. By Proposition 2.2, (4.13), (4.15), and (4.16) yield (5.26)–(5.28). It then suffices to apply part (i) of Theorem 5.4. ◻

6. Single nonlinearity, Kirchhoff transformation, and discussion. Obviously the analysis of sections 3, 4, and 5 can be much simplified whenever either of the two constitutive relations (1.2) and (1.3) is linear, for each of Problems 3.1_ε, 4.1, and 5.1 is then reduced to a single variational inequality, so that existence of a solution can be established without using compactness properties. In that case these problems are actually well-posed.

Here we assume that (1.3) is linear w.r.t. ∇u but not w.r.t. u, and we retrieve the homogenized conductivity tensor from the two-scale relation (4.5) via a well-known procedure. Afterwards we show how a class of nonlinear conduction laws may be reduced to linear form via a generalization of the classical Kirchhoff transformation. So let

$$\vec{\alpha}(\vec{z}, v, x, y, t) = K(v, x, y, t)\cdot\vec{z} \qquad \forall(\vec{z}, v) \in \mathbf{R}^3\times\mathbf{R} \text{ for a.e. } (x,y,t) \in \Omega_T\times\mathcal{Y}, \tag{6.1}$$

where K is a (possibly asymmetric) positive-definite 3×3-tensor function with entries k_{ij} that are Caratheodory functions. The relation (4.5) and (4.8) then read, respectively,

$$\vec{q}(x,y,t) = -K(u(x,t),x,y,t) \cdot [\nabla_x u(x,t) + \nabla_y u_\sharp(x,y,t)] \qquad \text{for a.e. } (x,y,t) \in \Omega_T \times \mathcal{Y}, \tag{6.2}$$

$$\nabla_y \cdot \big\{ K(u(x,t),x,y,t) \cdot [\nabla_x u(x,t) + \nabla_y u_\sharp(x,y,t)] \big\} = 0 \qquad \text{in } H^{-1}(\mathcal{Y}) \qquad \text{for a.e. } (x,t) \in \Omega_T. \tag{6.3}$$

We claim that (6.2) and (6.3) entail a homogenized relation of the form

$$\hat{\vec{q}}(x,t) = -K_0(u(x,t),x,t) \cdot \nabla_x u(x,t) \qquad \text{for a.e. } (x,t) \in \Omega_T, \tag{6.4}$$

where the tensor function K_0 is determined as follows. Along the lines of [2] and consistently with a classical approach to the homogenization of linear elliptic and parabolic equations (see, e.g., the works quoted in the introduction), we search for a family of auxiliary scalar functions $w_1(x,y,t), \dots, w_3(x,y,t)$ such that

$$u_\sharp(x,y,t) = \sum_{j=1,2,3} \frac{\partial u}{\partial x_j}(x,t) w_j(x,y,t) \qquad \text{for a.e. } (x,y,t) \in \Omega_T \times \mathcal{Y}.$$

Denoting by $\vec{e}_j$ the unit vector of the axis y_j ($j = 1,2,3$), (6.2) and (6.3) then read, respectively,

$$\begin{aligned} &\vec{q}(x,y,t) = -K(u(x,t),x,y,t) \cdot \sum_{j=1,2,3} \frac{\partial u}{\partial x_j}(x,t)[\vec{e}_j + \nabla_y w_j(x,y,t)] \\ &\qquad\qquad\qquad\qquad \text{for a.e. } (x,y,t) \in \Omega_T \times \mathcal{Y}, \\ &\nabla_y \cdot \left\{ K(u(x,t),x,y,t) \cdot \sum_{j=1,2,3} \frac{\partial u}{\partial x_j}(x,t)[\vec{e}_j + \nabla_y w_j(x,y,t)] \right\} = 0 \\ &\qquad\qquad\qquad\qquad \text{in } H^{-1}(\mathcal{Y}) \qquad \text{for a.e. } (x,t) \in \Omega_T, \end{aligned} \tag{6.5}$$

i.e., $\frac{\partial}{\partial y_\ell}\big\{k_{\ell m} \frac{\partial u}{\partial x_j}(\delta_{jm} + \frac{\partial w_j}{\partial y_m})\big\} = 0$, assuming that repeated indices are summed from 1 to 3 and denoting by $\{\delta_{jm}\}$ the 3×3 identity tensor. This equation is fulfilled whenever $w_1, \dots, w_3$ solve the following three families of *cell problems*:

$$\begin{aligned} &\nabla_y \cdot \big\{ K(u(x,t),x,y,t) \cdot [\vec{e}_j + \nabla_y w_j(x,y,t)] \big\} = 0 \\ &\text{in } H^{-1}(\mathcal{Y}) \qquad \text{for a.e. } (x,t) \in \Omega_T \text{ for } j = 1,2,3, \end{aligned} \tag{6.6}$$

i.e., $\frac{\partial}{\partial y_\ell}\big\{k_{\ell m}(\delta_{jm} + \frac{\partial w_j}{\partial y_m})\big\} = 0$. These linear elliptic problems have one and only one solution $w_1, \dots, w_3 \in L^2\big(\Omega_T; H^1(\mathcal{Y})\big)$ such that $\int_{\mathcal{Y}} w_j(\cdot,\cdot,y)dy = 0$ a.e. in Ω_T for $j = 1,2,3$. By integrating (6.5) over $\mathcal{Y}$, we then see that (6.4) is fulfilled if we define the tensor function K_0 with entries

$$\begin{aligned} k_{0ij}(v,x,t) &:= \int_{\mathcal{Y}} \vec{e}_i \cdot K(v,x,y,t) \cdot [\vec{e}_j + \nabla_y w_j(x,y,t)]\, dy \\ &= \int_{\mathcal{Y}} [\vec{e}_i + \nabla_y w_i(x,y,t)] \cdot K(v,x,y,t) \cdot [\vec{e}_j + \nabla_y w_j(x,y,t)]\, dy \\ &= \int_{\mathcal{Y}} \left[\delta_{i\ell} + \frac{\partial w_i}{\partial y_\ell}(x,y,t)\right] k_{\ell m}(v,x,y,t) \left[\delta_{jm} + \frac{\partial w_j}{\partial y_m}(x,y,t)\right] dy \end{aligned} \tag{6.7}$$

(still with the sum convention) for any $v \in \mathbf{R}$, a.e. $(x,t) \in \Omega_T$, and $i,j = 1,2,3$.

Next let us consider two special cases. If $K(v,x,y,t)$ might be factorized in the form

$$K(v,x,y,t) = K_1(v) \cdot K_2(x,y,t) \qquad \forall v \in \mathbf{R} \text{ for a.e. } (x,y,t) \in \Omega_T \times \mathcal{Y}, \tag{6.8}$$

then the problems (6.6) would be reduced to three families of cell problems that are independent of the function u:

$$\begin{aligned} &\nabla_y \cdot \big\{ K_2(x,y,t) \cdot [\vec{e}_j + \nabla_y w_j(x,y,t)] \big\} = 0 \\ &\text{in } H^{-1}(\mathcal{Y}) \qquad \text{for a.e. } (x,t) \in \Omega_T \text{ for } j = 1,2,3. \end{aligned} \tag{6.9}$$

Finally, if

$$K(v,x,y,t) = K_3(v,x,t) \cdot K_4(y) \qquad \forall v \in \mathbf{R} \text{ for a.e. } (x,y,t) \in \Omega_T \times \mathcal{Y}, \tag{6.10}$$

then the w_j's would be independent of (x,t) and the problems (6.6) would be reduced to three linear problems in the cell $\mathcal{Y}$:

$$\nabla_y \cdot \big\{ K_4(y) \cdot [\vec{e}_j + \nabla_y w_j(y)] \big\} = 0 \qquad \text{in } H^{-1}(\mathcal{Y}) \text{ for } j = 1,2,3. \tag{6.11}$$

The reader will notice that the problems (6.9) and (6.11) are independent of the solution u, at variance with (6.6). Moreover, the three linear problems (6.11) do not contain the parameters (x,t), at variance with (6.9). One may thus solve them once for all and then use the w_j's to construct the homogenized conductivity tensor K_0 via (6.7).

Remarks. (i) The tensor K_0 is symmetric whenever K is; however, the nondiagonal elements of K_0 need not vanish even if K is a diagonal tensor. An anisotropic medium may thus be the outcome of the homogenization of an isotropic one: this is quite natural, for the fine-scale arrangement need not be isotropic.

(ii) The tensor K need not be symmetric; namely, the function $\vec{v} \mapsto K \cdot \vec{v}$ need not be cyclically monotone. Nevertheless, Theorem 5.4 and Corollary 5.5 hold also in this case, with (6.4) and (6.7) in place of (5.9) in Problem 5.1.

(iii) If K is independent of the temperature, then Problem 3.1_ε is reduced to the weak formulation of the multidimensional extension of the Stefan problem. The same applies to the single-scale homogenized Problem 5.1 if (5.9) is replaced by (6.4) and (6.7).

(iv) If $K = k(y)I$ (k being a positive scalar function), we thus retrieve the results of [38].

The Kirchhoff transformation. For an isotropic and homogeneous material in the Fourier law (1.8), we have $K(u,x) = k(u)I$. In this case a simple but powerful tool for handling the heat balance equation is provided by the classical *Kirchhoff transformation*

$$\mathcal{K} : u \mapsto U := \int_0^u k(\xi)\, d\xi \tag{6.12}$$

that reduces (1.8) to a linear relation: $\vec{q} = -\nabla U$. Because of the strict monotonicity of the operator $\mathcal{K}$, (1.2) may be rewritten in the form $w \in \partial\tilde{\varphi}(U,x)$, where $\tilde{\varphi}(\cdot,x)$ is strictly convex whenever $\varphi(\cdot,x)$ is. One might then equivalently formulate the system (1.1)–(1.3) in terms of the triplet $(U,w,\vec{q})$ and thus with a single nonlinearity. However, in this paper we were concerned with an inhomogeneous material and hence with a space-dependent conductivity: $k = k(u, x)$. In this case if both k and $\nabla_x k(\cdot, x)$ were uniformly bounded, then one might still deal with a weak solution via the above transformation. But the boundedness of $\nabla_x k(\cdot, x)$ would exclude composite materials, which were our main concern; this author does not see how the Kirchhoff transformation might be used here without severe restrictions.

This transformation may also be extended to a class of anisotropic materials as follows. If

$$K(u,x) = G(x)h(u), \quad \text{i.e.,} \quad k_{ij}(u,x) = g_{ij}(x)h(u) \qquad \text{for } j = 1,2,3, \tag{6.13}$$

with $G = \{g_{ij}\}$ a positive-definite tensor function and h a positive real function, then

$$\vec{q} = -G(x)h(u) \cdot \nabla u = -G(x) \cdot \nabla V, \quad \text{where} \quad V := \int_0^u h(\xi)\, d\xi. \tag{6.14}$$

Equation (1.1) then reads

$$\frac{\partial w}{\partial t} - \nabla \cdot [G(x) \cdot \nabla V] = f \qquad \text{in } \mathcal{D}'(\Omega_T), \tag{6.15}$$

and V and w are related by a subdifferential condition analogous to (1.2), so that the developments of this paper are easily extended. This latter setting fits the framework that was already addressed in [17, 38].

Phase-dependent conductivity. The latter formulation also applies if the function h is discontinuous. As the phase depends discontinuously on the temperature, this approach may also be used in the physically relevant case in which the heat conductivity of the two phases are different, provided that the temperature u_* of phase equilibrium is independent of x.

More specifically, denoting by $\tilde{\phi}$ the pointwise content of latent heat (that we denoted by $L(x)$), the function $\phi := \tilde{\phi}/L(x)$ characterizes the phase, for

$$\phi = 0 \ \text{ in the solid phase}, \quad 0 < \phi < 1 \ \text{ in the mushy region}, \quad \phi = 1 \ \text{ in the liquid phase}. \tag{6.16}$$

Thus $\phi \in \tilde{H}(u - u_*)$ a.e. in Ω_T, with $\tilde{H}$ being the Heaviside graph:

$$\tilde{H}(v) = \{0\} \qquad \forall v \le 0, \qquad \tilde{H}(0) = [0,1], \qquad \tilde{H}(v) = \{1\} \qquad \forall v > 0. \tag{6.17}$$

If K can be factorized in the form

$$K(u,x) = G(x)h(u,\phi) \qquad \text{a.e. in } \Omega_T, \tag{6.18}$$

then $K(u,x) \in G(x)h(u, \tilde{H}(u - u_*))$ a.e. in Ω_T. Assuming that u_* is differentiable and setting

$$V := \int_0^u h(\xi, \tilde{H}(\xi - u_*))\, d\xi \qquad \forall u \in \mathbf{R}, \tag{6.19}$$

a simple computation yields $\nabla V = [h(u_*, 0) - h(u_*, 1)]\nabla u_* + h(u, \tilde{H}(u - u_*))\nabla u$. The Fourier law (1.8) then reads

$$\vec{q} = -h(u,\phi)G(x) \cdot \nabla u = -G(x) \cdot \nabla V + G(x) \cdot [h(u_*, 0) - h(u_*, 1)]\nabla u_* \qquad \text{a.e. in } \Omega_T. \tag{6.20}$$

The analysis of this work may thus easily be extended to this setting.

On the other hand, if $K(u,x)$ cannot be factorized as in (6.18), it is not clear how existence of a weak solution might be proved. Open questions also include the extension of the downscaling result, i.e., Theorem 5.4, to a noncyclically monotone function $\vec{\alpha}$.

Generalizations. Another class of doubly nonlinear parabolic problems also seems of interest:

$$w + \nabla \cdot \vec{q} = f \qquad \text{in } \mathcal{D}'(\Omega_T), \tag{6.21}$$

$$w \in \beta(\partial u/\partial t, x) \qquad \text{a.e. in } \Omega_T, \tag{6.22}$$

$$\vec{q} = \partial\psi(-\nabla u, u, x) \qquad \text{a.e. in } \Omega_T, \tag{6.23}$$

here with $\beta : \mathbf{R} \times \Omega \to \mathbf{R}$ and $\psi : \mathbf{R}^3 \times \mathbf{R} \times \Omega \to \mathbf{R} \cup \{+\infty\}$. (Notice the exchanged role of the subdifferential and of the maximal monotone function here and in (1.1)–(1.3).) Existence of a solution for associated initial- and boundary-value problems were studied in several works; see, e.g., [5, 34, 35, 78]. The homogenization of this system might be treated via techniques analogous to those of this work.

The developments of this paper might be extended in several directions. For instance, one might insert a relaxation dynamics into the phase evolution equation (1.2) and/or into the nonlinear conduction law (1.3):

$$\frac{\partial w}{\partial t} + w \in \partial\varphi(u, x, t) \qquad \text{a.e. in } \Omega_T, \tag{6.24}$$

$$\frac{\partial \vec{q}}{\partial t} + \vec{q} = \vec{\alpha}(-\nabla u, u, x, t) \qquad \text{a.e. in } \Omega_T. \tag{6.25}$$

The latter is a nonlinear extension of the classical Cattaneo–Fourier law. Existence of a weak solution is known for either case; see, e.g., [78]; the homogenization is less obvious but might be performed via techniques of [83].

It might also be of applicative interest to extend the present analysis to the homogenization of phase transitions in binary mixtures, to account for two time scales, and to homogenize a nonperiodic medium.

REFERENCES

[1] V. Alexiades and A. D. Solomon, *Mathematical Modeling of Melting and Freezing Processes*, Hemisphere Publishing, Washington, DC, 1993.

[2] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[3] G. Allaire, *Shape Optimization by the Homogenization Method*, Springer, New York, 2002.

[4] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[5] T. Arai, *On the existence of the solution for $\partial\varphi(u'(t)) + \partial\psi(u(t)) \ni f(t)$*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 26 (1979), pp. 75–96.

[6] T. Arbogast, J. Douglas, Jr., and U. Hornung, *Derivation of the double porosity model of single phase flow via homogenization theory*, SIAM J. Math. Anal., 21 (1990), pp. 823–836.

[7] H. Attouch, *Variational Convergence for Functions and Operators*, Pitman, Boston, 1984.

[8] J.-P. Aubin, *Un théorème de compacité*, C. R. Acad. Sci. Paris, 256 (1963), pp. 5042–5044.

[9] N. Bakhvalov and G. Panasenko, *Homogenisation: Averaging Processes in Periodic Media*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1989.

[10] A. Bamberger, *Étude d'une équation doublement non linéaire*, J. Functional Analysis, 24 (1977), pp. 148–155.

[11] G. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North–Holland, Amsterdam, 1978.

[12] D. Blanchard and G. A. Francfort, *Study of a doubly nonlinear heat equation with no growth assumptions on the parabolic term*, SIAM J. Math. Anal., 19 (1988), pp. 1032–1056.

[13] D. Blanchard and G. Francfort, *A few results on a class of degenerate parabolic equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 18 (1991), pp. 213–249.

[14] D. Blanchard and A. Porretta, *Stefan problems with nonlinear diffusion and convection*, J. Differential Equations, 210 (2005), pp. 383–428.

[15] L. Boccardo and F. Murat, *Almost everywhere convergence of the gradients of solutions to elliptic and parabolic equations*, Nonlinear Anal., 19 (1992), pp. 581–597.

[16] L. Boccardo, F. Murat, and J.-P. Puel, *Existence of bounded solutions for nonlinear elliptic unilateral problems*, Ann. Mat. Pura Appl. (4), 152 (1988), pp. 183–196.

[17] A. Bossavit and A. Damlamian, *Homogenization of the Stefan problem and application to magnetic composite media*, IMA J. Appl. Math., 27 (1981), pp. 319–334.

[18] A. Bourgeat, S. Luckhaus, and A. Mikelić, *Convergence of the homogenization process for a double-porosity model of immiscible two-phase flow*, SIAM J. Math. Anal., 27 (1996), pp. 1520–1543.

[19] A. Braides and A. Defranceschi, *Homogenization of Multiple Integrals*, Oxford University Press, Oxford, UK, 1998.

[20] H. Brezis, *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North–Holland, Amsterdam, 1973.

[21] M. Brokate and J. Sprekels, *Hysteresis and Phase Transitions*, Springer, Heidelberg, 1996.

[22] F. Browder, *Existence Theorems for Nonlinear Partial Differential Equations*, in Proceedings of Symposia in Pure Mathematics, Vol. XVI, S. Chern and S. Smale, eds., AMS, Providence, RI, 1970, pp. 1–60.

[23] L. Carbone and R. De Arcangelis, *Unbounded Functionals in the Calculus of Variations. Representation, Relaxation, and Homogenization*, Chapman and Hall/CRC, Boca Raton, FL, 2001.

[24] L. Carbone and C. Sbordone, *Some properties of Γ-limits of integral functionals*, Ann. Mat. Pura Appl. (4), 122 (1979), pp. 1–60.

[25] J. Carrillo, *Entropy solutions for nonlinear degenerate problems*, Arch. Ration. Mech. Anal., 147 (1999), pp. 269–361.

[26] J. Carrillo and P. Wittbold, *Uniqueness of renormalized solutions of degenerate elliptic-parabolic problems*, J. Differential Equations, 156 (1999), pp. 93–121.

[27] R. W. Carroll and R. E. Showalter, *Singular and Degenerate Cauchy Problems*, Academic Press, New York, 1976.

[28] B. Chalmers, *Principles of Solidification*, Wiley, New York, 1964.

[29] V. Chiadò Piat and G. V. Sandrakov, *Homogenization of some variational inequalities for elasto-plastic torsion problems*, Asymptot. Anal., 40 (2004), pp. 1–23.

[30] D. Cioranescu, A. Damlamian, and R. De Arcangelis, *Homogenization of nonlinear integrals via the periodic unfolding method*, C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 77–82.

[31] D. Cioranescu, A. Damlamian, and R. De Arcangelis, *Homogenization of quasiconvex integrals via the periodic unfolding method*, SIAM J. Math. Anal., 37 (2006), pp. 1435–1453.

[32] D. Cioranescu, A. Damlamian, and G. Griso, *Periodic unfolding and homogenization*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 99–104.

[33] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford University Press, New York, 1999.

[34] P. Colli, *On some doubly nonlinear evolution equations in Banach spaces*, Japan J. Indust. Appl. Math., 9 (1992), pp. 181–203.

[35] P. Colli and A. Visintin, *On a class of doubly nonlinear evolution problems*, Comm. Partial Differential Equations, 15 (1990), pp. 737–756.

[36] J. Crank, *Free and Moving Boundary Problems*, Clarendon Press, Oxford, UK, 1984.

[37] G. Dal Maso, *An Introduction to Γ-Convergence*, Birkhäuser, Boston, 1993.

[38] A. Damlamian, *How to homogenize a nonlinear diffusion equation: Stefan's problem*, SIAM J. Math. Anal., 12 (1981), pp. 306–313.

[39] E. De Giorgi and S. Spagnolo, *Sulla convergenza degli integrali dell'energia per operatori ellittici del secondo ordine*, Boll. Un. Mat. Ital. (4), 8 (1973), pp. 391–411.

[40] E. Di Benedetto and R. E. Showalter, *Implicit degenerate evolution equations and applications*, SIAM J. Math. Anal., 12 (1981), pp. 731–751.

[41] W. E, *Homogenization of linear and nonlinear transport equations*, Comm. Pure Appl. Math., 45 (1992), pp. 301–326.

[42] C. Eck, *Homogenization of a phase field model for binary mixtures*, Multiscale Model. Simul., 3 (2004), pp. 1–27.

[43] C. Eck, *Analysis of a two-scale phase field model for liquid-solid phase transitions with equiaxed dendritic microstructure*, Multiscale Model. Simul., 3 (2004), pp. 28–49.

[44] C. Eck, P. Knabner, and S. Korotov, *A two-scale method for the computation of solid-liquid phase transitions with dendritic microstructure*, J. Comput. Phys., 178 (2002), pp. 58–80.

[45] I. Ekeland and R. Temam, *Analyse Convexe et Problèmes Variationnelles*, Dunod, Gauthier–Villars, Paris, 1974.

[46] C. M. Elliott and J. R. Ockendon, *Weak and Variational Methods for Moving Boundary Problems*, Pitman, Boston, 1982.
[47] M. C. Flemings, *Solidification Processing*, McGraw–Hill, New York, 1973.
[48] A. Friedman, *Variational Principles and Free Boundary Problems*, Wiley, New York, 1982.
[49] O. Grange and F. Mignot, *Sur la résolution d'une équation et d'une inéquation paraboliques non linéaires*, J. Functional Analysis, 11 (1972), pp. 77–92.
[50] S. C. Gupta, *The Classical Stefan Problem*, Elsevier, Amsterdam, 2003.
[51] J.-B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Optimization Algorithms*, Springer, Berlin, 1993.
[52] U. Hornung, ed., *Homogenization and Porous Media*, Springer, New York, 1997.
[53] N. Igbida and J. M. Urbano, *Uniqueness for nonlinear degenerate problems*, NoDEA Nonlinear Differential Equations Appl., 10 (2003), pp. 287–307.
[54] A. D. Ioffe and V. M. Tihomirov, *Theory of Extremal Problems*, North–Holland, Amsterdam, 1979.
[55] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer, Berlin, 1994.
[56] W. Kurz and D. J. Fisher, *Fundamentals of Solidification*, Trans Tech, Aedermannsdorf, Switzerland, 1989.
[57] J. L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites non Linéaires*, Dunod, Paris, 1969.
[58] J. L. Lions, *Asymptotic behaviour of solutions of variational inequalities with highly oscillating coefficients*, in Applications of Methods of Functional Analysis to Problems in Mechanics, P. Germain and B. Nayroles, eds., Springer, Berlin, 1976, pp. 30–55.
[59] D. Lukkassen, G. Nguetseng, and P. Wall, *Two-scale convergence*, Int. J. Pure Appl. Math., 2 (2002), pp. 35–86.
[60] P. Marcellini, *Periodic solutions and homogenization of nonlinear variational problems*, Ann. Mat. Pura Appl. (4), 117 (1978), pp. 139–152.
[61] G. W. Milton, *The Theory of Composites*, Cambridge University Press, Cambridge, UK, 2002.
[62] A. M. Meirmanov, *The Stefan Problem*, De Gruyter, Berlin, 1992 (in English); Nauka, Novosibirsk, 1986 (in Russian).
[63] F. Murat and L. Tartar, *H-convergence*, in Topics in the Mathematical Modelling of Composite Materials, A. Cherkaev and R. Kohn, eds., Birkhäuser, Boston, 1997, pp. 21–43.
[64] A. K. Nandakumaran and M. Rajesh, *Homogenization of a nonlinear degenerate parabolic differential equation*, Electron. J. Differential Equations, 17 (2001).
[65] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.
[66] P. A. Raviart, *Sur la résolution de certaines équations paraboliques non linéaires*, J. Functional Analysis, 5 (1970), pp. 299–328.
[67] R. T. Rockafellar, *Integrals which are convex functionals*, Pacific J. Math., 24 (1968), pp. 525–539.
[68] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1969.
[69] J.-F. Rodrigues, *Free boundary convergence in the homogenization of the one-phase Stefan problem*, Trans. Amer. Math. Soc., 274 (1982), pp. 297–305.
[70] A. Romano, *Thermomechanics of Phase Transitions in Classical Field Theory*, World Scientific, River Edge, NJ, 1993.
[71] L. Rubinstein, *The Stefan Problem*, AMS, Providence, RI, 1971 (in English); Zvaigzne, Riga, 1967 (in Russian).
[72] E. Sanchez-Palencia, *Non-Homogeneous Media and Vibration Theory*, Springer, New York, 1980.
[73] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.
[74] L. Tartar, *Course Peccot*, Collège de France, Paris, 1977 (unpublished, partially written in *H-convergence*, in Topics in the Mathematical Modelling of Composite Materials, A. Cherkaev and R. Kohn, eds., Birkhäuser, Boston, 1997, pp. 21–43).
[75] D. A. Tarzia, *A Bibliography on Moving-Free Boundary Problems for the Heat-Diffusion Equation. The Stefan and Related Problems*, MAT Ser. A Conf. Semin. Trab. Mat. 2, Universidad Austral, Rosario, Argentina, 2000.
[76] M. Tsutsumi, *On solutions of some doubly nonlinear degenerate parabolic equations with absorption*, J. Math. Anal. Appl., 132 (1988), pp. 187–212.
[77] A. Visintin, *Strong convergence results related to strict convexity*, Comm. Partial Differential Equations, 9 (1984), pp. 439–466.
[78] A. Visintin, *Models of Phase Transitions*, Birkhäuser, Boston, 1996.

[79] A. Visintin, *Some properties of two-scale convergence*, Rend. Acc. Naz. Lincei, 15 (2004), pp. 93–107.
[80] A. Visintin, *Homogenization of doubly-nonlinear equations*, Rend. Lincei Mat. Appl., 17 (2006), pp. 211–222.
[81] A. Visintin, *Two-scale convergence of some integral functionals*, Calc. Var. Partial Differential Equations, 29 (2007), pp. 239–265.
[82] A. Visintin, *Two-scale convergence of first-order operators*, Z. Anal. Anwendungen, 26 (2007), pp. 133–164.
[83] A. Visintin, *Upscaling and Downscaling in Homogenization with Applications to Electromagnetism*, manuscript.
[84] A. Visintin, *Electromagnetic processes in doubly-nonlinear composites*, submitted.
[85] P. D. Woodruff, *The Solid-Liquid Interface*, Cambridge University Press, Cambridge, UK, 1973.

SIAM J. MATH. ANAL.
Vol. 39, No. 3, pp. 1018–1021

ERRATUM: "GLOBAL SOLUTIONS OF NONLINEAR TRANSPORT EQUATIONS FOR CHEMOSENSITIVE MOVEMENT" [SIAM J. MATH. ANAL. 36 (2005), PP. 1177–1199]*

HYUNG JU HWANG†, KYUNGKEUN KANG‡, AND ANGELA STEVENS§

Abstract. The purpose of this erratum is to correct Assumption 4.2 in [H. J. Hwang, K. Kang, and A. Stevens, *SIAM J. Math. Anal.*, 36 (2005), pp. 1177–1199]. We also modify some errors caused by the incorrectly stated assumption.

Key words. chemosensitive movement, drift-diffusion limit, Keller-Segel model

AMS subject classifications. 35K55, 45K05, 82C70, 92C17

DOI. 10.1137/070680813

First we correct Assumption 4.2 ([2], p. 1191, lines 3–4) by adding condition (0.3) given below

$$\phi_\epsilon^S[S] \geq \gamma \left(1 - \epsilon\Lambda(\|\nabla S\|_{W^{1,\infty}(\mathbb{R}^3)})\right) FF', \tag{0.1}$$

$$\int_V \frac{\phi_\epsilon^A[S]^2}{F\phi_\epsilon^S[S]} dv' \leq \epsilon^2 \Lambda\left(\|\nabla S\|_{W^{1,\infty}(\mathbb{R}^3)}\right), \tag{0.2}$$

$$\left\| \frac{(\nabla\phi_\epsilon^S[S])^2}{\phi_\epsilon^S[S]} \right\|_{L^{\frac{q}{2}}(\mathbb{R}^3)} + \left\| \frac{(\nabla\phi_\epsilon^A[S])^2}{\phi_\epsilon^S[S]} \right\|_{L^{\frac{q}{2}}(\mathbb{R}^3)} \leq C\epsilon^2 \|\nabla S\|^2_{W^{2,q}(\mathbb{R}^3)}, \quad 2 \leq q \leq \infty, \tag{0.3}$$

where $\phi_\epsilon^S[S]$ and $\phi_\epsilon^A[S]$ denote the symmetric and antisymmetric parts of the turning kernel, respectively, i.e., $\phi_\epsilon^S := \frac{1}{2}\left(T_\epsilon[S]F' + T_\epsilon^*[S]F\right)$ and $\phi_\epsilon^A := \frac{1}{2}\left(T_\epsilon[S]F' + T_\epsilon^*[S]F\right)$, and $\Lambda \in L^\infty_{\text{loc}}([0,\infty))$ is a nondecreasing function. Here F indicates a bounded velocity distribution satisfying (A0) in [1, p. 128, lines 18–23]. In what follows, we suppose, for simplicity, that the bounded velocity distribution fulfills $F = 1$. These conditions (0.1)–(0.3) hold for the example

$$T_\epsilon[S] = C + h\left(S(x+\epsilon v,t) - S(x-\epsilon v',t)\right) + h\left(\nabla S(x+\epsilon v,t) - \nabla S(x-\epsilon v',t)\right), \tag{0.4}$$

where $C > 0$ and h is a nonnegative and differentiable function satisfying $|h'(x)| \leq M$ and $C_1|x| \leq h(x) \leq C_2|x|$ for some $C_2 > C_1 > 0$ (see the appendix for details).

Our main purpose of this erratum is to fix the error in the proof of Theorem 4.3 in [2], where the estimate of ∇S ([2], p. 1191, line 5 from the bottom) is not correct. Here we consider only the case $\tau = 0, n = 3$, and $\beta = 0$. For the other cases similar arguments work.

Assume for the initial data $f_\epsilon^0, \nabla_x f_\epsilon^0 \in L^q(V \times \mathbb{R}^3)$. Our model equation is

$$\frac{\partial}{\partial t} f_\epsilon + \frac{1}{\epsilon} v \cdot \nabla_x f_\epsilon = -\frac{1}{\epsilon^2}\mathcal{T}[S_\epsilon](f_\epsilon), \tag{0.5}$$

*Received by the editors January 23, 2007; accepted for publication (in revised form) May 3, 2007; published electronically October 19, 2007.
http://www.siam.org/journals/sima/39-3/68081.html
†Department of Mathematics, Postech, Pohang 790-784, Republic of Korea (hjhwang@postech.ac.kr).
‡Department of Mathematics, Sungkyunkwan University, Suwon 440-746, Republic of Korea (kkang@skku.edu).
§Max-Planck-Institute for Mathematics in the Sciences, Inselstr. 22 - 26, D-04103 Leipzig, Germany (stevens@mis.mpg.de).

with

$$\mathcal{T}[S_\epsilon](f_\epsilon) = -\int_V \left(T[S_\epsilon]f'_\epsilon - T^*[S_\epsilon]f_\epsilon\right) dv'.$$

Let $q > 3$. Multiplying (0.5) with f_ϵ^{q-1} and integrating over $\mathbb{R}^3 \times V$, we obtain

$$\begin{aligned}
&\frac{1}{q}\frac{d}{dt}\int_{\mathbb{R}^3}\int_V f_\epsilon^q dv dx + \frac{1}{4\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \phi_\epsilon^S[S](f_\epsilon - f'_\epsilon)\left(f^{q-1} - (f')^{q-1}\right) dv' dv dx \\
&\quad \le \frac{1}{4\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \frac{\phi_\epsilon^A[S]^2}{\phi_\epsilon^S[S]}\frac{(f_\epsilon + f'_\epsilon)^2\left(f_\epsilon\left|f_\epsilon\right|^{q-2} - f'_\epsilon\left|f'_\epsilon\right|^{q-2}\right)}{f_\epsilon - f'_\epsilon} dv' dv dx \\
&\le C_q\Lambda\left(\left\|\nabla S_\epsilon\right\|_{W^{1,\infty}(\mathbb{R}^3)}\right)\int_{\mathbb{R}^3}\int_V f_\epsilon^q dv dx \le C_q\Lambda\left(\left\|\nabla f_\epsilon\right\|_{L^q(V\times\mathbb{R}^3)}\right)\int_{\mathbb{R}^3}\int_V f_\epsilon^q dv dx.
\end{aligned}$$

Here (0.1), (0.2), and Lemma 1 in [1, p. 129] are used, and $\left\|\nabla S_\epsilon\right\|_{W^{1,\infty}(\mathbb{R}^3)} \le C\left\|\nabla\rho_\epsilon\right\|_{L^q(\mathbb{R}^3)} \le C\left\|\nabla f_\epsilon\right\|_{L^q(V\times\mathbb{R}^3)}$. Therefore, we obtain

$$\frac{d}{dt}\left\|f_\epsilon\right\|^q_{L^q(V\times\mathbb{R}^3)} \le C_q\Lambda\left(\left\|\nabla f_\epsilon\right\|_{L^q(V\times\mathbb{R}^3)}\right)\left\|f_\epsilon\right\|^q_{L^q(V\times\mathbb{R}^3)}. \tag{0.6}$$

Now set $g_\epsilon = \partial_{x_j} f_\epsilon$, $j = 1, 2, 3$. Differentiating (0.5) with respect to x_j, we get

$$\frac{\partial}{\partial t}g_\epsilon + \frac{1}{\epsilon}v\cdot\nabla_x g_\epsilon = -\frac{1}{\epsilon^2}\mathcal{T}[S_\epsilon](g_\epsilon) - \frac{1}{\epsilon^2}\left(\partial_{x_j}\mathcal{T}[S_\epsilon]\right)(f_\epsilon),$$

where

$$\left(\partial_{x_j}\mathcal{T}[S_\epsilon]\right)(f_\epsilon) = -\int_V\left[\left(\partial_{x_j}T[S_\epsilon]\right)f'_\epsilon - \left(\partial_{x_j}T^*[S_\epsilon]\right)f_\epsilon\right] dv'.$$

Let $q > 3$. Multiplying with $\left|g_\epsilon\right|^{q-2} g_\epsilon$ and integrating over $\mathbb{R}^3 \times V$, we have

$$\begin{aligned}
\frac{1}{q}\frac{d}{dt}\int_{\mathbb{R}^3}\int_V \left|g_\epsilon\right|^q &= -\frac{1}{\epsilon^2}\int_{\mathbb{R}^3}\int_V \mathcal{T}[S_\epsilon](g_\epsilon)\left|g_\epsilon\right|^{q-2} g_\epsilon - \frac{1}{\epsilon^2}\int_{\mathbb{R}^3}\int_V\left(\partial_{x_j}\mathcal{T}[S_\epsilon]\right)(f_\epsilon)\left|g_\epsilon\right|^{q-2} g_\epsilon \\
&:= \mathrm{I} + \mathrm{II}.
\end{aligned}$$

Using again Lemma 1 in [1, p. 129], we note that I can be written as

$$\begin{aligned}
\mathrm{I} = &-\frac{1}{2\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \phi_\epsilon^S[S](g_\epsilon - g'_\epsilon)\left(g_\epsilon\left|g_\epsilon\right|^{q-2} - g'_\epsilon\left|g'_\epsilon\right|^{q-2}\right) dv' dv dx \\
&+ \frac{1}{2\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \phi_\epsilon^A[S](g_\epsilon + g'_\epsilon)\left(g_\epsilon\left|g_\epsilon\right|^{q-2} - g'_\epsilon\left|g'_\epsilon\right|^{q-2}\right) dv' dv dx.
\end{aligned}$$

Following a similar procedure as before, we obtain

$$\begin{aligned}
\mathrm{I} \le &-\frac{1}{4\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \phi_\epsilon^S[S](g_\epsilon - g'_\epsilon)\left(g_\epsilon\left|g_\epsilon\right|^{q-2} - g'_\epsilon\left|g'_\epsilon\right|^{q-2}\right) dv' dv dx \\
&+ C_q\left\|\nabla f_\epsilon\right\|_{L^q(\mathbb{R}^3\times V)}\int_{\mathbb{R}^3}\int_V \left|g_\epsilon\right|^q dv dx. \qquad (0.7)
\end{aligned}$$

Next we estimate II:

$$\begin{aligned}
\mathrm{II} = &-\frac{1}{2\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \left(\partial_{x_j}\phi_\epsilon^S[S]\right)(f_\epsilon - f'_\epsilon)\left(g_\epsilon\,|g_\epsilon|^{q-2} - g'_\epsilon\,|g'_\epsilon|^{q-2}\right) dv'dvdx \\
&+ \frac{1}{2\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \left(\partial_{x_j}\phi_\epsilon^A[S]\right)(f_\epsilon + f'_\epsilon)\left(g_\epsilon\,|g_\epsilon|^{q-2} - g'_\epsilon\,|g'_\epsilon|^{q-2}\right) dv'dvdx \\
&\le \frac{1}{8\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \phi_\epsilon^S[S](g_\epsilon - g'_\epsilon)\left(g_\epsilon\,|g_\epsilon|^{q-2} - g'_\epsilon\,|g'_\epsilon|^{q-2}\right) dv'dvdx \\
&+ \frac{1}{\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \frac{\left(\partial_{x_j}\phi_\epsilon^S[S]\right)^2}{\phi_\epsilon^S[S]}(f_\epsilon - f'_\epsilon)^2\left|\frac{\left(g_\epsilon\,|g_\epsilon|^{q-2} - g'_\epsilon\,|g'_\epsilon|^{q-2}\right)}{g-g'}\right| dv'dvdx \\
&+ \frac{1}{\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \frac{\left(\partial_{x_j}\phi_\epsilon^A[S]\right)^2}{\phi_\epsilon^S[S]}(f_\epsilon + f'_\epsilon)^2\left|\frac{\left(g_\epsilon\,|g_\epsilon|^{q-2} - g'_\epsilon\,|g'_\epsilon|^{q-2}\right)}{g-g'}\right| dv'dvdx \\
&\le \frac{1}{8\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \phi_\epsilon^S[S](g_\epsilon - g'_\epsilon)\left(g_\epsilon\,|g_\epsilon|^{q-2} - g'_\epsilon\,|g'_\epsilon|^{q-2}\right) dv'dvdx \\
&+ \frac{C_q}{\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \left(\frac{\left(\partial_{x_j}\phi_\epsilon^S[S]\right)^2}{\phi_\epsilon^S[S]} + \frac{\left(\partial_{x_j}\phi_\epsilon^A[S]\right)^2}{\phi_\epsilon^S[S]}\right)\left({f_\epsilon}^2 + {f'_\epsilon}^2\right)\left(|g_\epsilon|^{q-2} + |g'_\epsilon|^{q-2}\right) dv'dvdx.
\end{aligned}$$

For the last term, interchanging the order of integration and utilizing (0.3), Hölder's inequality, using the potential estimate for S_ϵ (see, e.g., [2, Lemma 4.1]), and Sobolev's embedding theorem, we can estimate

$$\begin{aligned}
&\frac{1}{\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \left(\frac{\left(\partial_{x_j}\phi_\epsilon^S[S]\right)^2}{\phi_\epsilon^S[S]} + \frac{\left(\partial_{x_j}\phi_\epsilon^A[S]\right)^2}{\phi_\epsilon^S[S]}\right)\left(f_\epsilon^2 + {f'_\epsilon}^2\right)\left(|g_\epsilon|^{q-2} + |g'_\epsilon|^{q-2}\right) dv'dvdx \\
&= \frac{1}{\epsilon^2}\int_V\int_V\int_{\mathbb{R}^3} \left(\frac{\left(\partial_{x_j}\phi_\epsilon^S[S]\right)^2}{\phi_\epsilon^S[S]} + \frac{\left(\partial_{x_j}\phi_\epsilon^A[S]\right)^2}{\phi_\epsilon^S[S]}\right)\left({f_\epsilon}^2 + {f'_\epsilon}^2\right)\left(|g_\epsilon|^{q-2} + |g'_\epsilon|^{q-2}\right) dxdv'dv \\
&\quad\le \frac{1}{\epsilon^2}\int_V\int_V \|f_\epsilon + f'_\epsilon\|^2_{L^\infty(\mathbb{R}^3)}\left\|\left(\frac{\left(\partial_{x_j}\phi_\epsilon^S[S]\right)^2}{\phi_\epsilon^S[S]}\right) + \left(\frac{\left(\partial_{x_j}\phi_\epsilon^A[S]\right)^2}{\phi_\epsilon^S[S]}\right)\right\|_{L^{\frac{q}{2}}(\mathbb{R}^3)} \\
&\quad\times \|(|g_\epsilon| + |g'_\epsilon|)\|^{q-2}_{L^q(\mathbb{R}^3)}\,dv'dv \\
&\quad\le C\,\|\nabla S_\epsilon\|^2_{W^{2,q}(\mathbb{R}^3)}\int_V\int_V \|f_\epsilon + f'_\epsilon\|^2_{W^{1,q}(\mathbb{R}^3)}\,\|(|g_\epsilon| + |g'_\epsilon|)\|^{q-2}_{L^q(\mathbb{R}^3)}\,dv'dv \\
&\quad\le C\,\|\nabla\rho_\epsilon\|^2_{L^q(\mathbb{R}^3)}\left(\|f_\epsilon\|^2_{L^q(V\times\mathbb{R}^3)} + \|\nabla f_\epsilon\|^2_{L^q(V\times\mathbb{R}^3)}\right)\|g_\epsilon\|^{q-2}_{L^q(V\times\mathbb{R}^3)} \\
&\quad\le C_q\left(\|f_\epsilon\|^{q+2}_{L^q(V\times\mathbb{R}^3)} + \|\nabla f_\epsilon\|^{q+2}_{L^q(V\times\mathbb{R}^3)}\right).
\end{aligned}$$

Summing up, we obtain

$$\begin{aligned}
\frac{1}{q}\frac{d}{dt}\int_{\mathbb{R}^3}\int_V |g_\epsilon|^q + \frac{1}{8\epsilon^2}\int_{\mathbb{R}^3}\int_V\int_V \phi_\epsilon^S[S](g_\epsilon - g'_\epsilon)\left(g_\epsilon\,|g_\epsilon|^{q-2} - g'_\epsilon\,|g'_\epsilon|^{q-2}\right) dv'dvdx \\
\le C_q\left(\|f_\epsilon\|^{q+2}_{L^q(V\times\mathbb{R}^3)} + \|\nabla f_\epsilon\|^{q+2}_{L^q(V\times\mathbb{R}^3)}\right).
\end{aligned}$$

The second term on the left-hand side of the above inequality is nonnegative, so it can be omitted. Adding up for $j = 1, 2, 3$, we obtain

$$\frac{d}{dt}\|\nabla f_\epsilon\|^q_{L^q(V\times\mathbb{R}^3)} \le C_q\left(\|f_\epsilon\|^{q+2}_{L^q(V\times\mathbb{R}^3)} + \|\nabla f_\epsilon\|^{q+2}_{L^q(V\times\mathbb{R}^3)}\right). \tag{0.8}$$

Let $X(t) := \|f_\epsilon\|^q_{L^q(V\times\mathbb{R}^3)} + \|\nabla f_\epsilon\|^q_{L^q(V\times\mathbb{R}^3)}$. Combining (0.6) and (0.8), we get

$$X'(t) \le C\Lambda\left(X^{\frac{1}{q}}(t)\right)X(t) + X^{\frac{q+2}{q}}(t).$$

Since $\Lambda \in L^\infty_{\text{loc}}([0,\infty))$, the above inequality ensures a short time existence of solutions, independent of ϵ. This completes the correction of the proof of Theorem 4.3.

On p. 1195 of [2] the following assumption has to be added between lines 4 and 5 from the bottom:

$$\left\|\frac{(\nabla\phi^S_\epsilon[S])^2}{\phi^S_\epsilon[S]}\right\|_{L^{\frac{q}{2}}(\mathbb{R}^3)} + \left\|\frac{(\nabla\phi^A_\epsilon[S])^2}{\phi^S_\epsilon[S]}\right\|_{L^{\frac{q}{2}}(\mathbb{R}^3)} \le C\epsilon^2\|\nabla S\|^2_{W^{2,q}(\mathbb{R}^3)}, \quad 2 \le q \le \infty.$$

Appendix. Here we show that the turning kernel (0.4) satisfies (0.1)–(0.3). Conditions (0.1) and (0.2) are straightforward. To show condition (0.3), denote for simplicity

$$\begin{aligned} S_v &:= S(x+\epsilon v) - S(x-\epsilon v'), & S_{v'} &:= S(x+\epsilon v') - S(x-\epsilon v),\\ \nabla S_v &:= \nabla S(x+\epsilon v) - \nabla S(x-\epsilon v'), & \nabla S_{v'} &:= \nabla S(x+\epsilon v') - \nabla S(x-\epsilon v). \end{aligned}$$

The symmetric and antisymmetric parts are given by

$$\begin{aligned} \phi^S_\epsilon[S] &= C + \frac{1}{2}\left(h(S_v) + h(S_{v'}) + h(\nabla S_v) + h(\nabla S_{v'})\right),\\ \phi^A_\epsilon[S] &= \frac{1}{2}\left(h(S_v) - h(S_{v'}) + h(\nabla S_v) - h(\nabla S_{v'})\right). \end{aligned}$$

Taking the derivative with respect to the x-variable, we get

$$\begin{aligned} \nabla\phi^S_\epsilon[S] &= \frac{1}{2}\left(h'(S_v)\nabla S_v + h'(S_{v'})\nabla S_{v'} + h'(\nabla S_v)\nabla^2 S_v + h'(\nabla S_{v'})\nabla^2 S_{V'}\right),\\ \nabla\phi^A_\epsilon[S] &= \frac{1}{2}\left(h'(S_v)\nabla S_v - h'(S_{v'})\nabla S_{v'} + h'(\nabla S_v)\nabla^2 S_v - h'(\nabla S_{v'})\nabla^2 S_{V'}\right). \end{aligned}$$

By means of the mean value theorem, we have

$$\begin{aligned} \nabla S_v &= \epsilon(v-v')\nabla^2 S\left(x+\epsilon\theta_1 v - \epsilon(1-\theta_1)v'\right), & 0 \le \theta_1 \le 1,\\ \nabla S_{v'} &= \epsilon(v'-v)\nabla^2 S\left(x+\epsilon\theta_2 v' - \epsilon(1-\theta_2)v\right), & 0 \le \theta_2 \le 1,\\ \nabla^2 S_v &= \epsilon(v-v')\nabla^3 S\left(x+\epsilon\theta_3 v - \epsilon(1-\theta_3)v'\right), & 0 \le \theta_3 \le 1,\\ \nabla^2 S_v &= \epsilon(v-v')\nabla^3 S\left(x+\epsilon\theta_4 v - \epsilon(1-\theta_4)v'\right), & 0 \le \theta_4 \le 1. \end{aligned}$$

Therefore, since $\|h'\|_{L^\infty} < M$ with a fixed constant M, we obtain

$$\left\|\frac{(\nabla\phi^S_\epsilon[S])^2}{\phi^S_\epsilon[S]}\right\|_{L^{\frac{q}{2}}(\mathbb{R}^3)} \le C\epsilon^2\left(\int_{\mathbb{R}^3}\left|\nabla^2 S\right|^q + \left|\nabla^3 S\right|^q dx\right)^{\frac{2}{q}} \le C\epsilon^2\|\nabla S\|^2_{W^{2,q}(\mathbb{R}^3)}.$$

The estimate for $\nabla\phi^A_\epsilon[S]$ follows by a similar procedure.

REFERENCES

[1] F. A. C. C. Chalub, P. Markowich, B. Perthame, and C. Schmeiser, *Kinetic models for chemotaxis and their drift-diffusion limits*, Monatsh. Math., 142 (2004), pp. 123–141.

[2] H. J. Hwang, K. Kang, and A. Stevens, *Global solutions of nonlinear transport equations for chemosensitive movement*, SIAM J. Math. Anal., 36 (2005), pp. 1177–1199.

SIAM J. MATH. ANAL.
Vol. 39, No. 4, pp. 1023–1069

STABILITY AND BIFURCATION ANALYSIS OF VOLTERRA FUNCTIONAL EQUATIONS IN THE LIGHT OF SUNS AND STARS*

ODO DIEKMANN†, PHILIPP GETTO‡, AND MATS GYLLENBERG§

Abstract. We show that the perturbation theory for dual semigroups (sun-star-calculus) that has proved useful for analyzing delay-differential equations is equally efficient for dealing with Volterra functional equations. In particular, we obtain both the stability and instability parts of the principle of linearized stability and the Hopf bifurcation theorem. Our results apply to situations in which the instability part has not been proved before. In applications to general physiologically structured populations even the stability part is new.

Key words. delay equations, dual semigroup, sun-star-calculus, Lipschitz perturbations, principle of linearized stability, center manifold, Hopf bifurcation, physiologically structured populations

AMS subject classifications. 39B82, 47D99, 92D25

DOI. 10.1137/060659211

1. Introduction.

Delay equations are rules for extending (in one direction) a function that is a priori defined on an interval. Usually, as in the books [23, 40], one considers delay *differential* equations of the type

$$\dot{x}(t) = F(x_t), \quad t \geq 0, \tag{1.1}$$

where, for some given $h > 0$,

$$x_t(\theta) := x(t + \theta) \tag{1.2}$$

for $\theta \in [-h, 0]$. Here, in contrast, we consider functional equations of Volterra type, so the extension rule prescribes the value of the function itself, rather than that of its derivative, in terms of the history. We thus study initial value problems of the form

$$x(t) = F(x_t), \quad t > 0, \tag{DE}$$

$$x_0(\theta) = \varphi(\theta), \quad \theta \in [-h, 0], \tag{IC}$$

with φ being a given function on $[-h, 0]$. The formula labels (DE), (IC) stand for delay equation and initial condition, respectively.

In [23], the main tool for analyzing the delay differential equation (1.1) is the perturbation theory for dual semigroups developed in [9, 10, 11, 12, 20], which under appropriate assumptions transforms the Cauchy problem (1.1) and (IC) into an abstract *semilinear* problem. This theory has proved to be equally efficient for treating

*Received by the editors May 5, 2006; accepted for publication February 27, 2007; published electronically October 24, 2007.
http://www.siam.org/journals/sima/39-4/65921.html

†Department of Mathematics, University of Utrecht, P.O. Box 80010, 3580 TA Utrecht, The Netherlands (O.Diekmann@math.uu.nl).

‡Department of Mathematics, University of Utrecht, P.O. Box 80010, 3580 TA Utrecht, The Netherlands. Current address: Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK (P.Getto@warwick.ac.uk). In the later stages of his research, this author's work was supported by project MRTN-CT-2004-503661.

§Corresponding author. Rolf Nevanlinna Institute, Department of Mathematics and Statistics, P.O. Box 68, FI-00014 University of Helsinki, Finland (mats.gyllenberg@helsinki.fi). The work of this author was supported by the Academy of Finland.

age-structured population models; see [9, 11] and various exercises in [23]. The aim of this paper is to show in detail that the same theory applies to functional equations of Volterra type (DE), the only difference being the choice of the underlying function space.

To give a feeling for the problems involved, we make a few formal manipulations. Let

$$u(t,\theta) := x_t(\theta), \quad t \geq 0,\ -h \leq \theta \leq 0. \tag{1.3}$$

The problem (DE), (IC) is equivalent to the following PDE with boundary and initial conditions:

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial \theta} = 0, \quad t > 0,\ -h \leq \theta \leq 0, \tag{1.4}$$

$$u(t,0) = F(u(t,\cdot)), \quad t \geq 0, \tag{1.5}$$

$$u(0,\theta) = \varphi(\theta), \quad -h \leq \theta \leq 0. \tag{1.6}$$

If $F = 0$, the problem reduces to an elementary linear problem. Its solution semigroup $T_0 = \{T_0(t)\}_{t\geq 0}$ is simply translation to the left with extension by zero:

$$(T_0(t)\varphi)(\theta) := \begin{cases} \varphi(t+\theta) & \text{for} \quad t+\theta \in [-h,0], \\ 0 & \text{for} \quad t+\theta > 0, \end{cases} \qquad t \geq 0,\ \theta \in [-h,0]. \tag{1.7}$$

Next we have to specify the state space (history space) on which the semigroup T_0 acts. The continuous functions will not do, because as can be seen from (1.7), $C[-h,0]$ is not invariant under T_0. A natural choice is $X = L^1[-h,0]$. With this choice of state space, the generator A_0 of T_0 is differentiation with the zero boundary condition entering into the domain of definition:

$$\mathcal{D}(A_0) = \{\varphi \in X : \varphi \in AC,\ \varphi(0) = 0\}, \tag{1.8}$$

$$A_0\varphi = \varphi', \tag{1.9}$$

where the notation $\varphi \in AC$ means that φ is absolutely continuous [2, p. 11].

The nonlinear problem (1.4)–(1.6) with $F \neq 0$ can now be written as the abstract Cauchy problem

$$\frac{du(t)}{dt} = A(u(t))u(t), \quad t > 0, \tag{1.10}$$

$$u(0) = \varphi \tag{1.11}$$

for $u(t) := u(t,\cdot) = x_t$, where the action of $A(u(t))$ is still differentiation, but the domain depends on the solution itself in a nonlinear way:

$$\mathcal{D}(A(u(t))) = \{\varphi \in X : \varphi \in AC, \varphi(0) = F(u(t))\}. \tag{1.12}$$

So the problem is *quasi-linear* and hence notoriously difficult [49]. A small trick, however, turns the quasi-linear problem into a *semilinear* one, that is, a problem in which the nonlinearity appears as an additive and relatively bounded perturbation of the linear operator A_0. Next we explain how this is done.

The space $L^1[-h,0]$ can be embedded into $\mathrm{NBV}(-h,0]$, the space of functions of bounded variation on $(-h,0]$ normalized to be zero at zero and continuous from the right, by integration $j : L^1[-h,0] \to \mathrm{NBV}(-h,0]$,

$$(j\varphi)(\theta) := -\int_\theta^0 \varphi(\tau)d\tau, \quad \theta \in (-h,0]. \tag{1.13}$$

The image of $L^1[-h,0]$ in $\mathrm{NBV}(-h,0]$ under j consists of all functions absolutely continuous on $[-h,0]$ and vanishing at 0 [46, sections IX. 2–4].

Integrating (1.4) from θ to 0 and taking the boundary condition (1.5) into account, one obtains the semilinear problem

$$\frac{d}{dt} ju(t) = A_0^{\odot *} ju(t) + F(u(t))H, \quad t > 0, \tag{1.14}$$

$$u(0) = \varphi, \tag{1.15}$$

where the operator $A_0^{\odot *}$ is differentiation on $\mathrm{NBV}(-h,0]$ with appropriate domain of definition (the $\odot *$-notation will be explained in section 2) and H is a Heaviside function defined by

$$H(\theta) := \begin{cases} -1 & \text{for} \quad \theta \in (-h,0), \\ 0 & \text{for} \quad \theta = 0. \end{cases} \tag{1.16}$$

The price one has to pay for the transformation of the quasi-linear problem into a semilinear one is that, while the unknown $u(t) = x_t$ belongs to $L^1[-h,0]$, the range of the perturbation lies in the bigger space $\mathrm{NBV}(-h,0]$ and actually outside $j\left(L^1[-h,0]\right)$ (note that $H \in \mathrm{NBV}(-h,0]$, but because of the discontinuity in 0 it is not absolutely continuous). The perturbation theory mentioned above was designed especially to have a general framework for such problems.

A key step is to replace the Cauchy problem (1.14) and (1.15) by an abstract integral equation of the variation-of-constants type, which is obtained from (1.14) and (1.15) by formal integration. The main point is that, in fact, this abstract integral equation is equivalent to the original problem (DE), (IC), while at the same time, it allows us to prove linearized stability and other properties in a standard manner. As these proofs are provided in detail in [23], we can concentrate here on the equivalence. Note, however, that in the present paper we shall always explicitly express the embedding operator j, while in [23] it is often suppressed with the understanding that one can identify X and $X^{\odot\odot}$ once and for all.

The mathematics of age-structured populations mentioned above has been extensively treated, for instance, in the books [17, 56]. Our main motivation comes from the theory of general physiologically structured populations [24, 25, 27, 45]. Individuals are distinguished from one another by their i-state (i for individual), which belongs to a measurable space Ω. The population state (p-state) is a measure m on Ω giving the distribution of i-states. Deterministic structured population models are defined in terms of ingredients prescribing i-state specific survival, reproduction, and i-state development, given the course of the *environmental condition* (or *input*) $I(t)$ and a feedback mechanism, which often is of the form

$$I(t) = \int_\Omega \gamma(\xi) m(t)(d\xi). \tag{1.17}$$

From the basic ingredients one can calculate the quantities $\mathcal{F}_{I_{|[t-a,t]}}(\xi,\omega)$ and $\lambda_{I_{|[t-a,t]}}(\xi,\omega)$ with the following interpretations: Let I be a given function of time, let $\xi \in \Omega$, and let ω be a measurable subset of Ω. Then we have the following.

- $\mathcal{F}_{I_{|[t-a,t]}}(\xi,\omega)$ is the probability that an individual who was born at time $t-a$ with i-state ξ is still alive at time t (when it has age a) and then has i-state in ω.
- $\lambda_{I_{|[t-a,t]}}(\xi,\omega)$ is the rate at which an individual who was born at time $t-a$ with i-state ξ produces offspring with state-at-birth in ω at time t (when it has age a).

The subscripts $I_{|[t-a,t]}$ of $\mathcal{F}$ and λ indicate that the quantities depend on the restriction of I to the interval $[t-a,t]$; that is, they depend only on the values of I during the lifetime of the individual in question.

Let $b(t)(\omega)$ denote the rate at which individuals are born with i-state in ω at time t. Assuming a maximal life span h, bookkeeping gives

$$b(t)(\omega) = \int_0^h \int_\Omega b(t-a)(d\xi)\lambda_{I_{|[t-a,t]}}(\xi,\omega)da, \tag{1.18}$$

$$m(t)(\omega) = \int_0^h \int_\Omega b(t-a)(d\xi)\mathcal{F}_{I_{|[t-a,t]}}(\xi,\omega)da. \tag{1.19}$$

Thus in this generality one has an abstract variant of (DE).

Often there is but one possible state-at-birth. Or, in particular when dealing with several interacting populations, there may be a finite number of possible states-at-birth. In such cases one may limit ω in (1.18) to points chosen from a finite set. If, in addition, $I(t)$ in (1.17) has only finitely many components, we can condense the essential information concerning the population problem into a finite dimensional equation (DE). Indeed, combining (1.18) and (1.19) with the feedback law (1.17), one finds that the value

$$x(t) = \begin{pmatrix} b(t) \\ I(t) \end{pmatrix} \tag{1.20}$$

is a nonlinear function of the history of x on $[t-h,t]$; that is, x satisfies a delay equation of the form (DE), with F being a function from $L^1\big([-h,0],\mathbf{R}^N\big)$ to $\mathbf{R}^N$ for some integer $N \geq 2$.

The population dynamical applications also motivate our choice of $L^1\left([-h,0],\mathbf{R}^N\right)$ as the history space. The components of b are *rates* at which individuals are born with certain i-states. While rates may be unbounded, *numbers* of individuals (integrals of rates) must remain finite.

The idea to use the history of I is new. The fact that in this manner we can use perturbation theory for dual semigroups to treat *general* physiologically structured population models, and not just age-structured models, triggered the writing of this paper. In a companion paper, to be written jointly with J. A. J. Metz, we shall elaborate in detail how the results of the present paper apply to population models.

In the present paper we shall consider only the case of finite delay. The reason is that in this case the semigroup defined by (1.7) has a desirable property called sun-reflexivity, which is lost in the case of infinite delay. However, our results can easily be extended to also encompass the case of infinite delay. In section 6 we briefly indicate how this can be done.

In this paper we follow a top-down approach. We start in section 2 by presenting the abstract perturbation theory for dual semigroups and then we formulate the principle of linearized stability which says that, under appropriate assumptions, local (in)stability of a steady state is completely determined by the spectral properties of the generator of the linearized semigroup. Under the extra assumption of finite dimensional range of the nonlinear perturbation G, we derive a characteristic equation, the roots of which are the spectral values of the generator of the linearized semigroup. We then give results on the stable, unstable, and center manifolds and on Hopf bifurcation. The results of section 2 are either known or slight modifications of known results. In section 3 we then specialize to the system (DE), (IC) and the associated unperturbed semigroup T_0 defined by (1.7) and verify that the assumptions made in section 2 indeed hold true. Models of structured populations often lead to delay equations coupled with delay differential equations. In section 4 we therefore consider such coupled systems. In section 5 we illustrate our theoretical results by two examples from population dynamics. We conclude in section 6 by relating our results to results by other authors and by discussing directions for future work.

2. Lipschitz perturbations in the sun-reflexive case. We start by briefly recalling the basic facts about dual semigroups. The books [4, 42, 47] are good general references, as are Chapter III and Appendix II of [23]. The theory of nonlinear Lipschitz continuous perturbations of generators of dual semigroups was first introduced in [11], where the principle of linearized stability was proved following [19]. The treatment of the stable, unstable, and center manifolds and of Hopf bifurcation follows [23].

2.1. Sun-reflexive dual semigroups. Let X be a real Banach space and $T_0 := \{T_0(t)\}_{t\geq 0}$ be a strongly continuous (i.e., the orbit $t \mapsto T_0(t)\varphi$ is continuous with respect to the norm topology on X for all initial values $\varphi \in X$) semigroup of bounded linear operators on X with infinitesimal generator A_0. Then $T_0^* := \{T_0^*(t)\}_{t\geq 0}$, where $T_0^*(t) : X^* \to X^*$ is the adjoint of $T_0(t)$, is a semigroup on the dual space X^* of X. T_0^* is called the *adjoint* or *dual* semigroup of T_0. If X is not reflexive, then T^* need not be strongly continuous. All one can say in general is that the orbits are continuous with respect to the weak* topology of X. At the level of generators this is reflected in the fact that the adjoint A_0^* of A_0 need not have dense domain and that A_0^* is the weak*-generator of T_0^*.

The maximal invariant subspace of X^* on which T_0^* is strongly continuous is denoted by $X^{\odot}$, that is,

$$X^{\odot} := \left\{ \varphi^* \in X^* : \lim_{t\downarrow 0} \|T_0^*(t)\varphi^* - \varphi^*\| = 0 \right\}. \tag{2.1}$$

Note that this so-called sun-subspace depends on the dynamical system one considers on the original space. It is known that

$$X^{\odot} = \overline{D(A_0^*)}, \tag{2.2}$$

where the bar denotes closure with respect to the norm topology of X^*. The operators $T_0^*(t)$, $t \geq 0$, leave $X^{\odot}$ invariant, and the restriction $T_0^{\odot}(t) := T_0^*(t)|_{X^{\odot}}$ of T_0^* to $X^{\odot}$ is a strongly continuous semigroup and its generator $A_0^{\odot}$ is the part of A_0^* in $X^{\odot}$; that is,

$$\mathcal{D}(A_0^{\odot}) := \{\varphi^{\odot} \in \mathcal{D}(A_0^*) : A_0^*\varphi^{\odot} \in X^{\odot}\}, \tag{2.3}$$

$$A_0^{\odot}\varphi^{\odot} := A_0^*\varphi^{\odot}. \tag{2.4}$$

We now have on $X^\odot$ exactly the same situation as we had on X at the outset. So in self-explanatory notation we obtain $X^{\odot *}$, $T_0^{\odot *}$, $A_0^{\odot *}$ and $X^{\odot\odot}$, $T_0^{\odot\odot}$, $A_0^{\odot\odot}$.

As usual, we denote the duality pairing between a Banach space X and its normed dual X^* by $\langle \cdot, \cdot \rangle$; that is, for $\varphi \in X$, $\varphi^* \in X^*$ we write $\langle \varphi, \varphi^* \rangle$ instead of $\varphi^*(\varphi)$. The formula

$$\langle \varphi^\odot, j\varphi \rangle = \langle \varphi, \varphi^\odot \rangle, \quad \varphi \in X, \ \varphi^\odot \in X^\odot \tag{2.5}$$

defines an embedding j of X into $X^{\odot *}$, the range of which lies in $X^{\odot\odot}$. Moreover, one has $T_0^{\odot *}(t)j = jT_0(t)$ for $t \geq 0$.

Definition 2.1. *A Banach space X is called* sun-reflexive *with respect to the strongly continuous linear semigroup T_0 if*

$$j(X) = X^{\odot\odot}.$$

From now on we shall always assume that X is sun-reflexive with respect to the unperturbed semigroup T_0.

2.2. Lipschitz perturbations and the nonlinear semigroup. Let $G : X \to X^{\odot *}$ be a nonlinear operator. The initial value problem

$$\frac{dju}{dt}(t) = A_0^{\odot *} ju(t) + G(u(t)), \quad t > 0, \tag{2.6}$$

$$u(0) = \varphi, \tag{2.7}$$

where u is an X-valued function, can be formally integrated to yield the abstract integral equation

$$u(t) = T_0(t)\varphi + j^{-1}\left(\int_0^t T_0^{\odot *}(t-s) G(u(s)) ds \right), \tag{AIE}$$

but we have to verify that the integral does indeed belong to $j(X)$.

The integral in (AIE) is to be interpreted in the weak*-sense. More precisely, if Z is a Banach space and $f : [a, b] \to Z^*$ is weakly*-continuous, then $\int_a^b f(t)dt$ is defined as the continuous linear functional on Z which takes $z \in Z$ to $\int_a^b \langle z, f(t) \rangle dt$. Note that $\int_a^b f(t)dt$ is an element of Z^*. For weak* integrals of the form

$$v(t) = \int_0^t T_0^{\odot *}(t-s) f(s) ds,$$

we have the following desirable result.

Proposition 2.2 (see [9, Theorem 3.2]). *If f is weakly*-continuous, then v is weakly*-continuous with values in $X^{\odot *}$. If f is norm continuous, then v is norm continuous as well and takes values in $X^{\odot\odot}$.*

We now consider (AIE). If G is globally Lipschitz continuous, then standard contraction mapping arguments yield existence and uniqueness of a solution $u(\cdot\,; \varphi) : \mathbf{R}_+ \to X$ of (AIE) for every $\varphi \in X$. The formula

$$\Sigma(t)\varphi := u(t; \varphi), \quad t \geq 0, \ \varphi \in X, \tag{2.8}$$

defines a strongly continuous nonlinear semigroup Σ on X. The generator of Σ, which we denote by C, is defined exactly as in the linear case: Its domain $\mathcal{D}(C)$ is the set of all $\varphi \in X$ for which the limit $\lim_{t\downarrow 0}(\Sigma(t)\varphi - \varphi)/t$ exists in the norm topology of X and $C\varphi$ is equal to this limit. The weak* generator $C^\times$ of Σ is defined as follows: $\varphi \in X$ belongs to $\mathcal{D}(C^\times)$ if $(j\Sigma(t)\varphi - j\varphi)/t$ converges to some $\varphi^{\odot *} \in X^{\odot *}$ as $t \downarrow 0$ and in this case $C^\times\varphi = \varphi^{\odot *}$.

THEOREM 2.3 (see [11, Theorems 3.2–3.6]).

(a) *$j\left(\mathcal{D}(C^\times)\right) = \mathcal{D}(A_0^{\odot *})$ and $C^\times\varphi = A_0^{\odot *}j\varphi + G(\varphi)$.*

(b) *C is the part of $C^\times$ in X, that is,*

$$\mathcal{D}(C) = \left\{\varphi \in X : \varphi \in \mathcal{D}(C^\times),\ C^\times\varphi \in j(X)\right\},$$

$$C\varphi = j^{-1}\left(C^\times\varphi\right).$$

(c) *If $\varphi \in \mathcal{D}(C)$ and if G is continuously Fréchet differentiable, then $t \mapsto u(t;\varphi) = \Sigma(t)\varphi$ is continuously differentiable and*

$$\frac{d}{dt}u(t;\varphi) = j^{-1}\left(A_0^{\odot *}ju(t;\varphi) + G(u(t;\varphi))\right).$$

2.3. Linearization around a steady state. In what follows we assume that the nonlinear operator $G : X \to X^{\odot *}$ is continuously Fréchet differentiable.

Assume that $\overline{\varphi} \in X$ is a steady state of the nonlinear dynamical system; that is,

$$\Sigma(t)\overline{\varphi} = \overline{\varphi} \tag{2.9}$$

for all $t \geq 0$. Equivalently, $j\overline{\varphi} \in \mathcal{D}(A_0^{\odot *})$ and

$$A_0^{\odot *}j\overline{\varphi} + G(\overline{\varphi}) = 0; \tag{2.10}$$

cf. Theorem 2.3(c). Because $G : X \to X^{\odot *}$ is Fréchet differentiable at $\overline{\varphi}$, its Fréchet derivative $B := G'(\overline{\varphi})$ is a bounded linear operator from X to $X^{\odot *}$. Formal linearization of (AIE) yields the following linear abstract integral equation:

$$T(t)\varphi = T_0(t)\varphi + j^{-1}\left(\int_0^t T_0^{\odot *}(t-\tau)BT(\tau)\varphi d\tau\right). \tag{LAIE}$$

For such equations the following result is known.

THEOREM 2.4 (see [9]). *The linear abstract integral equation* (LAIE) *uniquely defines a strongly continuous semigroup $T = \{T(t)\}_{t\geq 0}$ of bounded linear operators with generator A given by*

$$\mathcal{D}(A) = \{\varphi \in X : j\varphi \in \mathcal{D}(A_0^{\odot *}), A_0^{\odot *}j\varphi + B\varphi \in j(X)\},$$

$$A\varphi = j^{-1}(A_0^{\odot *}j\varphi + B\varphi).$$

That the formal linearization yields the desired result is the content of the following theorem.

THEOREM 2.5 (see [11]). *Let* (2.9) *hold and assume that the nonlinear operator $G : X \to X^{\odot *}$ is continuously Fréchet differentiable. Then for every $t > 0$ the nonlinear operator $\Sigma(t)$ is Fréchet differentiable at $\overline{\varphi}$. Its Fréchet derivative*

$$T(t) = (D\Sigma(t))(\overline{\varphi}) \tag{2.11}$$

defines a strongly continuous semigroup of bounded linear operators with generator A given by

$$\mathcal{D}(A) = \{\varphi \in X : j\varphi \in \mathcal{D}(A_0^{\odot *}), A_0^{\odot *}j\varphi + G'(\overline{\varphi})\varphi \in j(X)\},$$

$$A\varphi = j^{-1}(A_0^{\odot *}j\varphi + G'(\overline{\varphi})\varphi).$$

Moreover, for every $\varphi \in X$, $T(t)\varphi$ is the unique solution of (LAIE) *with $B = G'(\overline{\varphi})$.*

2.4. Eventual compactness and spectral analysis of the linearized semigroup. In subsection 2.6 we shall deal with criteria for the stability of a steady state. As is well known from the theory of ordinary differential equations (ODEs), *spectral analysis* of the linearized system is a most efficient tool for investigating stability. Therefore we shall in this subsection analyze the spectrum of the generator A of the semigroup T defined by (LAIE).

Our original nonlinear problem is meaningful only for *real* Banach spaces, whereas spectral analysis requires complex scalars. We therefore have to complexify X before doing spectral analysis. In the infinite dimensional case and, in particular, in our sun-star-framework, this is not a trivial task. We shall, however, omit the details because they can all be found in [23, section III.7].

As usual, we denote the resolvent set and the spectrum of a linear operator L by $\varrho(L)$ and $\sigma(L)$, respectively. The point spectrum of L, that is, the set of eigenvalues of L, is denoted by $P\sigma(L)$. The identity operator is denoted by E (to follow the tradition of Hilbert [14, formula (8), p. 5] and to avoid confusion with the input I of (1.17)), and Laplace transformation is denoted by $\widehat{\ }$:

$$\widehat{f}(\lambda) = \int_0^\infty e^{-\lambda t} f(t)dt. \tag{2.12}$$

$R(\lambda, L)$ denotes the *resolvent operator* of L:

$$R(\lambda, L) := (\lambda E - L)^{-1}, \quad \lambda \in \varrho(L). \tag{2.13}$$

Recall that $\lambda \mapsto R(\lambda, L)$ is a holomorphic operator-valued function on $\varrho(L)$. As for complex valued functions, an operator-valued function is *entire* if it is holomorphic in the whole complex plane.

The *growth bound* $\omega_0(T)$ of a semigroup T is defined by

$$\omega_0(T) = \inf\{\omega \in \mathbf{R} : \exists M_\omega \geq 1 \text{ such that } \|T(t)\| \leq M_\omega e^{\omega t} \text{ for all } t \geq 0\},$$

and the *spectral bound* $s(A)$ of its generator A is defined by

$$s(A) = \sup\{\operatorname{Re}\lambda : \lambda \in \sigma(A)\}.$$

One has $\sigma(A) = \sigma(A^*) = \sigma(A^\odot) = \sigma(A^{\odot *})$, $s(A) = s(A^*) = s(A^\odot) = s(A^{\odot *})$, and $\omega_0(T) = \omega_0(T^*) = \omega_0(T^\odot) = \omega_0(T^{\odot *})$ [26, Proposition 2.18, p. 262].

We start by characterizing the part of the point spectrum which belongs to $\varrho(A_0)$.

PROPOSITION 2.6. *Let A be the generator of the semigroup T defined by* (LAIE); *cf. Theorem 2.4. Then $\lambda \in \varrho(A_0)$ is an eigenvalue of A if and only if 1 is an eigenvalue of $j^{-1}R(\lambda, A_0^{\odot *})B$ and the corresponding eigenvectors are the same.*

Proof. Let $\psi \in X$. Using Theorem 2.4 we see that in the following sequence of identities, each implies both the preceding and the next one:

$$j^{-1}R(\lambda, A_0^{\odot *})B\psi = \psi,$$

$$j\psi \in \mathcal{D}(A_0^{\odot *}) \quad \text{and} \quad B\psi = \left(\lambda E - A_0^{\odot *}\right) j\psi,$$

$$j\psi \in \mathcal{D}(A_0^{\odot *}) \quad \text{and} \quad A_0^{\odot *} j\psi + B\psi = \lambda j\psi,$$

$$j\psi \in \mathcal{D}(A_0^{\odot *}) \quad \text{and} \quad j^{-1}\left(A_0^{\odot *} j\psi + B\psi\right) = \lambda\psi,$$

$$\psi \in \mathcal{D}(A) \quad \text{and} \quad A\psi = \lambda\psi. \qquad \square$$

If the semigroup T is *eventually compact*, that is, if the operators $T(t)$ are compact for all t greater than some $t_0 \geq 0$, then spectral analysis becomes as easy as one can possibly expect from an infinite dimensional system.

THEOREM 2.7. *Let A generate an eventually compact C_0-semigroup T on the Banach space X. Then*

$$\sigma(A) = P\sigma(A),$$

$$s(A) = \omega_0(T),$$

and every $\lambda \in \sigma(A)$ is a pole of the resolvent $R(\lambda, A)$ of finite algebraic multiplicity. Every right half-plane $\{\lambda \in \mathbf{C} : \alpha \leq \operatorname{Re}\lambda\}$ $(-\infty < \alpha)$ contains at most finitely many eigenvalues of A.

For a proof of this well-known result, see, e.g., [2, Theorem 2.1, p. 209].

Next we give a criterion for the eventual compactness of the perturbed semigroup which is easy to check and which applies to all our applications.

THEOREM 2.8. *Let T_0 be an eventually compact C_0-semigroup and let $B : X \to X^{\odot *}$ be compact. Then the C_0-semigroup T defined by* (LAIE) *is eventually compact.*

The corresponding result for the case in which B maps X into X is known [26, Proposition 1.14, p. 166], but Theorem 2.8 does not seem to have been stated in the literature yet. We therefore give a complete proof in the appendix.

2.5. Perturbations with finite dimensional range. If, as in the case of the delay problem (DE), (IC), the nonlinear perturbation G has finite dimensional range in $X^{\odot *}$, much of the analysis becomes considerably simpler, in fact, essentially finite dimensional. We therefore have a closer look at this special case. So let $G : X \to X^{\odot *}$ have the form

$$G(\varphi) = \sum_{i=1}^{N} F_i(\varphi) r_i^{\odot *}, \quad \varphi \in X, \tag{2.14}$$

where $F = (F_1, F_2, \ldots, F_N)$ is a mapping from X to $\mathbf{R}^N$ and $\{r_1^{\odot *}, r_2^{\odot *}, \ldots, r_N^{\odot *}\}$ is a linearly independent set in $X^{\odot *}$.

Note. In what follows we shall use the letter j both as a summation index and, as before, to denote the canonical embedding of X into $X^{\odot *}$, sometimes even in the same formula. This should not lead to any misunderstanding.

Clearly G is Fréchet differentiable at $\overline{\varphi}$ if and only if F is Fréchet differentiable at $\overline{\varphi}$, which is the case if and only if all the components F_i are Fréchet differentiable at $\overline{\varphi}$. So when G is Fréchet differentiable at $\overline{\varphi}$, there exist elements $r_1^*, r_2^*, \ldots, r_N^*$ of X^* such that the derivative $G'(\overline{\varphi})$ is the linear operator $B : X \to X^{\odot *}$ given by

$$B\varphi = \sum_{i=1}^{N} \langle \varphi, r_i^* \rangle r_i^{\odot *}, \quad \varphi \in X. \tag{2.15}$$

In order to exploit the finite dimensional structure of the perturbation we define

$$r_i(\lambda) = j^{-1} R(\lambda, A_0^{\odot *}) r_i^{\odot *}, \quad \lambda \in \varrho(A_0), \tag{2.16}$$

$$r_i^{\odot}(\lambda) = R(\lambda, A_0^*) r_i^*, \quad \lambda \in \varrho(A_0), \tag{2.17}$$

and let $M(\lambda)$ be the matrix with entries

$$M_{ij}(\lambda) = \langle r_j(\lambda), r_i^* \rangle, \quad \lambda \in \varrho(A_0). \tag{2.18}$$

Note that the matrix-valued function M is defined in $\varrho(A_0)$ only. When the real part of λ is greater than the growth bound of T_0, we can express $r_i(\lambda)$ and $r_i^{\odot}(\lambda)$ using the Laplace transform representation of the resolvent [26, Theorem 1.10, p. 55]:

$$r_i(\lambda) = j^{-1} \int_0^\infty e^{-\lambda t} T_0^{\odot *}(t) r_i^{\odot *} dt, \quad \operatorname{Re}\lambda > \omega_0(T_0), \tag{2.19}$$

$$r_i^{\odot}(\lambda) = \int_0^\infty e^{-\lambda t} T_0^*(t) r_i^* dt, \quad \operatorname{Re}\lambda > \omega_0(T_0). \tag{2.20}$$

We start with a few lemmas.

LEMMA 2.9. *Let M be the matrix-valued function defined by* (2.18). *Then*

$$M_{ij}(\lambda) = \left\langle r_i^{\odot}(\lambda), r_j^{\odot *} \right\rangle, \quad \lambda \in \varrho(A_0). \tag{2.21}$$

Proof. We first prove the claim for $\operatorname{Re}\lambda > \omega_0(T_0)$ using the representations (2.19) and (2.20). If $r_i^* \in X^{\odot}$, the equality of the right-hand sides of (2.18) and (2.21) is clear from the definition of the weak*-integral. Next we approximate r_i^* by

$$\varphi_s^{\odot} = \frac{1}{s} \int_0^s T_0^*(\tau) r_i^* d\tau, \quad s > 0.$$

It follows from Proposition 2.2 (just interchange the roles of X and $X^{\odot}$) that $\varphi_s^{\odot} \in X^{\odot}$ for all $s > 0$. By the observation made above, one has

$$\left\langle j^{-1} \int_0^\infty e^{-\lambda t} T_0^{\odot *}(t) r_i^{\odot *} dt, \varphi_s^{\odot} \right\rangle = \left\langle \int_0^\infty e^{-\lambda t} T_0^{\odot}(t) \varphi_s^{\odot} dt, r_i^{\odot *} \right\rangle \tag{2.22}$$

for all $s > 0$. A straightforward calculation (see the proof of Lemma 2.17 in [23, p. 61] for a very similar case) shows that the left-hand side of (2.22) converges to (2.18) and that the right-hand side of (2.22) converges to the right-hand side of (2.21). This proves the assertion for the case $\operatorname{Re}\lambda > \omega_0(T_0)$. The general case follows from the resolvent identity. ∎

When B has finite dimensional range we get a more detailed description of the point spectrum of A than we do in Proposition 2.6.

LEMMA 2.10. *Let A be the generator of the semigroup T defined by* (LAIE) *and assume that B has the form* (2.15). *If $\lambda \in \varrho(A_0)$ and $\psi \in X$, then*

$$A\psi = \lambda\psi \tag{2.23}$$

if and only if

$$\psi = \sum_{i=1}^N c_i r_i(\lambda), \tag{2.24}$$

where the coefficients c_i are the components of a vector c satisfying

$$M(\lambda)c = c \tag{2.25}$$

and $M(\lambda)$ *is the matrix defined by* (2.18).

Proof. By Proposition 2.6, $A\psi = \psi$ if and only if

$$j^{-1}R(\lambda, A_0^{\odot *})B\psi = \psi. \tag{2.26}$$

Because the vectors $r_1^{\odot *}, r_2^{\odot *}, \dots, r_N^{\odot *}$ are linearly independent and $j^{-1}R(\lambda, A_0^{\odot *})$ is one-to-one, the definition (2.16) shows that also the vectors $r_1(\lambda), r_2(\lambda), \dots, r_N(\lambda)$ are linearly independent. Equation (2.26) shows that ψ belongs to the subspace spanned by the vectors $r_i(\lambda)$, $i = 1, 2, \dots, N$; that is, ψ is of the form (2.24). Substituting (2.24) into (2.26), one obtains (2.25). ☐

The following dual version of Lemma 2.10 is proved analogously.

LEMMA 2.11. *Let A be the generator of the semigroup T defined by* (LAIE) *and assume that B has the form* (2.15). *If $\lambda \in \varrho(A_0)$ and $\psi^\odot \in X^\odot$, then*

$$A^*\psi^\odot = \lambda\psi^\odot \tag{2.27}$$

if and only if

$$\psi^\odot = \sum_{i=1}^{N} d_i r_i^\odot(\lambda), \tag{2.28}$$

where the coefficients d_i are the components of a row vector d satisfying

$$dM(\lambda) = d \tag{2.29}$$

and $M(\lambda)$ *is the matrix defined by* (2.18).

LEMMA 2.12. *The mapping $\lambda \mapsto \langle r_j(\lambda), r_i^*\rangle$ is holomorphic in $\varrho(A_0)$ and*

$$\frac{d}{d\lambda}\langle r_j(\lambda), r_i^*\rangle = -\left\langle r_j(\lambda), r_i^\odot(\lambda)\right\rangle, \quad \lambda \in \varrho(A_0). \tag{2.30}$$

Proof. Using the resolvent identity, one finds that

$$\begin{aligned}\frac{1}{\lambda-\mu}\left(\langle r_j(\lambda), r_i^*\rangle - \langle r_j(\mu), r_i^*\rangle\right) &= -\left\langle j^{-1}R(\lambda, A_0^{\odot *})R(\mu, A_0^{\odot *})r_j^{\odot *}, r_i^*\right\rangle \\ &= -\left\langle j^{-1}R(\mu, A_0^{\odot *})r_j^{\odot *}, R(\lambda, A_0^*)r_i^*\right\rangle \\ &= -\left\langle r_j(\mu), r_i^\odot(\lambda)\right\rangle,\end{aligned}$$

which proves the assertion. ☐

As a direct consequence of the three preceding lemmas we obtain the following result.

COROLLARY 2.13. *The matrix-valued function $\lambda \mapsto M(\lambda)$ is holomorphic in $\varrho(A_0)$, and if λ is an eigenvalue of A with eigenvector ψ and adjoint eigenvector $\psi^\odot$, then*

$$\left\langle \psi, \psi^\odot\right\rangle = -dM'(\lambda)c, \quad \lambda \in \varrho(A_0), \tag{2.31}$$

where c and d are as described in Lemmas 2.10 *and* 2.11, *respectively.*

Corollary 2.13 provides a convenient criterion for the simplicity of an eigenvalue, which we shall use in the context of the Hopf bifurcation theorem to be treated in subsection 2.8. In the present subsection we shall show that when B has finite dimensional range, there exists a so-called characteristic equation, the roots of which are the eigenvalues of the generator of the perturbed semigroup. It turns out that the order of λ as a root of the characteristic equation equals the algebraic multiplicity of λ as an eigenvalue of A. An easy way to show this is to use the theory of Weinstein–Aronszajn determinants; see [43, section IV.6] for an account of the general theory and [22] for an application to perturbed dual semigroups. Before we can present the Weinstein–Aronszajn formula we have to define the multiplicity functions for closed operators and meromorphic functions.

Let L be a closed operator in a Banach space. For every isolated point λ of $\sigma(L)$ we denote the spectral projection onto the corresponding generalized eigenspace by P_λ. The *multiplicity function* $\widetilde{\nu}(\lambda, L)$ of L is defined as

$$\widetilde{\nu}(\lambda, L) = \begin{cases} 0 & \text{if } \lambda \in \varrho(L), \\ \dim \mathcal{R}(P_\lambda) & \text{if } \lambda \text{ is an isolated point of } \sigma(L), \\ \infty & \text{in all other cases.} \end{cases} \tag{2.32}$$

The multiplicity function of a (numerical) meromorphic function f is defined as

$$\nu(\lambda, f) = \begin{cases} k & \text{if } \lambda \text{ is a zero of order } k \text{ of } f, \\ -k & \text{if } \lambda \text{ is a pole of order } k \text{ of } f, \\ 0 & \text{otherwise.} \end{cases} \tag{2.33}$$

THEOREM 2.14 (Weinstein–Aronszajn formula). *Let A be the generator of the semigroup T defined by* (LAIE), *assume that B has the form* (2.15), *and let $M(\lambda)$ be the matrix-valued function defined in* (2.18). *Then*

$$\widetilde{\nu}(\lambda, A) = \widetilde{\nu}(\lambda, A_0) + \nu\left(\lambda, \det\left(E - M(\lambda)\right)\right). \tag{2.34}$$

Proof. Because B has finite dimensional range one can unambiguously define the so-called Weinstein–Aronszajn determinant $\det\left(E - BR\left(\lambda, A_0^{\odot *}\right)\right)$ as the determinant of the restriction of $E - BR\left(\lambda, A_0^{\odot *}\right)$ to $\mathcal{R}(B)$. The definition of M together with Lemma 2.9 shows that

$$\det\left(E - BR\left(\lambda, A_0^{\odot *}\right)\right) = \det\left(E - M(\lambda)\right).$$

The Weinstein–Aronszajn formula [43, Theorem IV.6.2] now yields

$$\widetilde{\nu}(\lambda, A^{\odot *}) = \widetilde{\nu}(\lambda, A_0^{\odot *}) + \nu\left(\lambda, \det\left(E - M(\lambda)\right)\right). \tag{2.35}$$

But A_0 (resp., A) is the part of $A_0^{\odot *}$ (resp., $A^{\odot *}$) in $j(X)$, and hence it follows from [26, Lemma 1.15, p. 245 and Proposition 2.17, p. 261] that

$$\widetilde{\nu}(\lambda, A_0^{\odot *}) = \widetilde{\nu}(\lambda, A_0),$$

$$\widetilde{\nu}(\lambda, A^{\odot *}) = \widetilde{\nu}(\lambda, A),$$

from which the conclusion (2.34) follows. ∎

We are now ready to prove the following theorem.

THEOREM 2.15. *Let A be the generator of the semigroup T defined by* (LAIE). *Suppose that B has the form* (2.15) *and let M be the corresponding matrix-valued function defined by* (2.18). *Then $\lambda \in \varrho(A_0)$ is in $\sigma(A)$ if and only if*

$$\det(E - M(\lambda)) = 0, \tag{2.36}$$

where E denotes the $N \times N$ identity matrix. Moreover, when this is the case, λ belongs to $P\sigma(A)$ and the algebraic multiplicity of λ equals the order of λ as a root of (2.36).

In particular, if $\sigma(A_0) = \emptyset$, then

$$\sigma(A) = P\sigma(A) = \{\lambda \in \mathbf{C} : \det(E - M(\lambda)) = 0\}. \tag{2.37}$$

Proof. Taking the Laplace transform of (LAIE), one obtains

$$R(\lambda, A) = R(\lambda, A_0) + j^{-1}R(\lambda, A_0^{\odot *})BR(\lambda, A)$$

or

$$\left(E - j^{-1}R(\lambda, A_0^{\odot *})B\right) R(\lambda, A) = R(\lambda, A_0). \tag{2.38}$$

From (2.38) we deduce that if $\lambda \in \varrho(A_0)$, then $\lambda \in \sigma(A)$ if and only if

$$E - j^{-1}R(\lambda, A_0^{\odot *})B$$

is not invertible. But because $j^{-1}R(\lambda, A_0^{\odot *})B$ is a bounded linear operator $X \to X$ with finite dimensional range, this is the case if and only if 1 is an eigenvalue of $j^{-1}R(\lambda, A_0^{\odot *})B$. According to Proposition 2.6, this in turn is equivalent to λ being an eigenvalue of A, which by Lemma 2.10 is equivalent to 1 being an eigenvalue of $M(\lambda)$. This shows that $\lambda \in \sigma(A)$ if and only if λ is a root of (2.36) and that then $\lambda \in P\sigma(A)$.

Since $\widetilde{\nu}(\lambda, A_0)$ is zero, the assertion concerning the multiplicity of λ follows from Theorem 2.14.

The final assertion is obvious, because if $\sigma(A_0)$ is empty, then the basic assumption $\lambda \in \varrho(A_0)$ is automatically satisfied. □

Equation (2.36) is called the *characteristic equation*.

Remark 2.16. It was shown in [9, Lemma 5.1] that there exists a matrix-valued function $k \in L^\infty_{\text{loc}}(\mathbf{R}_+, \mathbf{R}^{N\times N})$ such that

$$\left\langle j^{-1}\left(\int_0^t T_0^{\odot *}(t-\tau)r_j^{\odot *}\eta(\tau)d\tau\right), r_i^*\right\rangle = \int_0^t k_{ij}(t-\tau)\eta(\tau)d\tau$$

for all $\eta \in L^1_{\text{loc}}(\mathbf{R}_+)$. From this it follows easily that

$$\widehat{k}(\lambda) = M(\lambda), \quad \lambda \in \varrho(A_0),$$

where M is the matrix-valued function defined by (2.18). The characteristic equation can thus be rewritten as

$$\det\left(E - \widehat{k}(\lambda)\right) = 0. \tag{2.39}$$

In subsection 3.4 we shall compute the matrix k explicitly in the concrete case connected to (DE).

Using properties of the Laplace transform and holomorphic functions (in particular, the Riemann–Lebesgue lemma and the fact that the zeros of holomorphic functions have no limit points) it is possible to prove directly (in the case $\sigma(A_0) = \emptyset$) that there are only finitely many eigenvalues in each right half-plane. But because we obtain this result from Theorem 2.7 in all our applications, we have refrained from stating it in Theorem 2.15.

As the proof of Theorem 2.15 shows, the existence of a characteristic equation depends on two facts: the analyticity in the whole complex plane of the resolvent of the generator A_0 of the unperturbed semigroup and the finite dimensionality of the range of the perturbation. We shall later encounter applications where $R(\lambda, A_0)$ has a simple pole at the origin. Anticipating this situation, we next show that the corresponding singularity of $R(\lambda, A)$ is removable and that we still get a characteristic equation.

THEOREM 2.17. *Let B be given by* (2.15) *and assume that*

$$R(\lambda, A_0^{\odot *}) = \frac{1}{\lambda} P H_1(\lambda) + (E - P) H_2(\lambda), \tag{2.40}$$

where $P : X^{\odot *} \to X^{\odot *}$ is a projection with finite dimensional range in $j(X)$, H_1 and H_2 are entire functions with values in $\mathcal{L}(X^{\odot *})$, and the range of $H_2(\lambda)$ is in $j(X)$. Then $\sigma(A) = P\sigma(A)$, and there exists an entire matrix-valued function Δ such that $\lambda \in \sigma(A)$ if and only if

$$\det \Delta(\lambda) = 0. \tag{2.41}$$

Proof. The assumption (2.40) implies that

$$j^{-1} R(\lambda, A_0^{\odot *}) B = \left(E - j^{-1} P j + \frac{1}{\lambda} j^{-1} P j \right) K(\lambda) \tag{2.42}$$

for the entire function K defined by

$$K(\lambda)\varphi = \sum_{i=1}^{N} \langle \varphi, r_i^* \rangle \, j^{-1} \left(P H_1(\lambda) + (E - P) H_2(\lambda) \right) r_i^{\odot *}$$

with values in the subspace of finite rank operators of $\mathcal{L}(X)$. It follows (cf. (2.38)) that

$$\begin{aligned} &\left(E - \left(E - j^{-1} P j + \frac{1}{\lambda} j^{-1} P j \right) K(\lambda) \right) R(\lambda, A) \\ &\qquad = j^{-1} \left(\frac{1}{\lambda} P H_1(\lambda) + (E - P) H_2(\lambda) \right) j. \end{aligned} \tag{2.43}$$

If one multiplies (2.43) by $E - j^{-1} P j + \lambda j^{-1} P j$, one obtains

$$(E - j^{-1} P j + \lambda j^{-1} P j - K(\lambda)) R(\lambda, A) = j^{-1} \left(P H_1(\lambda) + (E - P) H_2(\lambda) \right) j. \tag{2.44}$$

Because the right-hand side of (2.44) is entire, $R(\lambda, A)$ is holomorphic everywhere except at the points where $(E - j^{-1} P j + \lambda j^{-1} P j - K(\lambda))$ is not invertible. Because $j^{-1} P j - \lambda j^{-1} P j + K(\lambda)$ has finite dimensional range and is everywhere holomorphic, it follows as in the proof of Theorem 2.15 that there is an entire matrix $\Delta(\lambda)$ such that $(E - j^{-1} P j + \lambda j^{-1} P j - K(\lambda))$ is not invertible if and only if $\Delta(\lambda)$ is not invertible, that is, if and only if (2.41) holds. ∎

2.6. Linearized stability. Recall that a steady state $\overline{\varphi}$ of Σ is (locally) *stable* if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\|\Sigma(t)\varphi - \overline{\varphi}\| \leq \varepsilon \quad \text{for all } t \geq 0$$

whenever $\|\varphi - \overline{\varphi}\| \leq \delta$. If $\overline{\varphi}$ is not stable, it is *unstable*. It is (locally) *exponentially stable* if there exist numbers $\delta > 0,\ K > 0,\ \alpha > 0$ such that

$$\|\Sigma(t)\varphi - \overline{\varphi}\| \leq Ke^{-\alpha t}, \quad t \geq 0,$$

for all φ with $\|\varphi - \overline{\varphi}\| \leq \delta$.

The next result is called the *principle of linearized stability.* It has two parts. The first states that if the growth bound of the linearized semigroup is negative, then the steady state is exponentially stable. The second part states that if the generator of the linearized semigroup has at least one eigenvalue of finite multiplicity with positive real part, then the steady state is not stable.

Theorem 2.18 (see [19], [11, Theorems 4.2 and 4.3], [23, Corollary 5.12]). *Let Σ be a strongly continuous nonlinear semigroup. Let $\overline{\varphi}$ be a steady state of Σ and assume that for each $t \geq 0$, $\Sigma(t)$ has a (uniform) Fréchet derivative $T(t)$ at $\overline{\varphi}$. Let A be the infinitesimal generator of T. Assume further that X admits a decomposition*

$$X = X_- \oplus X_+$$

into two $T(t)$-invariant subspaces X_- and X_+ such that

(i) *X_+ is finite dimensional,*

(ii) *the restriction of $T(t)$ to X_- converges exponentially to* 0 *as $t \to \infty$.*

Then $\overline{\varphi}$ is

(a) *(locally) exponentially stable if* $\operatorname{Re} \lambda < 0$ *for all $\lambda \in \sigma(A\big|_{X_+})$,*

(b) *unstable if there exists a $\lambda \in \sigma(A\big|_{X_+})$ with* $\operatorname{Re} \lambda > 0$.

Note that if $\omega_0(T) < 0$, then (i) and (ii) are satisfied with X_+ equal to the trivial subspace $\{0\}$, and hence $\overline{\varphi}$ is exponentially stable because $\sigma(A\big|_{X_+})$ is empty.

Theorem 2.5 shows that the differentiability assumption of Theorem 2.18 is indeed satisfied for the semigroup Σ generated by the abstract integral equation (AIE).

When applied to the nonlinear semigroup Σ generated by the abstract integral equation (AIE), Theorem 2.18 becomes particularly simple to apply if T_0 is eventually compact and $G'(\overline{\varphi})$ is compact (in particular if G has finite dimensional range). Indeed, Theorem 2.7 immediately implies the following corollary.

Corollary 2.19. *Assume that $G : X \to X^{\odot *}$ is continuously Fréchet differentiable. Let Σ be the nonlinear semigroup "generated" by* (AIE) *(i.e., defined through* (2.8)*) and let A be the generator of the linearized semigroup T as in Theorem* 2.5. *Let $\overline{\varphi}$ be a steady state of Σ.*

If T_0 is eventually compact and if $G'(\overline{\varphi})$ is compact, then $\overline{\varphi}$ is locally exponentially stable if all $\lambda \in \sigma(A)$ have real part less than zero, whereas if there exists at least one $\lambda \in \sigma(A)$ with positive real part, then $\overline{\varphi}$ is unstable.

Proof. By Theorem 2.8 T is eventually compact, and hence the growth bound equals the spectral bound (Theorem 2.7). Thus, if all $\lambda \in \sigma(A) = P\sigma(A)$ have negative real part, then $\omega_0(T) < 0$ and, as noted after Theorem 2.18, $\overline{\varphi}$ is exponentially stable. If there exists an eigenvalue with positive real part, there exist finitely many eigenvalues with positive real part, and they all have generalized eigenspaces of finite dimension (Theorem 2.7). Therefore there exists a decomposition as in Theorem 2.18

with X_+ finite dimensional and a $\lambda \in \sigma(A\big|_{X_+})$ with $\operatorname{Re}\lambda > 0$. An application of Theorem 2.18 completes the proof. $\square$

We close this section by applying a version of the argument principle [52, Theorem 10.43(a)], also known as Nyquist's theorem, to derive a very convenient criterion for the (in)stability of steady states in the case where the perturbation has finite dimensional range and we have a characteristic equation. Nyquist's theorem states that if the matrix-valued function k belongs to $L^1(\mathbf{R}_+)$ and satisfies $\det\big(E - \widehat{k}(i\omega)\big) \neq 0$ for all $\omega \in \mathbf{R}$, then the number of zeros of $\det\big(E - \widehat{k}(\lambda)\big)$ in the open right half-plane $\operatorname{Re}\lambda > 0$, counted according to their multiplicities, equals the index $\operatorname{Ind}_\Gamma(0)$ of the curve $\Gamma : \omega \mapsto \det\big(E - \widehat{k}(i\omega)\big)$, where ω runs from $+\infty$ to $-\infty$ [29, Theorem 6.3, p. 61]. Recall the geometrical interpretation of the index $\operatorname{Ind}_\Gamma(0)$: it is the number of times the curve Γ winds counterclockwise around the origin as ω runs from $+\infty$ to $-\infty$.

COROLLARY 2.20. *Assume in addition to the hypotheses of Corollary* 2.19 *that* T_0 *is nilpotent and that* G *has finite dimensional range. Let* $M(\lambda) = \widehat{k}(\lambda)$ *be the matrix-valued function as defined by* (2.18) *and Remark* 2.16 *and let* Γ *be the curve defined above. If the characteristic equation* (2.39) *has no roots on the imaginary axis, then* $\overline{\varphi}$ *is exponentially stable if* $\operatorname{Ind}_\Gamma(0) = 0$ *and unstable if* $\operatorname{Ind}_\Gamma(0) > 0$.

Proof. The nilpotency of T_0 implies that k has compact support and hence (being locally L^∞) belongs to $L^1(\mathbf{R}_+)$. The conclusion now follows from Nyquist's theorem. $\square$

The assumption that T_0 is nilpotent is much stronger than is actually needed, but it is a convenient assumption that is satisfied in many applications (including structured populations with a maximum individual life span). The key point is that when we extend the argument principle from integration along closed curves to integration along the imaginary axis, we need to control the behavior of the integrand at infinity. The assumption $k \in L^1$ makes the Riemann–Lebesgue lemma valid and gives an easy estimate of the behavior at infinity.

The stability criterion of Corollary 2.20 is easy to implement numerically and even graphically. By the Riemann–Lebesgue lemma, $\widehat{k}(i\omega)$ tends to 0 as $\omega \to \pm\infty$, and hence $\det\big(E - \widehat{k}(i\omega)\big)$ tends to 1 as $\omega \to \pm\infty$. Choose ω_0 so large that $\det\big(E - \widehat{k}(i\omega_0)\big)$ is close to 1 for $|\omega| > \omega_0$ and plot $\det\big(E - \widehat{k}(i\omega)\big)$ as ω runs from $+i\omega_0$ to $-i\omega_0$. If the plotted curve does not wind around the origin, then $\overline{\varphi}$ is exponentially stable; otherwise it is unstable. If the curve passes through the origin, the test does not give any information.

2.7. The unstable, stable, and center manifolds. It is possible to give a more detailed description of the behavior near an unstable steady state. For the linearized semigroup, one has, provided that the characteristic equation has no roots on the imaginary axis, a direct sum spectral decomposition into a finite dimensional unstable subspace X_+ and an infinite dimensional stable subspace X_-. On X_+ one can go backwards in time. As a matter of fact, X_+ is *characterized* by the property that the orbit through a point in X_+ can be extended in the negative time-direction to $-\infty$ and that the α-limit set equals $\{0\}$. Similarly, X_- consists of precisely those points that have $\{0\}$ as ω-limit set. A general orbit shows saddle-point behavior: It may come close to 0 but will eventually move far away and, if it can be extended in the negative time-direction, it will also move far away in that direction.

One can construct a finite dimensional local unstable manifold $\mathcal{W}_u$ as the graph of a smooth function from X_+ to X_-, shifted to $\overline{\varphi}$. The manifold $\mathcal{W}_u$ is invariant, and the tangent space at $\overline{\varphi}$ is exactly X_+. Moreover, an orbit starting in a sufficiently small ball around $\overline{\varphi}$ can be extended to $t = -\infty$ with α-limit set equal to $\{\overline{\varphi}\}$ if and only if it starts (and hence remains) in $\mathcal{W}_u$. We refer to [23, Chapter VIII] for precise formulations (see in particular Theorems 4.4 and 4.7 and Corollary 4.11). Similarly, one can construct and characterize the local stable manifold $\mathcal{W}_s$ [23, Chapter VIII, Theorem 6.1].

If A does have a spectrum on the imaginary axis, the spectral decomposition involves a third component X_0, which in the setting of Theorem 2.15 or Theorem 2.17 is finite dimensional. The orbits of the linearized semigroup that start in X_0 are characterized by the fact that they grow at most polynomially as $t \to \pm\infty$ (note that in X_0 orbits can be extended to $t = -\infty$). As this characterization is more difficult to work with, the construction of the corresponding center manifold for the nonlinear semigroup (and the proof of its smoothness) is much more involved. Moreover, modification of the nonlinearity outside a small ball around $\overline{\varphi}$ plays a role in the construction and as a consequence the center manifold is not unique (yet it will contain all solutions which are defined for all times and remain inside the small ball for all times). We refer to [23, Chapter IX] for detailed formulations and proofs that apply verbatim to the setting of Theorem 2.15 or Theorem 2.17.

A situation of particular interest is the case that the nonlinear semigroup depends on a parameter and that for a specific value of this parameter, the characteristic equation (2.39) has a pair of simple roots on the imaginary axis (note that since the kernel k takes on real values, $\widehat{k}(-i\omega) = \overline{\widehat{k}(i\omega)}$, and hence complex roots of (2.39) occur in conjugate pairs). Under some further mild genericity conditions one then finds periodic orbits for nearby parameter values. Chapter X of [23] gives a detailed treatment of this so-called Hopf bifurcation in the setting of exactly the abstract integral equation (AIE) that we consider here. We present the main result in the next subsection and at the end of subsection 3.4 we shall briefly indicate how to obtain a corollary for Volterra functional equations.

2.8. Hopf bifurcation. In this subsection we consider Hopf bifurcation under the assumption that the nonlinear perturbation $G : X \to X^{\odot *}$ has finite dimensional range, which does *not* depend on the bifurcation parameter θ. So G is of the form

$$G(\varphi, \theta) = \sum_{i=1}^{N} F_i(\varphi, \theta) r_i^{\odot *}, \tag{2.45}$$

and its derivative with respect to φ at 0 is

$$B(\theta)\varphi = \sum_{i=1}^{N} \langle \varphi, r_i^*(\theta) \rangle \, r_i^{\odot *}. \tag{2.46}$$

Note carefully that now the vector r_i^* depends on the bifurcation parameter θ, as do the vector $r_i^{\odot}(\lambda)$ and the matrix $M(\lambda)$ introduced in (2.17) and (2.18), respectively:

$$M_{ij}(\lambda, \theta) = \langle r_j(\lambda), r_i^*(\theta) \rangle = \left\langle r_i^{\odot}(\lambda, \theta), r_j^{\odot *} \right\rangle. \tag{2.47}$$

In order to have Hopf bifurcation, we need to make sure that a conjugate pair $\pm i\omega_0$ of simple eigenvalues crosses the imaginary axis with positive speed as the bifurcation parameter θ passes some value θ_0. (Note: The real number ω_0 used in this subsection has of course nothing to do with the growth bound of a semigroup. We use the same symbol to denote two unrelated numbers because in both cases the usage conforms with common practice. No confusion is expected to arise.)

The simplicity of the eigenvalues is, by Corollary 2.13 and Theorem 2.15, ensured by the condition

$$\left\langle \psi(\theta_0), \psi^{\odot}(\theta_0) \right\rangle = -d(\theta_0)\frac{\partial M}{\partial \lambda}(i\omega_0, \theta_0)c(\theta_0) \neq 0. \tag{2.48}$$

The condition of crossing the imaginary axis with positive speed means more precisely that

$$\operatorname{Re} \lambda'(\theta_0) \neq 0, \tag{2.49}$$

where $\lambda(\theta)$ is a branch of eigenvalues through $i\omega_0$ at $\theta = \theta_0$. To derive a verifiable form of this condition, let $c(\theta)$ and $d(\theta)$ be the right and left eigenvectors, respectively, of $M(\lambda, \theta)$ normalized by

$$|c(\theta)| = \sum_{i=1}^{N} |c_i(\theta)| = 1, \tag{2.50}$$

$$d(\theta)c(\theta) = 1. \tag{2.51}$$

Differentiating the equation

$$d(\theta)M(\lambda, \theta)c(\theta) = 1$$

implicitly with respect to θ, one obtains

$$\frac{d}{d\theta}\left(d(\theta)c(\theta)\right) + d(\theta)\frac{\partial M}{\partial \theta}c(\theta) + d(\theta)\frac{\partial M}{\partial \lambda}c(\theta)\lambda'(\theta) = 0. \tag{2.52}$$

It now follows from (2.51) and Corollary 2.13 that

$$d(\theta)\frac{\partial M}{\partial \theta}(\lambda(\theta), \theta)c(\theta) = \left\langle \psi(\theta), \psi^{\odot}(\theta) \right\rangle \lambda'(\theta). \tag{2.53}$$

From (2.48) and (2.53) we deduce that (2.49) holds if and only if

$$\operatorname{Re} d(\theta_0)\frac{\partial M}{\partial \theta}(i\omega_0, \theta_0)c(\theta_0) \neq 0. \tag{2.54}$$

We are now ready to formulate the Hopf bifurcation theorem.

THEOREM 2.21 (Hopf bifurcation theorem, [23, Theorem 2.6, p. 290]). *Consider the abstract integral equation*

$$u(t) = T_0(t-s)u(s) + j^{-1}\int_s^t T_0^{\odot *}(t-\tau)G(u(\tau), \theta)d\tau, \quad -\infty < s \leq t < \infty, \tag{AIE}$$

and assume that the following hold:

(H1) $G(\varphi, \theta) = \sum_{i=1}^{N} F_i(\varphi, \theta)r_i^{\odot *}$, $F : X \times \mathbf{R} \to \mathbf{R}^N$ *is* C^k, $k \geq 2$.

(H2) $F(0, \theta) = 0$ *for all* θ.

(H3) $D_1G(0, \theta) = B(\theta)$ *with* $B(\theta)$ *defined by* (2.46). *The corresponding matrix* $M(\lambda, \theta)$ *has for* $\lambda = \pm i\omega_0$, $\theta = \theta_0$, *eigenvalue* 1 *with right eigenvector* $c(\theta_0)$ *and left eigenvector* $d(\theta_0)$ *and* $d(\theta_0)\frac{\partial}{\partial \lambda}M(i\omega_0, \theta_0)c(\theta_0) \neq 0$. *For* $\theta = \theta_0$, *no root of the characteristic equation* $\det(E - M(\lambda)) = 0$ *other than* $\lambda = \pm i\omega_0$ *belongs to* $i\omega_0\mathbf{Z}$.

(H4) $\operatorname{Re} d(\theta_0)\frac{\partial}{\partial\theta}M(i\omega_0,\theta_0)c(\theta_0) \neq 0$.

There exist C^{k-1} functions $\varepsilon \mapsto \widetilde{\theta}(\varepsilon)$, $\varepsilon \mapsto \widetilde{\psi}(\varepsilon)$, and $\varepsilon \mapsto \widetilde{\omega}(\varepsilon)$ with values in $\mathbf{R}$, X, and $\mathbf{R}$, respectively, defined for ε sufficiently small, such that the solution of (AIE) *with $u(0) = \widetilde{\psi}(\varepsilon)$ is $2\pi/\widetilde{\omega}(\varepsilon)$ periodic. Moreover, $\widetilde{\theta}$ and $\widetilde{\omega}$ are even functions, $\widetilde{\theta}(0) = \theta_0$, $\widetilde{\omega}(0) = \omega_0$, $\widetilde{\psi}(-\varepsilon) = \widetilde{\psi}(\varepsilon + \frac{\pi}{\widetilde{\omega}(\varepsilon)})$. If $u(t)$ is any small periodic solution of* (AIE) *for θ close to θ_0 and period close to $2\pi/\omega_0$, then necessarily $\theta = \widetilde{\theta}(\varepsilon)$ for some ε and there exists $\sigma \in [0, 2\pi/\widetilde{\omega}(\varepsilon))$ such that $u(\sigma) = \widetilde{\psi}(\varepsilon)$. If for $\theta = \theta_0$ all roots λ of the characteristic equation*

$$\det(E - M(\lambda, \theta_0)) = 0$$

other than $\pm i\omega_0$ lie in the left half-plane and $\operatorname{Re} d(\theta_0)\frac{\partial}{\partial\theta}M(i\omega_0,\theta_0)c(\theta_0) < 0$, then the periodic solution is, for ε sufficiently small, asymptotically stable with asymptotic phase if $\widetilde{\theta}(\varepsilon) > \theta_0$ and unstable if $\widetilde{\theta}(\varepsilon) < \theta_0$.

Remark 2.22. In order to determine the direction of the bifurcation one has to compute the second derivative of $\widetilde{\theta}$. How this is done is explained in [23, section X.3].

3. Volterra functional equations.

3.1. Unperturbed semigroup for systems of delay equations. We let h denote a positive real number and N a positive integer. $X = L^1\left([-h, 0]; \mathbf{R}^N\right)$ is the space of all (equivalence classes of) $\mathbf{R}^N$-valued measurable functions φ defined and absolutely integrable on $[-h, 0]$ (i.e., each component φ_i, $i = 1, 2, \ldots, N$, is absolutely integrable) with norm

$$\|\varphi\|_1 := \sum_{i=1}^{N} \|\varphi_i\|_1. \tag{3.1}$$

The dual space X^* of $X = L^1\left([-h, 0]; \mathbf{R}^N\right)$ is represented by $L^\infty\left([0, h]; \mathbf{R}^N\right)$, that is, the space of (equivalence classes of) $\mathbf{R}^N$-valued essentially bounded measurable functions g with norm

$$\|g\|_\infty := \max_{1\le i\le N} \|g_i\|_\infty \tag{3.2}$$

via the duality pairing

$$\langle \varphi, g\rangle = \sum_{i=1}^{N} \int_{-h}^{0} \varphi_i(\theta) g_i(-\theta)\, d\theta, \quad \varphi \in X,\ g \in X^*. \tag{3.3}$$

In order to apply the general linear theory summarized in section 2, we take X as above and consider the strongly continuous semigroup T_0 defined by (1.7):

$$(T_0(t)\varphi)(\theta) := \begin{cases} \varphi(t+\theta) & \text{for} \quad t+\theta \in [-h, 0], \\ 0 & \text{for} \quad t+\theta > 0, \end{cases} \quad \varphi \in X,\, t \ge 0,\, \theta \in [-h, 0]. \tag{3.4}$$

Note that T_0 is nilpotent ($T_0(t) = 0$ for $t > h$). In particular, T_0 is eventually compact.

Remark 3.1. The explicit formula (3.4) makes it clear that equivalence classes are mapped to equivalence classes, such that $T_0(t)$ is indeed an operator mapping X into X. In line with common praxis, we will be sloppy when it comes to distinguishing elements of L^1, namely equivalence classes, from their representatives. It is, however, important to note that an equivalence class is by definition absolutely continuous if it contains an absolutely continuous function (that is, a function all the components of which are absolutely continuous). We shall always use this absolutely continuous function to represent an absolutely continuous equivalence class. As in the introduction we shall use the notation $\varphi \in AC$ to indicate that φ is absolutely continuous.

The following characterization of the generator of T_0 is well known, at least in the case of scalar-valued functions [2, p. 11]. As the vector-valued case is not more difficult, we present it without proof.

PROPOSITION 3.2. *The generator A_0 of T_0 is given by*

$$\begin{aligned} \mathcal{D}(A_0) &= \{\varphi \in X : \varphi \in AC,\ \varphi(0) = 0\}, \\ A_0\varphi &= \varphi'. \end{aligned}$$

Our next task is to characterize $X^{\odot *}$ and $T_0^{\odot *}$ and prove sun-reflexivity of X with respect to T_0 so that we can give a precise meaning to the abstract integral equation (AIE) for the specific application we are considering. This is a rather straightforward exercise. In the case of scalar-valued functions it is essentially carried out in [9], the only difference being the way in which the spaces X^*, $X^{\odot}$, $X^{\odot *}$ are represented. Because the smoothness and boundary conditions entering into the domains of definition of the generators are defined componentwise, the vector-valued case does not present any extra difficulties [27, Chapter 3]. We shall therefore give only a brief sketch of the construction of $X^{\odot *}$ and $T_0^{\odot *}$ and a precise formulation of the result that we need.

With the chosen representation of X^*, the adjoint semigroup T_0^* is translation to the left with extension by zero. Translation is clearly *not* continuous in L^∞ (to see this, just consider translation of any discontinuous function). The maximal subspace on which T_0^* is strongly continuous is $X^{\odot} = C_0\left([0,h);\mathbf{R}^N\right)$, the space of all continuous $\mathbf{R}^N$-valued functions vanishing at h. This last condition derives, of course, from the extension by zero of the translated function.

By the Riesz representation theorem, the dual $X^{\odot *}$ can be represented by $\mathrm{NBV}((-h,0];\ \mathbf{R}^N)$, the space of all $\mathbf{R}^N$-valued functions f, all the components of which are of bounded variation, are continuous from the right, and vanish at 0. Note, in particular, that $f \in \mathrm{NBV}\big((-h,0];\mathbf{R}^N\big)$ does *not* have a jump in $-h$ and that this is indicated by the half-open interval $(-h,0]$ of definition of f. The duality pairing between $X^{\odot} = C_0\big([0,h);\mathbf{R}^N\big)$ and $X^{\odot *} = \mathrm{NBV}\big((-h,0];\mathbf{R}^N\big)$ is given by the sum of Riemann–Stieltjes integrals

$$\langle g, f\rangle = \sum_{i=1}^{N} \int_{-h}^{0} g_i(-\theta) f_i(d\theta), \quad g \in X^{\odot},\ f \in X^{\odot *}, \tag{3.5}$$

and the norm on $X^{\odot *}$ by

$$\|f\|_{\mathrm{NBV}} := \sum_{i=1}^{N} \|f_i\|_{\mathrm{NBV}}, \tag{3.6}$$

where on the right-hand side $\|f_i\|_{\mathrm{NBV}}$ denotes the total variation of f_i.

The semigroup $T_0^{\odot *}$ is again translation to the left with extension by zero and it is not strongly continuous on $X^{\odot *}$. It is strongly continuous precisely on $X^{\odot\odot} = \{f \in X^{\odot *} : f \in AC\}$ [4, 9]. By the definition (2.5) of the canonical injection j :

$X \to X^{\odot *}$ and the definitions (3.3) and (3.5) of the pairings between our particular representations of X and $X^{\odot}$ and $X^{\odot}$ and $X^{\odot *}$, one obtains

$$\sum_{i=1}^{N} \int_{-h}^{0} g_i(-\theta)(j\varphi)_i(d\theta) = \langle g, j\varphi \rangle = \langle \varphi, g \rangle = \sum_{i=1}^{N} \int_{-h}^{0} \varphi_i(\theta) g_i(-\theta) d\theta,$$

from which it follows that $(j\varphi)' = \varphi$ or, equivalently,

$$j(\varphi)(\theta) = -\int_{\theta}^{0} \varphi(\tau) d\tau, \quad \theta \in (-h, 0]. \tag{3.7}$$

Now it is well known [52, Theorem 8.18] that a function of bounded variation is absolutely continuous if and only if it is the primitive of an L^1-function. Thus $j\,(X) = X^{\odot\odot}$; that is, X is sun-reflexive. We formulate the main conclusions as the following proposition.

PROPOSITION 3.3. *The space* $X = L^1\left([-h, 0]; \mathbf{R}^N\right)$ *is sun-reflexive with respect to the strongly continuous semigroup* T_0 *of bounded linear operators defined by* (3.4). *For* $\psi \in X^{\odot *} = \mathrm{NBV}((-h, 0]; \mathbf{R}^N)$, $t \geq 0$, *and* $\theta \in [-h, 0]$ *one has*

$$(T_0^{\odot *}(t)\psi)(\theta) = \begin{cases} \psi(t+\theta) & \text{for} \quad t + \theta \in [-h, 0), \\ 0 & \text{for} \quad t + \theta \geq 0. \end{cases} \tag{3.8}$$

The generator $A_0^{\odot *}$ *of* $T_0^{\odot *}$ *is given by*

$$\begin{aligned} \mathcal{D}(A_0^{\odot *}) = \{\varphi \in X^{\odot *} :\ & \varphi(\theta) = \int_{\theta}^{0} \psi(\alpha) d\alpha \text{ for all } \theta \in [-h, 0] \\ & \text{and some } \psi \in X^{\odot *}\}, \end{aligned} \tag{3.9}$$

$$A_0^{\odot *}\varphi = -\psi \tag{3.10}$$

or, in shorthand notation, $A_0^{\odot *}\varphi = \varphi'$.

As a corollary to Proposition 3.3 we get a formula for the resolvent of $A_0^{\odot *}$ which we state for later use.

COROLLARY 3.4. *For* $f \in X^{\odot *} = \mathrm{NBV}((-h, 0]; \mathbf{R}^N)$ *and* $\lambda \in \mathbf{C}$ *we have*

$$\left(j^{-1} R\left(\lambda, A_0^{\odot *}\right) f\right)(\theta) = \int_{\theta}^{0} e^{\lambda(\theta - \tau)} f(d\tau), \quad \theta \in [-h, 0].$$

Proof. By definition, $R\left(\lambda, A_0^{\odot *}\right) f$ is the unique element $\varphi \in \mathcal{D}(A_0^{\odot *})$ which satisfies the equation

$$(\lambda E - A_0^{\odot *})\varphi = f. \tag{3.11}$$

By Proposition 3.3 there exists a $\psi \in X^{\odot *}$ such that

$$\varphi(\theta) = \int_{\theta}^{0} \psi(\alpha) d\alpha, \quad \theta \in [-h, 0], \tag{3.12}$$

and

$$A_0^{\odot *}\varphi = -\psi. \tag{3.13}$$

Equation (3.11) therefore becomes

$$\lambda \int_\theta^0 \psi(\alpha)d\alpha + \psi(\theta) = f(\theta), \quad \theta \in [-h, 0], \tag{3.14}$$

which has the unique solution

$$\psi(\theta) = -\int_\theta^0 e^{\lambda(\theta-\tau)} f(d\tau), \quad \theta \in [-h, 0]. \tag{3.15}$$

The inverse of the canonical injection j defined by (3.7) is clearly differentiation. Therefore

$$j^{-1} R\left(\lambda, A_0^{\odot *}\right) f = j^{-1}\varphi = \varphi' = -\psi. \tag{3.16}$$

The assertion now follows from (3.15) and (3.16). ◻

3.2. The perturbed problem. In this subsection we show that with a specific choice of perturbation $G : X \to X^{\odot *}$, the perturbed problem, which, as we have shown in section 2, amounts to the abstract integral equation (AIE), is equivalent to the originally given delay equation (DE) and initial condition (IC). To this end, we let $F : X \to \mathbf{R}^N$ be a nonlinear mapping and define $G : X \to X^{\odot *}$ by

$$G(\varphi) = \sum_{i=1}^N F_i(\varphi) H_i, \tag{3.17}$$

where F_i denotes the ith component of F for $i = 1, \dots, N$ and H_i is defined by

$$H_i(\theta) := \begin{cases} e_i & \text{for} \quad \theta \in (-h, 0), \\ 0 & \text{for} \quad \theta = 0. \end{cases} \tag{3.18}$$

Here and in what follows $\{e_1, e_2, \dots, e_N\}$ is the standard basis of $\mathbf{R}^N$. Notice that G has finite dimensional range spanned by $\{H_1, H_2, \dots, H_N\}$ in $X^{\odot *}$.

Next we compute the weak* integral in (AIE) when G is defined through (3.17) and (3.18).

LEMMA 3.5. *Let T_0 be the strongly continuous semigroup defined by* (3.4). *Then for every $\eta \in L^1_{\mathrm{loc}}(\mathbf{R}_+)$ and $t \geq 0$ one has*

$$\left(\int_0^t T_0^{\odot *}(t-\tau)\eta(\tau) H_i d\tau\right)(\theta) = -e_i \int_{t+\max\{-t,\theta\}}^t \eta(\sigma)d\sigma, \quad \theta \in (-h, 0].$$

Proof. First notice that for $0 \leq s < h$ one has

$$\left(T_0^{\odot *}(s) H_i\right)(\theta) = \begin{cases} -e_i & \text{for} \quad -h \leq \theta < -s, \\ 0 & \text{for} \quad -s \leq \theta \leq 0. \end{cases} \tag{3.19}$$

The NBV function $T_0^{\odot *}(s) H_i$ thus has a unit jump at $\theta = -s$, and hence

$$\left\langle T_0^{\odot *}(t) H_i, g\right\rangle = \int_{-h}^0 g(-\theta)\left(T_0^{\odot *}(s) H_i\right)(d\theta) = g_i(s) \tag{3.20}$$

for any continuous g. It follows that for $0 \le t \le h$

$$\left\langle \int_0^t T_0^{\odot *}(t-s)\eta(s)H_i ds, g \right\rangle = \int_0^t \eta(s)g_i(t-s)ds$$

$$= \int_{-t}^0 \eta(t+s)g_i(-s)ds = \langle y, g\rangle,$$

where y is the absolutely continuous NBV function defined by

$$y(\theta) = \begin{cases} -\int_{t+\theta}^t \eta(s)ds & \text{for} \quad \theta \le 0 \le 0, \\ -\int_0^t \eta(s)ds & \text{for} \quad -h \le \theta < -t, \end{cases} \tag{3.21}$$

and the conclusion follows. □

Applying this result to $\eta(t) = F_i(u(t))$, we get the following corollary.

COROLLARY 3.6. *Let T_0 be the strongly continuous semigroup defined by* (3.4) *and let $G : X \to X^{\odot *}$ be defined by* (3.17) *and* (3.18). *If $u : [0, t) \to X$ is continuous, then*

$$\left(\int_0^t T_0^{\odot *}(t-s)G(u(s))ds\right)(\theta) = -\int_{t+\max\{-t,\theta\}}^t F(u(s))ds \tag{3.22}$$

for all $\theta \in [-h, 0]$.

Proof. Using Lemma 3.5, one computes

$$\left(\int_0^t T_0^{\odot *}(t-s)G(u(s))ds\right)(\theta) = \sum_{i=1}^N \left(\int_0^t T_0^{\odot *}(t-s)F_i(u(s))H_i ds\right)(\theta)$$

$$= -\sum_{i=1}^N e_i \int_{t+\max\{-t,\theta\}}^t F_i(u(s))ds = -\int_{t+\max\{-t,\theta\}}^t F(u(s))ds. \qquad \square$$

We are now ready to prove equivalence of solutions of the abstract integral equation

$$u(t) = T_0(t)\varphi + j^{-1}\left(\int_0^t T_0^{\odot *}(t-s)G(u(s))ds\right) \tag{AIE}$$

and the delay problem

$$x(t) = F(x_t), \quad t > 0, \tag{DE}$$

$$x_0(\theta) = \varphi(\theta), \quad \theta \in [-h, 0]. \tag{IC}$$

For ease of formulation we consider global solutions, i.e., solutions defined for all future times. It should, however, be evident that one can formulate and prove an analogous result concerning local solutions.

THEOREM 3.7. *Let $\varphi \in X = L^1\left([-h, 0]; \mathbf{R}^N\right)$ be given.*

(a) *Suppose that $x \in L^1_{\text{loc}}\left([-h, \infty); \mathbf{R}^N\right)$ satisfies* (DE) *and* (IC). *Then the function $u : [0, \infty) \to X$ defined by $u(t) := x_t$ is continuous and satisfies* (AIE).

(b) *Suppose that there is a continuous map* $u : [0,\infty) \to X$ *that satisfies* (AIE). *Then the function* x *defined as*

$$x(t) := \begin{cases} \varphi(t) & \text{for} \quad t \in [-h, 0), \\ u(t)(0) & \text{for} \quad t \geq 0 \end{cases} \tag{3.23}$$

is an element of $L^1_{\text{loc}}\left([-h,\infty);\mathbf{R}^N\right)$ *and satisfies* (DE) *and* (IC).

Proof. (a) First, note that the continuity assertion follows from the fact that translation is continuous in L^1. Then, by (DE) and (IC) we get

$$u(t)(\theta) - (T_0(t)\varphi)(\theta) = \begin{cases} 0 & \text{for} \quad t+\theta \in [-h,0), \\ x(t+\theta) & \text{for} \quad t+\theta \geq 0 \end{cases} \tag{3.24}$$

$$= \begin{cases} 0 \\ F(x_{t+\theta}) \end{cases} = \begin{cases} 0 & \text{for} \quad t+\theta \in [-h,0), \\ F(u(t+\theta)) & \text{for} \quad t+\theta \geq 0. \end{cases}$$

On the other hand, by Corollary 3.6 one gets

$$j^{-1}\left(\int_0^t T_0^{\odot *}(t-s)G(u(s))ds\right)(\theta) = \begin{cases} 0 & \text{for} \quad t+\theta \in [-h,0), \\ F(u(t+\theta)) & \text{for} \quad t+\theta \geq 0, \end{cases}$$

which equals (3.24), and therefore (AIE) holds.

(b) Suppose now that u satisfies (AIE). Then by Corollary 3.6 for $t \geq 0$ one has

$$\begin{aligned} x(t) = u(t)(0) &= j^{-1}\left(\int_0^t T_0^{\odot *}(t-s)G(u(s))ds\right)(0) \\ &= -\frac{d}{da}\int_{t+\max\{-t,a\}}^t F(u(s))ds|_{a=0} \qquad (3.25) \\ &= F(u(t)). \end{aligned}$$

Hence it remains to be shown that $u(t) = x_t$. Using (AIE), Corollary 3.6, and (3.25), one computes for $\theta \in [-h, 0]$ that

$$u(t)(\theta) = \begin{cases} \varphi(t+\theta) & \text{for} \quad t+\theta \in [-h,0), \\ j^{-1}\left(\int_0^t T_0^{\odot *}(t-s)G(u(s))ds\right)(\theta) & \text{for} \quad t+\theta \geq 0 \end{cases}$$

$$= \begin{cases} \varphi(t+\theta) & \text{for} \quad t+\theta \in [-h,0), \\ j^{-1}\left(\int_{t+\max\{-t,\cdot\}}^t F(u(s))ds\right)(\theta) & \text{for} \quad t+\theta \geq 0 \end{cases}$$

$$= \begin{cases} \varphi(t+\theta) & \text{for} \quad t+\theta \in [-h,0), \\ F(u(t+\theta)) & \text{for} \quad t+\theta \geq 0. \end{cases}$$

Thus one has $u(t) = x_t$, and (b) is also proved. ∎

As is clear from the results of section 2, the abstract integral equation approach is ideal for deriving results concerning the qualitative behavior of solutions, such as stability and bifurcation. On the other hand, for proving regularity of solutions it is usually easier to attack the problem (DE) and (IC) directly. This is shown in the proof of the next theorem (which is not the sharpest possible result; indeed, the conclusion holds even if F is only locally Lipschitz, but then the proof is a bit more technical).

One of the advantages of the equivalence result of Theorem 3.7 is that we can freely choose between the abstract and the concrete, according to our needs.

THEOREM 3.8. *Let $F : L^1\left([-h,0];\mathbf{R}^N\right) \to \mathbf{R}^N$ be globally Lipschitz continuous. Then the unique solution $x : [-h,\infty) \to \mathbf{R}^N$ of* (DE), (IC) *with $\varphi \in L^1\left([-h,0];\mathbf{R}^N\right)$ is continuous in $[0,\infty)$.*

Proof. Let $\varphi \in L^1\left([-h,0];\mathbf{R}^N\right)$ and $\ell > 0$. Define

$$Z = \left\{ y \in C\left([0,\ell];\mathbf{R}^N\right) : y(0) = F(\varphi) \right\}.$$

Then Z is a closed subset of the Banach space $C\left([0,\ell];\mathbf{R}^N\right)$ and thus a complete metric space. Define for each $y \in Z$ the function $\Phi(y)$ on $[0,\ell]$ by

$$\left(\Phi y\right)(t) = F(z_t^y), \quad 0 \le t \le \ell,$$

where z^y is the function defined by

$$z^y(\tau) = \begin{cases} \varphi(\tau) & \text{for} \quad -h \le \tau < 0, \\ y(\tau) & \text{for} \quad \le \tau \le \ell \end{cases}$$

and z_t^y is the translate of z^y as in (1.2). Clearly z^y belongs to L^1. Because translation is continuous when regarded as a mapping from an interval to L^1 and F is continuous on L^1, it follows that Φy is continuous. Moreover, $\left(\Phi y\right)(0) = F(z_0^y) = F(\varphi)$, and hence Φy belongs to Z. Next we show that Φ is a contraction on Z for ℓ sufficiently small. Because F is globally Lipschitz continuous we have for y_1, $y_2 \in Z$

$$\begin{aligned} \left|\left(\Phi y_1\right)(t) - \left(\Phi y_2\right)(t)\right| &= \left|F\left(z_t^{y_1}\right) - F\left(z_t^{y_2}\right)\right| \\ &\le L \left\|z_t^{y_1} - z_t^{y_2}\right\|_{L^1([-h,0];\mathbf{R}^N)} \\ &\le L \int_0^\ell |y_1(\tau) - y_2(\tau)| \, d\tau. \end{aligned}$$

Hence Φ has, for ℓ sufficiently small, a unique fixed point. The fixed point is obviously a solution of (DE) and (IC). This proves the assertion. ∎

The present way to associate a dynamical system with a Volterra integral equation is dual to the way studied in [21], where, of course, "dual" is precisely defined only in the linear case. The advantage of the present approach is that we also cover autonomous nonlinear problems that are *not* of convolution type, while [21] is restricted to convolution equations (see subsection 3.5 below).

3.3. Steady states. In this subsection we characterize the steady states of the nonlinear semigroup Σ generated by the abstract integral equation (AIE) in terms of constant solutions of (DE) and (IC).

THEOREM 3.9. (a) *Suppose $\overline{\varphi}$ is a steady state of Σ. Then $\overline{\varphi}$ is a constant function*

$$\overline{\varphi}(\theta) = \overline{x}, \quad \theta \in [-h,0], \tag{3.26}$$

and

$$\overline{x} = F\left(\overline{\varphi}\right). \tag{3.27}$$

(b) *Conversely, if the constant function $\overline{\varphi}$ given by* (3.26) *satisfies* (3.27)*, then it is a steady state of Σ.*

Proof. (a) Let $\overline{\varphi}$ be a steady state of Σ, i.e., $\Sigma(t)\overline{\varphi} = \overline{\varphi}$ for all $t \geq 0$. From (AIE) we then get

$$\overline{\varphi} = T_0(t)\overline{\varphi} + j^{-1}\left(\int_0^t T_0^{\odot*}(t-\tau)G(\overline{\varphi})d\tau\right), \quad t \geq 0.$$

Because $T_0(t) = 0$ for $t > h$, it follows that

$$j\overline{\varphi} = \int_0^t T_0^{\odot*}(t-\tau)G(\overline{\varphi})d\tau = \sum_{i=1}^N \int_0^t T_0^{\odot*}(\tau)H_iF_i(\overline{\varphi})d\tau.$$

Using Lemma 3.5 we then deduce that for $t > h$

$$(j\overline{\varphi}) = \sum_{i=1}^N -e_i \int_{t+\theta}^t F_i(\overline{\varphi})d\theta = \sum_{i=1}^N e_i\theta F_i(\overline{\varphi}).$$

But because j is integration, this means precisely that

$$\overline{\varphi}(\theta) = \sum_{i=1}^N e_iF_i(\overline{\varphi}) = F(\overline{\varphi}), \quad \theta \in [-h, 0];$$

that is, $\overline{\varphi}$ is a constant function and (3.27) holds.

The proof of (b) is similar. ∎

From the equivalence of (AIE) and (DE), (IC) (Theorem 3.7) it is clear that a function $\overline{\varphi}$ that takes the constant value $\overline{x} \in \mathbf{R}^N$ on $[-h, 0]$ is a steady state of Σ if and only if the constant function $x(t) = \overline{x},\ t \in [-h, \infty)$ is the solution of (DE), (IC).

Remark 3.10. In what follows we shall abuse notation and denote both the constant function $\overline{\varphi}$ on $[-h, 0]$ and the corresponding constant function on $[-h, \infty)$ by the same symbol as the constant value they take, viz. $\overline{x}$.

Because the constant solutions of (DE), (IC) are steady states of the dynamical system Σ, we have well-defined notions of stability at our disposal. It follows immediately from Theorem 3.8 that the constant solution $\overline{x}$ of (DE), (IC) is (locally) stable if and only if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\int_{-h}^0 |x(t) - \overline{x}|dt \leq \delta \quad \Rightarrow \quad |x(t) - \overline{x}| \leq \varepsilon \text{ for all } t > 0$$

and (locally) *exponentially stable* if there exist numbers $\delta > 0,\ K > 0,\ \alpha > 0$ such that

$$\int_{-h}^0 |x(t) - \overline{x}|dt \leq \delta \quad \Rightarrow \quad |x(t) - \overline{x}| \leq Ke^{-\alpha t} \text{ for all } t > 0.$$

3.4. The characteristic equation. In section 2.5 we showed that whenever $\sigma(A_0)$ is empty (in particular, when T_0 is nilpotent) and the perturbation has finite dimensional range, the spectrum $\sigma(A)$ of the perturbed generator consists entirely of eigenvalues and there exists a characteristic equation

$$\det(E - M(\lambda)) = 0,$$

the roots of which are exactly the eigenvalues. The characteristic equation contains all the information about asymptotic behavior, Hopf bifurcation, etc. In this section we identify the matrix $M(\lambda)$ for the special case in which the unperturbed semigroup T_0 is given by (3.4) and the perturbation G is of the form (3.17).

If G is differentiable at $\overline{\varphi}$, there exist functions $k_{ij} \in L^\infty([0,h];\mathbf{R})$ such that

$$G'(\overline{\varphi})\varphi = \sum_{i=1}^{N}\left(\sum_{j=1}^{N}\int_{-h}^{0} k_{ij}(-\theta)\varphi_j(\theta)d\theta\right)H_i.$$

$B = G'(\overline{\varphi})$ is thus of the form (2.15), with $r_i^{\odot *} = H_i$ and $r_i^* = k_i = \{k_{ij}\}_{j=1}^N$. Corollary 3.4 now yields

$$\begin{aligned}(r_j(\lambda))(\theta) &= \left(j^{-1}R(\lambda, A_0^{\odot *})H_j\right)(\theta) \\ &= \int_\theta^0 e^{\lambda(\theta-\tau)}H_j(d\tau) = e^{\lambda\theta}e_j, \quad \theta \in [-h,0],\end{aligned} \tag{3.28}$$

and hence

$$M_{ij}(\lambda) = \langle r_j(\lambda), k_i\rangle = \int_{-h}^{0} e^{\lambda\theta}k_{ij}(-\theta)d\theta = \int_0^h e^{-\lambda\theta}k_{ij}(\theta)d\theta = \widehat{k}_{ij}(\lambda).$$

Denoting the matrix with entries k_{ij} by k, the characteristic equation thus takes the form

$$\det\left(E - \widehat{k}(\lambda)\right) = 0. \tag{3.29}$$

The results of subsection 2.6 now tell us that if all the roots of the characteristic equation (3.29) have negative real part, then the steady state is exponentially stable, whereas it is unstable if at least one root has positive real part. Note that the hypotheses of Corollary 2.20 are fulfilled, so Nyquist's criterion for (in)stability is applicable. It is also a straightforward fill-in exercise to translate Theorem 2.21 into a result for delay equations (generalizing Theorem 11.1 in [21] to include equations which are not of convolution type, and being the analogue of Theorem X.2.7 in [23], which applies to delay differential equations).

3.5. Differentiability for three important classes of nonlinearity. In order to apply the general results on stability and bifurcation to the system (DE), (IC) we have to give conditions that ensure that the map $G : X \to X^{\odot *}$ is Fréchet differentiable with $X = L^1\left([-h,0];\mathbf{R}^N\right)$ and G of finite dimensional range given by (3.17) or, more generally, by (2.14). As noted in subsection 2.5, G is differentiable if and only if F is differentiable from $L^1\left([-h,0];\mathbf{R}^N\right)$ to $\mathbf{R}^N$. This leads us to have a closer look at differentiability criteria for functions from $L^1\left([-h,0];\mathbf{R}^N\right)$ to $\mathbf{R}^N$.

There is one obvious class of differentiable mappings from $L^1\left([-h,0];\mathbf{R}^N\right)$ to $\mathbf{R}^N$, consisting of those mappings of the form $\varphi \mapsto (g\circ\Lambda)\varphi$, where $\Lambda : L^1\left([-h,0];\mathbf{R}^N\right) \to \mathbf{R}^N$ is a bounded linear map and g is a smooth function from $\mathbf{R}^N$ to $\mathbf{R}^N$.

A map F that occurs frequently in applications is $F = \Lambda \circ N_g$, where Λ is a bounded linear map from $L^1\left([-h,0];\mathbf{R}^N\right)$ to $\mathbf{R}^N$ and N_g is the Nemytskiĭ operator induced by a smooth function $g : \mathbf{R}^N \to \mathbf{R}^N$ as follows:

$$(N_g(\varphi))(\theta) = g(\varphi(\theta)). \tag{3.30}$$

For instance, the nonlinear Volterra convolution equation

$$x(t) = \int_0^h k(s)g(x(t-s))ds, \quad t > 0, \tag{3.31}$$

$$x(t) = \varphi(t), \quad -h \le t \le 0, \tag{3.32}$$

is of the form (DE), (IC) with $F = \Lambda \circ N_g$ and $\Lambda\varphi = \int_0^h k(\theta)\varphi(-\theta)d\theta$.

It may come as a surprise that the Nemytskiĭ operator from $L^1\left([-h,0];\mathbf{R}^N\right)$ to $L^1\left([-h,0];\mathbf{R}^N\right)$ generated by a differentiable, globally Lipschitz continuous function $g:\mathbf{R}^N \to \mathbf{R}^N$ is *not* Fréchet differentiable unless g is affine (a constant plus a linear operator), but in hindsight this is easy to understand. An indication of the reason is that in the formal Taylor series expansion

$$\begin{aligned}(N_g(\varphi))(\theta) &= g(\varphi(\theta)) \\ &= g(\overline{\varphi}(\theta)) + g'(\overline{\varphi}(\theta))(\overline{\varphi}(\theta)-\varphi(\theta)) + \frac{1}{2}g''(\overline{\varphi}(\theta))(\overline{\varphi}(\theta)-\varphi(\theta))^2 + \cdots\end{aligned} \tag{3.33}$$

around an element $\overline{\varphi} \in L^1$, the higher order terms contain powers of φ which need not belong to L^1. So showing that the higher order terms are small cannot be done in the standard way (and, in fact, cannot be done at all).

The above result may seem disastrous for our theory because it appears as if the important case of the nonlinear Volterra convolution equation (3.31) would not be covered by it. Fortunately, a simple transformation saves our bacon.

Consider the Volterra functional equation

$$x(t) = \Lambda N_g(x_t) \tag{3.34}$$

with initial condition

$$x_0 = \varphi. \tag{3.35}$$

Applying the function g to both sides of (3.34) and (3.35), one obtains

$$(g\circ x)(t) = g\left(\Lambda N_g(x_t)\right) \tag{3.36}$$

and

$$g(x_0) = g(\varphi). \tag{3.37}$$

But

$$(g\circ x)_t(\theta) = g(x(t+\theta)) = g(x_t(\theta)) = (N_g(x_t))(\theta),$$

that is,

$$(g\circ x)_t = N_g(x_t), \tag{3.38}$$

and hence (3.36) and (3.37) take the form

$$y(t) = (g\circ\Lambda)\left(y_t\right), \tag{3.39}$$

$$y_0 = \psi \tag{3.40}$$

with

$$y(t) = (g \circ x)(t), \tag{3.41}$$

$$\psi = g \circ \varphi. \tag{3.42}$$

But $g \circ \Lambda$ is differentiable, and thus our theory applies to the transformed problem (3.39), (3.40): A constant solution $\overline{y}$ of (3.39), (3.40) is exponentially stable if all the roots λ of the characteristic equation

$$\det\left(E - g'(\overline{x})\widehat{k}(\lambda)\right) = 0 \tag{3.43}$$

satisfy $\mathrm{Re}\lambda < 0$ and unstable if there exists at least one root with positive real part.

We recover the solution x of our original problem (3.34), (3.35), because (3.34), (3.41), and (3.38) together show that

$$x(t) = \Lambda y_t. \tag{3.44}$$

It remains to be shown that the stability properties of the transformed problem determine those of the original problem. For this the differentiability of g is irrelevant; we assume only global Lipschitz continuity as this guarantees that the Nemytskiĭ operator N_g maps L^1 into L^1.

THEOREM 3.11. *Let $\Lambda : L^1\left([-h,0];\mathbf{R}^N\right) \to \mathbf{R}^N$ be a bounded linear operator and let $g : \mathbf{R}^N \to \mathbf{R}^N$ be globally Lipschitz continuous with Lipschitz constant L. Let $\overline{x}$ be a constant solution of* (3.34), (3.35) *and let $\overline{y} = g(\overline{x})$ be the corresponding constant solution of* (3.39), (3.40). *Then the following hold:*

(a) *If $\overline{y}$ is [exponentially] stable, then so is $\overline{x}$.*

(b) *If $\overline{y}$ is unstable, then so is $\overline{x}$.*

Proof. The estimate

$$|x(t) - \overline{x}| = |\Lambda y_t - \Lambda g(\overline{x})| \leq \|\Lambda\| \|y_t - \overline{y}\|_1 \tag{3.45}$$

proves (a). Assume now that $\overline{x}$ is stable and let y be the solution of (3.39), (3.40). Define

$$x(t) = \Lambda y_t, \quad t \geq 0. \tag{3.46}$$

We know by Theorem 3.8 that y is continuous for $t \geq 0$; it follows that $x_h \in L^1\left([-h,0];\mathbf{R}^N\right)$. So, for $t \geq h$, y may be regarded as the solution of (3.39) with the initial condition (3.40) replaced by

$$y_h = g \circ x_h. \tag{3.47}$$

Because the mapping that takes ψ to y_t is a strongly continuous (nonlinear) semigroup, $\sup_{0 \leq t \leq h} \|y_t - \overline{y}\|_1$ can be made arbitrarily small by choosing $\|\psi - \overline{y}\|_1$ sufficiently small. Now (3.45) shows that $x_h - \overline{x}$ also can be made arbitrarily small.

Let $\varepsilon > 0$ be arbitrary. Because $\overline{x}$ is stable one can choose $\delta > 0$ such that $\|x_h - \overline{x}\| < \delta$ implies $|x(t) - \overline{x}| < \varepsilon/L$ for all $t > h$. It follows that

$$|y(t) - \overline{y}| = |g(x(t)) - g(\overline{x})| \leq L|x(t) - \overline{x}| < \varepsilon$$

for all $t > 0$ provided that $\|\psi - \overline{y}\|_1$ is sufficiently small, that is, $\overline{y}$ is stable. ∎

Note that if we linearize the nonlinear Volterra integral equation (3.31) in $\mathbf{R}^N$ (as opposed to linearizing $F : L^1 \to \mathbf{R}^N$ in the delay equation), we obtain

$$x(t) = \int_0^t k(s)g'(\overline{x})x(t-s)ds + G(x)(t), \tag{3.48}$$

where $G(x)(t)$ stands for the higher order terms. In the theory of Volterra integral equations [29] one associates the characteristic equation

$$\det\left(E - \widehat{k}(\lambda)g'(\overline{x})\right) = 0 \tag{3.49}$$

with (3.48). Clearly, (3.43) and (3.49) have exactly the same roots.

A third class of differentiable maps from $L^1\left([-h,0];\mathbf{R}^N\right)$ to $\mathbf{R}^N$ is obtained by composing a map from the space $C\left([0,\ell];\mathbf{R}^N\right)$ of continuous functions on some interval $[0,\ell]$ to $\mathbf{R}^N$ with a linear (or affine) map from $L^1\left([-h,0];\mathbf{R}^N\right)$ to $C\left([0,\ell];\mathbf{R}^N\right)$. The reason for this detour via $C\left([0,\ell];\mathbf{R}^N\right)$ is that, roughly speaking, it is much easier for a function to be differentiable if it is defined on $C\left([0,\ell];\mathbf{R}^N\right)$ than if it is defined on L^1. Indeed, the product of two continuous functions which are small in the supremum norm is continuous, and the supremum norm of the product is of quadratic order. In particular, the expansion (3.33) applied to a continuous function φ shows that the Nemytskiĭ operator is differentiable in $C\left([0,\ell];\mathbf{R}^N\right)$ and that

$$N_g'(\overline{\varphi}) = N_{g'}(\overline{\varphi}). \tag{3.50}$$

This observation is important in applications to, for instance, population dynamics. Let us illustrate it by an age-structured model of the type first studied in [30]. Assume that the age-specific per capita death rate depends on the present value $I(t)$ of the (one-dimensional) environmental condition in the following way:

$$\mu(a, I(t)) = \mu_0(a) + \mu_1(a)I(t) \tag{3.51}$$

(where μ_0 and μ_1 are nonnegative functions). Then the probability $\mathcal{F}(a;\varphi)$ that an individual that was born a time units ago is still alive, given the history φ of the environmental condition, is the solution of the ODE initial value problem

$$\frac{d}{d\alpha}\mathcal{F}(\alpha;\varphi) = -\mu(\alpha, \varphi(\alpha - a))\mathcal{F}(\alpha,\varphi), \tag{3.52}$$

$$\mathcal{F}(0) = 1 \tag{3.53}$$

at $\alpha = a$, that is,

$$\mathcal{F}(a;\varphi) = \exp\left(-\int_0^a \left(\mu_0(\alpha) + \mu_1(\alpha)\varphi(\alpha - a)\right)d\alpha\right). \tag{3.54}$$

The theory presented in this paper presupposes a maximum life span h. This is achieved by assuming that μ_0 has a nonintegrable singularity at h:

$$\int_0^h \mu_0(a)da = \infty,$$

because then the survival probability

$$\mathcal{F}_0(a) = \exp\left(-\int_0^a \mu_0(\alpha)d\alpha\right) \tag{3.55}$$

with respect to density-independent effects vanishes at h.

If $\beta(a, I(t))$ is the age-specific fecundity, then the integral equations (1.18), (1.19) combined with the feedback law (1.17) yield

$$b(t) = \int_0^h \beta(a, I(t))\mathcal{F}(a; I_t)b(t-a)da, \tag{3.56}$$

$$I(t) = \int_0^h \gamma(a)\mathcal{F}(a; I_t)b(t-a)da, \tag{3.57}$$

which is a delay equation of the type (DE). More specifically, we have

$$\begin{pmatrix} b(t) \\ I(t) \end{pmatrix} = F\begin{pmatrix} b_t \\ I_t \end{pmatrix} \tag{3.58}$$

with F given by

$$F\begin{pmatrix} \psi \\ \varphi \end{pmatrix} = \begin{pmatrix} \int_0^h \beta\left(a, \int_0^h \gamma(\alpha)\mathcal{F}(\alpha;\varphi)\psi(-\alpha)d\alpha\right)\mathcal{F}(a;\varphi)\psi(-a)da \\ \int_0^h \gamma(a)\mathcal{F}(a;\varphi)\psi(-a)da \end{pmatrix}. \tag{3.59}$$

F is a well-defined mapping from $L^1\left([-h,0];\mathbf{R}^2\right)$ to $\mathbf{R}^2$ if $\gamma\mathcal{F}_0 \in L^\infty[0,h]$ and $\beta(\cdot,\overline{I})\mathcal{F}_0 \in L^\infty[0,h]$ for all $\overline{I} \in \mathbf{R}$, and hence we make this assumption.

We want to show that F is differentiable. First notice that the argument of the exponential function in formula (3.54) is an affine map taking $\varphi \in L^1[-h,0]$ to $C[0,h]$. The mapping $\varphi \mapsto \mathcal{F}(\cdot;\varphi)$ is thus obtained by composing the Nemytskiĭ operator induced in $C[0,h]$ by the exponential function with an affine map. As we already saw, this map is Fréchet differentiable. For fixed ψ, the second component F_2 of F is now obtained by applying a continuous linear mapping to the differentiable map $\varphi \mapsto \mathcal{F}(\cdot;\varphi)$. Hence F_2 is differentiable in φ. Because F_2 is linear in ψ it is also differentiable in ψ. Because $F_2(\psi,\varphi)$ appears as the second argument of β in the expression for F_1, the chain rule implies that F_2 also is differentiable provided that $\beta : \mathbf{R}^2 \to \mathbf{R}$ is differentiable in its second argument.

The derivative of F can be computed explicitly. A straightforward but tedious computation yields F' at a steady state $(\overline{b},\overline{I})$:

$$\left[F'\begin{pmatrix} \overline{b} \\ \overline{I} \end{pmatrix}\right]\begin{pmatrix} \psi \\ \varphi \end{pmatrix} = \int_0^h k(a)\begin{pmatrix} \psi \\ \varphi \end{pmatrix}(-a)da, \tag{3.60}$$

where k is a 2×2 matrix-valued function with entries

$$k_{11}(a) = \left(\gamma(a)\int_0^h \partial_2\beta(\tau,\overline{I})\mathcal{F}(\tau;\overline{I})d\tau\,\overline{b} + \beta(a,\overline{I})\right)\mathcal{F}(a;\overline{I}), \tag{3.61}$$

$$\begin{aligned} k_{12}(a) = {} & \int_0^h \partial_2\beta(\tau,\overline{I})\mathcal{F}(\tau;\overline{I})d\tau\,\overline{b}\,k_{22}(a) \\ & - \int_0^{h-a} \mu_1(\alpha)\beta(\alpha+a,\overline{I})\mathcal{F}(\alpha+a;\overline{I})d\alpha\,\overline{b}, \end{aligned} \tag{3.62}$$

$$k_{21}(a) = \gamma(a)\mathcal{F}(a;\overline{I}), \tag{3.63}$$

$$k_{22}(a) = -\int_0^{h-a} \mu_1(\alpha)\gamma(\alpha+a)\mathcal{F}(\alpha+a;\overline{I})d\alpha\,\overline{b}. \tag{3.64}$$

The characteristic equation is (3.29) with the matrix k defined by (3.61)–(3.64). It is easy to check that the resulting stability criterion is equivalent (as it should be) to the one given in [30] for $\gamma \equiv 1$ and in [32] and [50] for the general case.

The steady environmental condition $\overline{I}$ of a nontrivial equilibrium $(\bar{b}, \overline{I}) \neq (0,0)$ is a solution (there may be many) of the steady state condition

$$1 = \int_0^h \beta(a, \overline{I})\mathcal{F}(a; \overline{I})da. \tag{3.65}$$

Once $\overline{I}$ has been solved from (3.65), the corresponding steady birth rate is obtained from

$$\bar{b} = \frac{\overline{I}}{\int_0^h \gamma(a)\mathcal{F}(a; \overline{I})da}. \tag{3.66}$$

On the other hand, for the population-free, or trivial, steady state $(\bar{b}, \overline{I}) = (0,0)$, the characteristic equation (3.29) reduces to the scalar equation

$$1 = \int_0^h e^{-\lambda a}\beta(a, 0)\mathcal{F}_0(a)da. \tag{3.67}$$

As a consequence, the population-free steady state is exponentially stable if

$$R_0 := \int_0^h \beta(a, 0)\mathcal{F}_0(a)da < 1$$

and unstable if

$$R_0 > 1.$$

In subsection 5.1 we shall elaborate on this a bit more in the context of a model for an age-structured population with cannibalistic behavior.

4. Volterra functional equations coupled with delay differential equations. In applications to structured population dynamics, one encounters models that take the form of a Volterra functional equation coupled with a delay differential equation [31, 32]. In this section we therefore briefly consider systems of the following type:

$$x(t) = F_1(x_t, y_t), \tag{4.1}$$

$$\dot{y}(t) = F_2(x_t, y_t). \tag{4.2}$$

For the component x of the delay equation (4.1), we choose as before $X = L^1\left([-h, 0]; \mathbf{R}^N\right)$ as state space, whereas the natural state space for the component y of the delay differential equation (4.2) is $Y = C\left([-h, 0]; \mathbf{R}^M\right)$ (see [23]). We therefore have to assume that the mappings $F_1 : X \times Y \to \mathbf{R}^N$ and $F_2 : X \times Y \to \mathbf{R}^M$ are at least Lipschitz continuous. Equations (4.1) and (4.2) must, of course, be supplemented by initial conditions

$$x(\theta) = \varphi(\theta), \quad -h \leq \theta \leq 0, \tag{4.3}$$

$$y(\theta) = \psi(\theta), \quad -h \leq \theta \leq 0. \tag{4.4}$$

In section 3 we showed in detail how a Volterra functional equation could be written as a semilinear abstract integral equation. The same program has been carried out for delay differential equations in the book [23] (see also [40]). It is now an easy exercise to combine the two procedures for the coupled system (4.1)–(4.4).

Let T_{10} be the C_0-semigroup defined on X by (3.4) and define the C_0-semigroup T_{20} on Y by

$$(4.5)\quad (T_{20}(t)\psi)(\theta) := \begin{cases} \psi(t+\theta) & \text{for} \quad t+\theta \in [-h,0], \\ \psi(0) & \text{for} \quad t+\theta \geq 0, \end{cases} \quad \psi \in Y,\, t \geq 0,\, \theta \in [-h,0].$$

The two semigroups T_{10} and T_{20} induce in an obvious way a semigroup T_0 on $X \times Y$:

$$(4.6)\qquad T_0(t) = \begin{pmatrix} T_{10}(t) & 0 \\ 0 & T_{20}(t) \end{pmatrix}.$$

It was shown in [23] that $Y^{\odot *}$ has the representation $\mathbf{R}^M \times L^\infty\left([-h,0];\mathbf{R}^M\right)$ and that Y is $\odot$-reflexive with respect to T_{20}. Because X is $\odot$-reflexive with respect to T_{10}, as shown in section 3, it is plain that $X \times Y$ is $\odot$-reflexive with respect to T_0, that $(X \times Y)^{\odot *}$ is (isometrically isomorphic to) $X^{\odot *} \times Y^{\odot *}$, and that $j_{X\times Y}(X \times Y) = j_X(X) \times j_Y(Y)$ (here, of course, j_Z denotes the canonical embedding of Z into $Z^{\odot *}$). Note also that for $t > h$, the range of $T_{20}(t)$ lies in the subspace of Y consisting of the constant functions, which is finite dimensional. In particular, $T_{20}(t)$ is eventually compact. As T_{10} is nilpotent, the semigroup T_0 on $X \times Y$ is eventually compact.

The system (4.1)–(4.4) is equivalent to the abstract integral equation

$$(\text{AIE})\qquad u(t) = T_0(t)\begin{pmatrix} \varphi \\ \psi \end{pmatrix} + j^{-1}\left(\int_0^t T_0^{\odot *}(t-s)G(u(s))ds\right),$$

where $G : X \times Y \to X^{\odot *} \times Y^{\odot *}$ is defined by

$$(4.7)\qquad G(\varphi,\psi) = \sum_{i=1}^{N} F_{1i}(\varphi,\psi)\begin{pmatrix} r_i^{\odot *} \\ 0 \end{pmatrix} + \sum_{i=1}^{M} F_{2i}(\varphi,\psi)\begin{pmatrix} 0 \\ s_i^{\odot *} \end{pmatrix}.$$

Here $r_i^{\odot *} \in X^{\odot *}$ is the Heaviside function (3.18), and $s_i^{\odot *} = (f_i, 0) \in Y^{\odot *}$, where $\{f_1, f_2, \ldots, f_M\}$ is the standard basis of $\mathbf{R}^M$ and 0 is the zero element of $L^\infty\left([-h,0];\mathbf{R}^M\right)$. We are now exactly in the situation described in section 2.5.

The resolvent $R\left(\lambda, A_{10}^{\odot *}\right)$ of $T_{10}^{\odot *}$ was calculated in Corollary 3.4 and the associated vector $r_i(\lambda)$ in (3.28). An analogous computation for $T_{20}^{\odot *}$ shows that $\sigma(A_{20}) = \sigma(A_{20}^{\odot *}) = \{0\}$ and that the resolvent of $A_{20}^{\odot *}$ is given by

$$(4.8)\qquad \left(R\left(\lambda, A_{20}^{\odot *}\right)(\alpha,\psi)\right)(\theta)$$
$$= \frac{1}{\lambda}e^{\lambda\theta}\alpha + \int_\theta^0 e^{\lambda(\theta-\tau)}\psi(\tau)d\tau, \quad (\alpha,\psi) \in Y^{\odot *}, \quad \theta \in [-h,0].$$

In particular,

$$(4.9)\qquad s_i(\lambda) := \left(j^{-1}R\left(\lambda, A_{20}^{\odot *}\right)s_i^{\odot *}\right)(\theta) = \frac{1}{\lambda}e^{\lambda\theta}f_i, \quad \theta \in [-h,0].$$

If F is Fréchet differentiable, its derivative can be represented by the $(N+M) \times (N+M)$ matrix

$$(4.10)\qquad \begin{pmatrix} k_{11} & m_{12} \\ k_{21} & m_{22} \end{pmatrix},$$

where k_{11} and k_{21} are $N \times N$ (resp., $N \times M$) matrices of elements of $L^\infty([0,h]$ and m_{12} and m_{22} are $M \times N$ (resp., $M \times M$) matrices of elements in $NBV[0,h]$. The interpretation of (4.10) is that

$$F'(\overline{\varphi},\overline{\psi})\begin{pmatrix}\varphi\\ \psi\end{pmatrix} = \begin{pmatrix}\int_0^h k_{11}(\theta)\varphi(-\theta)d\theta + \int_0^h m_{12}(d\theta)\psi(-\theta)\\ \int_0^h k_{21}(\theta)\varphi(-\theta)d\theta + \int_0^h m_{22}(d\theta)\psi(-\theta)\end{pmatrix}. \tag{4.11}$$

Using the expressions (3.28) and (4.9) for $r_i(\lambda)$ and $s_i(\lambda)$, respectively, and the definition (2.18) of the matrix $M(\lambda)$, we deduce that

$$M(\lambda) = \begin{pmatrix}\widehat{k}_{11}(\lambda) & \frac{1}{\lambda}\widehat{dm}_{12}(\lambda)\\ \widehat{k}_{21}(\lambda) & \frac{1}{\lambda}\widehat{dm}_{22}(\lambda)\end{pmatrix}, \quad \lambda \neq 0, \tag{4.12}$$

where, as before, $\widehat{k}$ denotes the Laplace transform of k and $\widehat{dm}$ denotes the Laplace–Stieltjes transform of m,

$$\widehat{dm}(\lambda) = \int_0^h e^{-\lambda\theta} m(d\theta). \tag{4.13}$$

It now follows from Theorem 2.15 that $\lambda \neq 0$ is an eigenvalue of the generator of the linearized equation (LAIE) if and only if

$$\det\left(E - M(\lambda)\right) = 0 \tag{4.14}$$

and that the algebraic multiplicity of λ coincides with the order of λ as a root of (4.14). Clearly, for $\lambda \neq 0$, (4.14) is equivalent to

$$\det\left(\begin{pmatrix} E & 0\\ 0 & \lambda E\end{pmatrix} - \begin{pmatrix}\widehat{k}_{11}(\lambda) & \widehat{dm}_{12}(\lambda)\\ \widehat{k}_{21}(\lambda) & \widehat{dm}_{22}(\lambda)\end{pmatrix}\right) = 0. \tag{4.15}$$

As Theorem 2.17 shows that the singularity at $\lambda = 0$ is removable, we conclude that (4.15) is the characteristic equation for the (AIE) with T_0 and G as specified above.

5. Examples.

5.1. Cannibalistic interaction. Even though *size* is the more natural individual state variable used to describe cannibalistic interaction, we shall here use *age* as a substitute, while referring to [25, section 4.1] and [28] for size-structured models. We assume that individuals turn adult and start to reproduce upon reaching age $\overline{a}$. Furthermore, only adults practice cannibalism and their victims are juveniles. The vulnerability for intraspecific predation is defined by a function c of age, the support of which lies in $[0,\overline{a})$.

Let $\mathcal{F}_0(a)$ be the survival probability to at least age a with respect to causes of death other than cannibalism. Let $b(t)$ be the population birth rate at time t and $I_1(t)$ the total number of adults at time t. We assume that "standard" adult food (that is, food other than juveniles of their own kind) is available at a constant density and that an adult produces, from this food, offspring at a rate Z. Let $I_2(t)$ denote the rate at which an adult produces offspring at time t on the basis of the energy provided by its cannibalistic actions. Then, by definition,

$$b(t) = \left(Z + I_2(t)\right) I_1(t), \tag{5.1}$$

$$I_1(t) = \int_{\overline{a}}^\infty b(t-a)\mathcal{F}_0(a)e^{-\int_0^a c(\alpha)I_1(t-a+\alpha)d\alpha}da. \tag{5.2}$$

To these equations we add

$$I_2(t) = \int_0^{\overline{a}} b(t-a)\mathcal{F}_0(a)e^{-\int_0^a c(\alpha)I_1(t-a+\alpha)d\alpha}c(a)E(a)da, \tag{5.3}$$

expressing that the (instantaneous) offspring yield resulting from the consumption of an individual of age a is given by $E(a)$.

The system (5.1)–(5.3) is of the form (3.54)–(3.57) (albeit with two instead of one interaction variable), and thus the arguments provided in section 3.5 establish that the system is a (DE) on L^1 with a C^1-map F. To guarantee that the maximum delay is finite, we assume that $\mathcal{F}_0$ drops to zero at a finite age h (or, equivalently, that the μ_0 of (3.55) has a nonintegrable singularity at h).

By elementary manipulations one can eliminate $\overline{I}_2$ and $\overline{b}$ from the equations for nontrivial steady states to arrive at a single equation

$$Z = \frac{e^{C\overline{I}_1}}{\int_{\overline{a}}^h \mathcal{F}_0(a)da}\left(1 - \overline{I}_1\int_0^{\overline{a}} c(a)E(a)\mathcal{F}_0(a)e^{-\overline{I}_1\int_0^a c(\alpha)d\alpha}da\right) \tag{5.4}$$

for the unknown $\overline{I}_1$. Here

$$C := \int_0^{\overline{a}} c(a)da. \tag{5.5}$$

Next we consider Z (thus, in essence, the density of the standard food) as a bifurcation parameter. The formula (5.4) is an explicit expression for Z as a function of $\overline{I}_1$. If we insert $\overline{I}_1 = 0$ at the right-hand side of (5.4), we obtain the critical value

$$Z_{\text{crit}} = \frac{1}{\int_{\overline{a}}^h \mathcal{F}_0(a)da} \tag{5.6}$$

such that newborn individuals, on average, produce exactly one offspring. In the absence of cannibalism (i.e., for $c \equiv 0$), a nontrivial steady state exists if and only if $Z > Z_{\text{crit}}$. By computing the derivative of Z with respect to $\overline{I}_1$ from (5.4) and evaluating at $\overline{I}_1 = 0$, one concludes that the condition

$$\int_0^{\overline{a}} (E(a)\mathcal{F}_0(a) - 1)\, c(a)da > 0 \tag{5.7}$$

guarantees that the bifurcation from the trivial steady state is subcritical in the sense that $\overline{I}_1$ is positive for values of Z slightly less than Z_{crit}. Thus if (5.7) holds, cannibalism allows the population to persist at levels of the standard food that are, by themselves, insufficient to sustain a consumer population. We refer once more to [25, section 4.1] and [28] for the biological interpretation and further elaborations.

A characteristic equation can now be derived as for the system (3.54)–(3.57) treated in section 3.5. The stability of the trivial steady state is governed by the position of the roots of

$$1 = Z\int_{\overline{a}}^h e^{-\lambda a}\mathcal{F}_0(a)\,da \tag{5.8}$$

in the complex plane. Hence the trivial solution is stable for $Z < Z_{\text{crit}}$ and unstable for $Z > Z_{\text{crit}}$. According to the principle of exchange of stability (see [15, 16, 44] and [3] for an application to population dynamics), the branch of positive steady states described by (5.4) is locally (i.e., for Z near Z_{crit}) stable if the bifurcation is supercritical and unstable if it is subcritical.

For the nontrivial steady states a detailed analysis of the global shape of the curve defined by (5.4) and the changes in the position of the roots of the associated characteristic equation along this curve requires a considerable effort and is beyond the scope of this paper. The point, however, is that the results of this paper allow one to derive conclusions about (in)stability and Hopf bifurcation from the appropriate information about these roots.

5.2. A structured metapopulation model. In this subsection we consider a metapopulation model first introduced in [33] and later modified and analyzed in [34, 35, 36, 37, 39, 41]. The model considers an infinite collection of identical patches that can support local populations. The structuring variable is the size x of a local population. Local populations may go extinct due to a catastrophe, but the vacated patch is immediately recolonized by migrants arriving from other patches. In PDE-formulation, the model is described by

$$\frac{\partial}{\partial t}n(t,x) + \frac{\partial}{\partial x}\left(f(x,D(t))n(t,x)\right) = -\mu(x)n(t,x), \quad t>0,\ x>0, \tag{5.9}$$

$$f(0,D(t))n(t,0) = \int_{\mathbf{R}_+} \mu(x)n(t,x)\,dx, \tag{5.10}$$

$$\frac{d}{dt}D(t) = -(\alpha+\nu)D(t) + \int_{\mathbf{R}_+} \gamma(x)n(t,x)\,dx, \quad t>0, \tag{5.11}$$

supplemented, of course, by appropriate initial conditions.

In (5.9)–(5.11), $n(t,\cdot)$ is the size-distribution of local populations at time t, and $D(t)$ is the density of dispersers at time t. $\gamma(x) = k(x)x$ is the emigration rate ($k(x)$ is the *per capita* emigration rate), α is the rate at which dispersers immigrate into a patch, and ν is the death rate during dispersal. $f(x,D)$ is the growth rate of a local population of size x when the density of dispersers is D. It is given by

$$f(x,D) = g(x) + \alpha D = r(x)x - k(x)x + \alpha D, \tag{5.12}$$

where $r(x)$ is the difference between the per capita birth and death rates when the local population size is x. Finally, $\mu(x)$ is the size-specific catastrophe rate of local populations.

Next we rewrite the equations (5.9)–(5.11) as a coupled system of the form (4.1)–(4.2). By the *age* of the local population of a patch we shall mean the time elapsed since the last catastrophe. Hence a local population of age a at time t had size zero at time $t-a$. The dynamics of such a local population is therefore described by the scalar ODE

$$\frac{d}{d\tau}x(\tau) = g(x(\tau)) + \alpha D_t(\tau - a), \quad 0<\tau\le a, \tag{5.13}$$

$$x(0) = 0. \tag{5.14}$$

For the solution of (5.13)–(5.14) we use the notation

$$x(\tau) = X(\tau, a, D_t), \quad 0\le\tau\le a. \tag{5.15}$$

The probability that a local population survives to age a, given the history of D, is

$$\mathcal{F}(a, D_t) = e^{-\int_0^a \mu(X(\tau,a,D_t))\,d\tau}. \tag{5.16}$$

The results formulated in this paper require a finite maximum life span. In the present model, this could be achieved by assuming that the catastrophe rate has a nonintegrable singularity at some finite local population size. However, in nature it is often the case that large local populations are much less prone to extinction than small ones, which experience a high risk of extinction due to demographic stochasticity. If this is the case, μ should rather be a decreasing function of local population size instead of blowing up. Also, exponentially distributed lifetimes (corresponding to constant catastrophe rates μ) occur frequently in applications. Fortunately, our theory carries over almost *verbatim* to the case of infinite delay (see section 6). In this example we shall therefore not make the assumption of a finite maximum life span. In particular, we shall allow the catastrophe rate μ to be constant.

We can now express the age-distribution

$$m(t, a) = f(X(a, a, D_t), D(t))n(t, X(a, a, D_t)) \tag{5.17}$$

of local populations in terms of the histories of the disperser density D and the birth rate

$$b(t) = f(0, D(t))n(t, 0) \tag{5.18}$$

of local populations as follows:

$$m(t, a) = b(t-a)\mathcal{F}(a, D_t) = b_t(-a)\mathcal{F}(a, D_t), \quad t \geq 0,\ 0 \leq a. \tag{5.19}$$

Equations (5.10) and (5.11) now yield the following system of a delay equation coupled with a delay differential equation:

$$b(t) = \int_0^\infty \mu(X(a, a, D_t))\mathcal{F}(a, D_t)b_t(-a)\,da, \tag{5.20}$$

$$\frac{d}{dt}D(t) = -(\alpha+\nu)D(t) + \int_0^\infty \gamma(X(a, a, D_t))\mathcal{F}(a, D_t)b_t(-a)\,da. \tag{5.21}$$

Here D plays the role of the environmental interaction variable. As we saw in section 4, the state space of b_t should be taken as L^1 and the space of D_t as C.

The steady state equation for (5.20), (5.21) is readily found. For constant functions b and D, (5.20) becomes an identity because

$$\int_0^\infty \mu(X(a, a, \overline{D}))\mathcal{F}(a, \overline{D})da = 1. \tag{5.22}$$

The identity (5.22) reflects the conservation of local populations: After a catastrophe, the patch is immediately recolonized. If we normalize the total amount of patches to 1, then

$$\overline{b} = \frac{1}{\int_0^\infty \mathcal{F}(a, \overline{D})da} \tag{5.23}$$

and the steady state condition becomes

$$\overline{D} = \frac{1}{\alpha+\nu} \cdot \frac{\int_0^\infty \gamma(X(a,a,\overline{D}))\mathcal{F}(a,\overline{D})da}{\int_0^\infty \mathcal{F}(a,\overline{D})da}. \tag{5.24}$$

The numerator on the right-hand side of (5.24) is the expected number of dispersers produced by a local population during its lifetime. When divided by the expected lifetime $\int_0^\infty \mathcal{F}(a,\overline{D})da$, it yields the average rate of dispersers produced by a patch, and when this rate is multiplied by the expected sojourn time $1/(\alpha+\nu)$ in the disperser pool, one gets the local population's contribution to the disperser pool. Equation (5.24) says that at equilibrium this contribution equals the steady disperser density (i.e., dispersers per patch).

In order to derive a characteristic equation and apply our theory, we have to show that the right-hand sides of (5.20) and (5.21) are differentiable in b_t and D_t. As they are linear in b_t, we only have to prove differentiability of $\varphi \mapsto X(\tau,a,\varphi)$ as a mapping on C. The differentiability of $\varphi \mapsto \mathcal{F}(a,\varphi)$ then follows immediately, and to obtain the desired result we only have to assume differentiability of the real functions μ and γ.

Assume that g is differentiable. Differentiating the integrated form of (5.13), (5.14),

$$X(\tau,a,\varphi) = \int_0^\tau g(X(\sigma,a,\varphi))\,d\sigma + \alpha\int_0^\tau \varphi(\sigma-a)\,d\sigma, \tag{5.25}$$

with respect to φ at $\overline{D}$, one obtains the linear equation

$$\frac{\partial}{\partial\varphi}X(\tau,a,\overline{D})\varphi = \int_0^\tau g'(X(\sigma,a,\overline{D}))\frac{\partial}{\partial\varphi}X(\sigma,a,\overline{D})\varphi\,d\sigma + \alpha\int_0^\tau \varphi(\sigma-a)\,d\sigma, \tag{5.26}$$

the solution of which is

$$\frac{\partial}{\partial\varphi}X(\tau,a,\overline{D})\varphi = \alpha\int_0^\tau e^{\int_\sigma^\tau g'(X(s,a,\overline{D}))\,ds}\varphi(\sigma-a)\,d\sigma. \tag{5.27}$$

Let us now assume that the catastrophe rate μ and the per capita emigration rate k are constant. The survival probability then becomes independent of φ: $\mathcal{F}(a) = \exp(-\mu a)$, and the equations (5.20), (5.21) simplify to

$$b(t) = \int_0^\infty \mu e^{-\mu a} b_t(-a)\,da, \tag{5.28}$$

$$\frac{d}{dt}D(t) = -(\alpha+\nu)D(t)$$
$$+\, k\int_0^\infty X(a,a,D_t)e^{-\mu a}b_t(-a)\,da, \tag{5.29}$$

while the steady state condition (5.24) simplifies to

$$\overline{D} = \frac{\mu k}{\alpha+\nu}\int_0^\infty X(a,a,\overline{D})e^{-\mu a}da. \tag{5.30}$$

We take the per capita emigration rate k as a bifurcation parameter. Note that $X(a,a,\overline{D})$, being the solution of $dx/da = r(x)x - kx + \alpha\overline{D}$, $x(0)=0$, depends on k, so in general one cannot solve (5.30) explicitly for k as a function of $\overline{D}$.

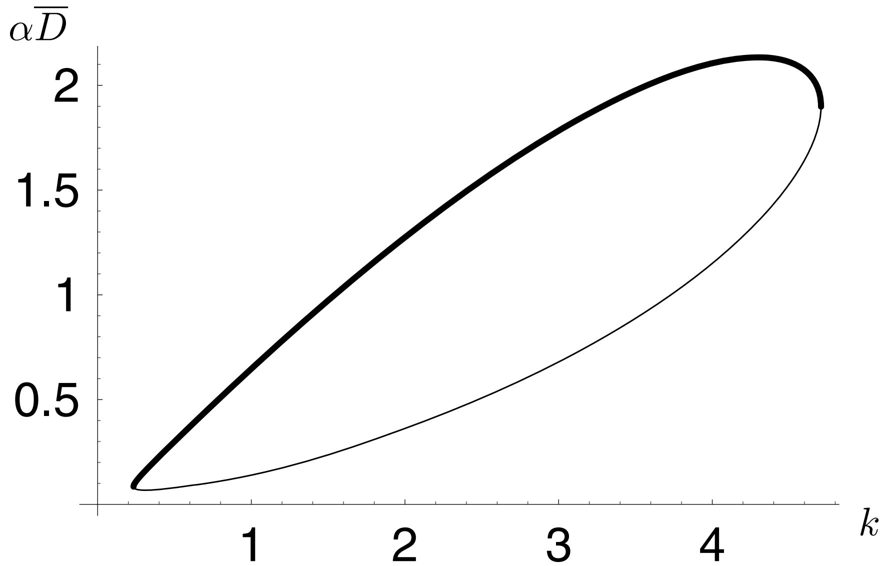


FIG. 1. *Equilibrium values for the immigration rate $\alpha\overline{D}$ in the case of an Allee effect with $f(x, D)$ given by* (5.32). *Parameters: $\alpha = 0.5$, $\mu = 0.2$, $\nu = 0.1$, $H = 1$, $\beta = 18$, $c = 1$, $d = 8$.*

Next we assume that there is an *Allee effect*, that is, that small local populations have a negative intrinsic growth rate [1] and therefore cannot persist without a sufficiently large immigration rate. We model this by assuming that the per capita birth rate depends on the local population size x as

$$\frac{\beta x}{H + x} \tag{5.31}$$

for some positive constants β and H. For a discussion of the rationale for this choice and its biological interpretation we refer to [18, 38]. Furthermore, if we make the standard assumption of density-dependent death rate as in the logistic equation, we end up with

$$f(x, D) = \left(\frac{\beta x}{H + x} - c - dx \right) x - kx + \alpha D \tag{5.32}$$

for some positive constants c and d.

It is clear that with the choice (5.32), the curve defined by (5.30) in the $k\overline{D}$-plane does not touch the axis $\overline{D} = 0$. As a matter of fact, as shown in [39], equation (5.30) defines a closed curve like the one depicted in Figure 1, at least for some choices of parameter values. As seen in Figure 1, there is a saddle-node bifurcation at $k \approx 0.2$ and another one at $k \approx 4.9$. In contrast to the situation with the transcritical bifurcation treated in subsection 5.1, we cannot allude here to the principle of exchange of stability to determine which of the two branches is stable and which is not. That information has to be deduced from the characteristic equation, which we now derive.

The linearized version of (5.28), (5.29) is

$$\psi(t) = \int_0^\infty \mu e^{-\mu a} \psi(t - a)\, da, \tag{5.33}$$

$$\frac{d}{dt}\varphi(t) = -(\alpha+\nu)\varphi(t) + k\int_0^\infty Y(a,\overline{D})\psi(t-a)\,da$$
$$+\,\alpha k\int_0^\infty Z(a,\overline{D})\varphi(t-a)\,da, \tag{5.34}$$

where

$$Y(a,\overline{D}) = X(a,a,\overline{D})e^{-\mu a}, \tag{5.35}$$

$$Z(a,\overline{D}) = \mu\int_a^\infty e^{\int_{\sigma-a}^{\sigma} g'(X(\tau,\sigma,\overline{D}))d\tau} e^{-\mu\sigma}\,d\sigma. \tag{5.36}$$

Taking the Laplace transform of (5.33), (5.34), one obtains

$$\widehat{\psi}(\lambda) = \frac{\mu}{\mu+\lambda}\widehat{\psi}(\lambda), \tag{5.37}$$

$$\lambda\widehat{\varphi}(\lambda) - \varphi(0) = -(\alpha+\nu)\widehat{\varphi}(\lambda) + k\widehat{Y}(\lambda,\overline{D})\widehat{\psi}(\lambda) + \alpha k\widehat{Z}(\lambda,\overline{D})\widehat{\varphi}(\lambda). \tag{5.38}$$

Hence the characteristic equation is

$$\det\begin{pmatrix} 1-\frac{\mu}{\mu+\lambda} & 0 \\ -k\widehat{Y}(\lambda,\overline{D}) & \lambda+\alpha+\nu-\alpha k\widehat{Z}(\lambda,\overline{D}) \end{pmatrix} = 0. \tag{5.39}$$

$\lambda = 0$ is always a root of (5.39). The reason is the indeterminacy of $\overline{b}$ explained above. The situation is analogous to the simple ODE SIS-model of mathematical epidemiology. If one treats the SIS model as a two-dimensional ODE, zero is an eigenvalue, which disappears after the substitution $S = N - I$ (N is the total population). Similarly, in our case the stability of the steady state is determined by the location in the complex plane of the roots of the equation

$$\lambda+\alpha+\nu-\alpha k\widehat{Z}(\lambda,\overline{D}) = 0. \tag{5.40}$$

For $\lambda \neq -(\alpha+\nu)$, (5.40) is equivalent to

$$1-\alpha k\frac{\widehat{Z}(\lambda,\overline{D})}{\lambda+\alpha+\nu} = 0, \tag{5.41}$$

and to this equation we can apply Nyquist's criterion (Corollary 2.20). The (numerical) results show that the upper branch (the thick line in Figure 1) is stable, while the lower branch (thin line) is unstable.

6. Discussion. The principle of linearized stability and the Hopf bifurcation theorem are among the fundamental results of the theory of ODEs. In the past three decades they have been generalized in various ways to infinite dimensional dynamical systems. In this paper we have used perturbation theory of adjoint semigroups (sun-star-calculus) to prove the principle of linearized stability and the Hopf bifurcation theorem for Volterra functional equations. The sun-star-framework made it possible to treat fully nonlinear functional equations as semilinear problems by transforming the original equation into an abstract integral equation of variation-of-constants type.

The transformation of the fully nonlinear problem into a seminlinear problem was made possible by extending the originally given state space. The idea that one should extend the state space when dealing with Hopf bifurcation for delay differential equations was introduced by Chow and Mallet-Paret in 1977 in a pioneering paper [7]. The sun-star-framework provides a functional analytic elaboration of this idea.

The principle of linearized stability consists of two parts. The first part concerns stability and says that if all roots of the so-called characteristic equation associated with a steady state have negative real part, then the steady state is exponentially stable. The second part states that if at least one characteristic root has positive real part, then the steady state is unstable.

The proof of the stability part of the principle of linearized stability is relatively simple as it uses only standard estimates and Gronwall's inequality, and therefore this part can be rather easily generalized from the ODE setting to infinite dimensional systems. In contrast, the proof of the instability part is geometric in nature and is even in the finite dimensional case much more difficult than the proof of the stability part. As a consequence, infinite dimensional generalizations of the instability part are comparatively rare in the literature. In many cases authors hint that the instability part is valid, but without giving a formal proof.

In the important paper [19], Desch and Schappacher proved both the stability and instability parts of the principle of linearized stability for nonlinear perturbations of generators of strongly continuous semigroups. Following their proof, Clément et al. [11] proved both parts within the context of adjoint semigroups and Thieme [53] within the framework of integrated semigroups. In the book [23] sun-star-calculus was systematically used for stability and bifurcation analysis of delay differential equations.

Our main motivation comes from structured population dynamics. In their seminal paper [30], Gurtin and MacCamy proved the stability part of the principle of linearized stability for age-structured populations but passed the instability part with silence. The same applies to most of the papers published in the early 1980s (e.g., [31, 32]). In the first comprehensive book [56] on the mathematical theory of age-structured population dynamics, Webb treated both the stability and instability parts using semigroup methods. Finally, in a somewhat neglected paper [50], Prüß proved both the stability and instability parts in a very general setting of several interacting age-structured populations.

When one moves from age-structured models to general physiologically structured models, even results on stability become rare. Tucker and Zimmermann [55] proved the stability part for a class of models, which, however, did not allow for a finite number of states-at-birth. Calsina and Saldaña [5] considered a size-structured model in which all individuals are born with the same size and gave conditions for the existence of a global attractor. They also gave sufficient conditions for conditional convergence to a steady state. Here conditional convergence means that the size distribution converges to a steady distribution in L^1, given that the total population converges.

There is also a vast literature on the stability of Volterra integral equations

$$x(t) = \int_0^t k(s)x(t-s)ds + G(x)(t); \tag{6.1}$$

see [29] and the references and historical remarks therein. These results are usually based on a classical theorem of Paley and Wiener [48] or generalizations thereof. In its basic form, the Paley–Wiener theorem says that if the kernel k belongs to $L^1(\mathbf{R}_+)$, then its resolvent kernel r is in $L^1(\mathbf{R}_+)$ if the characteristic equation

$$\det\left(E - \widehat{k}(\lambda)\right) = 0 \tag{6.2}$$

has no roots in the closed half-plane $\{\lambda \in \mathbf{C} : \mathrm{Re}\lambda \geq 0\}$. Using the fact that the solution x of (6.1) satisfies

$$x(t) = G(x)(t) + \int_0^t r(t-s)G(x)(s)ds, \tag{6.3}$$

it is easy to show that if $G(x)$ is of higher order, then the zero solution of (6.1) is stable. If (6.2) has no roots in $\{\lambda \in \mathbf{C} : \mathrm{Re}\lambda \geq -\varepsilon\}$ for some $\varepsilon > 0$ (this is the case, for instance, if k has compact support), then 0 is exponentially stable. So the stability part of the principle of linearized stability is well known for Volterra integral equations. On the other hand, a clear statement of the instability part seems to be lacking in the literature (however, see [21]). In section 3.5 we showed that our general theory applies to equations of the type (6.1) (at least if k has compact support), and hence it provides the instability part of the principle of linearized stability for Volterra integral equations.

In some respects the theory presented in this paper is not general enough. It does not, for instance, encompass all population dynamical applications that we want to consider. First of all, we have made the assumption of a finite delay h. In applications to population dynamics this corresponds to the assumption of a maximum individual life span. Although true in nature, it disregards the (mathematically) important case of exponentially distributed lifetimes. However, this is not a serious defect. The assumption was made to have sun-reflexivity, which simplified analysis for the following reason: For a norm continuous function $f : [0, \infty) \to X^{\odot *}$, the weak*- integral

$$\int_0^t T_0^{\odot *}(t-\sigma)f(\sigma)d\sigma \tag{6.4}$$

takes values in $X^{\odot\odot}$ (Proposition 2.2). The key advantage of assuming sun-reflexivity is that then the integral automatically takes values in $j(X)$, so that we can apply j^{-1} to obtain an element of X. If, by lack of compactness, we do not have sun-reflexivity, it may still be the case that this integral takes values in $j(X)$ if we restrict f to take values in a certain subspace of $X^{\odot *}$. For (nonlinear) perturbation operators taking values in such a subspace, the complete machinery retains its strength and all the results carry through. We intend to elaborate on this very useful remark in detail in a separate publication, with two motivating examples: infinite delay and a continuum of birth states.

Secondly, the unknown $x(t)$, which in population dynamical applications is a vector consisting of the components of the birth rate and the environmental interaction variables, is a vector in $\mathbf{R}^N$. There are important applications, for instance, models of size-dependent cannibalism [8], which require an infinite dimensional environmental condition. Prüß [50] treated an age-structured model, and Calsina and Saldaña [6] a size-structured model with an infinite dimensional environmental condition by other means, but it is unclear how the results of the present paper could be extended to cover that situation.

Thirdly, because the Nemytskiĭ operator from $L^1([-h, 0]; \mathbf{R}^N)$ to $L^1([-h, 0]; \mathbf{R}^N)$ generated by a smooth function $g : \mathbf{R}^N \to \mathbf{R}^N$ is Fréchet differentiable if and only if g is affine, we have to assume in applications to population dynamics that, for instance, the death rate is of the form $\mu(\xi, I) = \mu_0(\xi) + \mu_1(\xi)I$, where ξ is the individual state variable and I the interaction variable. Interestingly, this affine form, which corresponds to mass action interaction, is biologically the most relevant. In the future we shall investigate this aspect in detail in collaboration with J. A. J. Metz.

Appendix. Proof of Theorem 2.8. In this appendix we prove that if T_0 is eventually compact and if the perturbation $B : X \to X^{\odot *}$ is a compact operator, then the semigroup T defined by (LAIE) is eventually compact. This does not seem to have been stated in the literature yet. Clément et al. [13] proved the eventual compactness of the perturbed semigroup under the slightly weaker assumption that $R(\lambda, A_0^{\odot *})B$ is compact, but in addition to that they needed the assumption that

$$t \mapsto \int_0^t T_0^{\odot *}(t-\tau)BT(\tau)d\tau$$

is eventually uniformly continuous (that is, continuous from $[t_0, \infty)$ to $\mathcal{L}(X)$ equipped with the uniform operator topology, for some t_0).

The corresponding result for the case in which B maps X into X is known [26, Proposition 1.14, p. 166]. Without the compactness assumption on B the statement is false [26, Example 1.15, p. 166]. Therefore, the task in Exercise 2.5 of [23, p. 57] is impossible.

The proof in [26] is rather opaque, as it is based on statements like "without loss of generality . . . we may . . . assume that [X] is $C[0,1]$" [26, p. 525]. The proof provided here, which also covers the case in which the range of B lies in X, is more straightforward as it depends only on basic properties of semigroups and integrals.

Note. After we had finished this paper, Horst Thieme pointed out to us that Theorem 2.8 is an easy consequence of [54, Theorem 3], which he proved using the theory of integrated semigroups.

PROPOSITION A.1. *Let $B : X \to X^{\odot *}$ be compact. Then $j^{-1} \int_0^t T_0^{\odot *}(\tau)Bd\tau$ is a compact operator from X to X.*

Proof. By Schauder's theorem, $B^* : X^{\odot **} \to X^*$ is compact, and hence so is its "restriction" to $X^\odot$. Because the composition of a compact operator and a bounded operator is compact, it follows that $B^* \int_0^t T_0^\odot(\tau)d\tau : X^\odot \to X^*$ is compact. Using Schauder's theorem once more, we conclude that

$$\left(\left(B^* \int_0^t T_0^\odot(\tau)d\tau\right)^*\right)\Big|_X = j^{-1} \int_0^t T_0^{\odot *}(\tau)Bd\tau$$

is compact, as asserted. $\square$

Let V be a subset of a Banach space. In what follows, $\overline{\mathrm{con}}\, V$ denotes the closed convex hull of V, that is, the smallest closed convex set that contains V. Without any specifications, closedness refers to the norm topology. When other topologies are considered, the topology is indicated by a subscript. For instance, if $V \subset X^*$, then $\overline{\mathrm{con}}_{\sigma(X^*,X)}\, V$ is the smallest weakly* closed convex set that contains V.

The closed ball of radius r with center at x is denoted by $U(x, r)$.

THEOREM A.2. *Let $B : X \to X^{\odot *}$ be compact. Then $j^{-1} \int_0^t T_0^{\odot *}(\tau)BT(t-\tau)d\tau$ is a compact operator from X to X.*

Proof. Because T is a strongly continuous semigroup on X, the function $y : \tau \mapsto T(t-\tau)x$ is continuous from $[0, t]$ to X, and its range belongs to $U(0, M)$ for all $x \in U(0, 1)$ for some $M \geq 1$. Because $j^{-1} \int_0^t T_0^{\odot *}(\tau)Bd\tau$ is compact,

$$j^{-1} \int_0^t T_0^{\odot *}(\tau)Bd\tau(U(0, M))$$

is relatively compact, and hence

$$\overline{\mathrm{con}}\, j^{-1} \int_0^t T_0^{\odot *}(\tau) B d\tau (U(0,M))$$

is compact [51, Theorem 3.25, p. 72]. The proof is therefore completed if we can show that

$$j^{-1} \int_0^t T_0^{\odot *}(\tau) B T(t-\tau) x d\tau \in t\, \overline{\mathrm{con}}\, j^{-1} \int_0^t T_0^{\odot *}(\tau) B d\tau (U(0,M)).$$

This last statement is proved in the next lemmas. ◻

LEMMA A.3. *Let $t \mapsto x^*(t)$ be a weakly* continuous function from $[a,b]$ to X^*. Then*

$$\int_a^b x^*(t) dt \in (b-a)\, \overline{\mathrm{con}}_{\sigma(X^*,X)}\, x^*([a,b]).$$

Proof. First note that by the uniform boundedness principle a weakly* continuous function is norm bounded on compact intervals. By the definition of the weak* integral, one has that $\int_a^b x^*(t)dt$ belongs to the ball

$$U\left(0, (b-a) \sup_{a \le t \le b} \|x^*(t)\|\right),$$

which is weakly* compact by the Banach–Alaoglu theorem. Clearly $x^*([a,b]) \subset U\left(0, \sup_{a\le t\le b} \|x^*(t)\|\right)$, which is convex and weakly* compact. Hence

$$\overline{\mathrm{con}}_{\sigma(X^*,X)}\, x^*([a,b])$$

is weakly* compact. Theorem 3.27 of [51] now implies the assertion. ◻

LEMMA A.4. *Let $y : [0,t] \to X$ be continuous. Then*

$$x_0 := j^{-1} \int_0^t T_0^{\odot *}(\tau) B y(t-\tau) d\tau \in t\, \overline{\mathrm{con}}\, j^{-1} \int_0^t T_0^{\odot *}(\tau) B d\tau\, y([0,t]).$$

Proof. Because $j^{-1} \int_0^t T_0^{\odot *}(\tau) B d\tau$ is a compact operator from X to X (Proposition A.1), the set $V := j^{-1} \int_0^t T_0^{\odot *}(\tau) B d\tau\, y([0,t])$ is relatively compact in X, and hence $t\, \overline{\mathrm{con}}\, V$ is compact [51, Theorem 3.25, p. 72]. It follows that $t\, \overline{\mathrm{con}}\, j(V) = j(t\, \overline{\mathrm{con}}\, V)$ is compact in $X^{\odot *}$. Because $\sigma(X^{\odot *}, X^{\odot})$ is weaker than the norm topology of $X^{\odot *}$, the set $t\, \overline{\mathrm{con}}\, j(V)$ is also $\sigma(X^{\odot *}, X^{\odot})$-compact.

Suppose x_0 does *not* belong to $t\, \overline{\mathrm{con}}\, V$ or, equivalently, $jx_0 \notin t\, \overline{\mathrm{con}}\, j(V)$. A version of the Hahn–Banach theorem [51, Theorem 3.4, p. 58] then implies that there exist $x^{\odot} \in X^{\odot}$ and $\gamma \in \mathbf{R}$ such that

$$\mathrm{Re}\, \langle x^{\odot}, jx_0 \rangle < \gamma < \mathrm{Re}\, \langle x^{\odot}, x^{\odot *} \rangle \quad \text{for all } x^{\odot *} \in t\, \overline{\mathrm{con}}\, j(V).$$

So

$$\left\{ x^{\odot *} \in X^{\odot *} : \mathrm{Re} \langle x^{\odot}, x^{\odot *} - jx_0 \rangle < \gamma \right\}$$

is a $\sigma(X^{\odot *}, X^{\odot})$-neighborhood of jx_0 which does *not* intersect $t\, \overline{\mathrm{con}}\, j(V)$. Hence $jx_0 \notin t\, \overline{\mathrm{con}}_{\sigma(X^{\odot *}, X^{\odot})}\, j(V)$. But this contradicts Lemma A.3. ◻

Theorem 2.8 is now an immediate corollary of Theorem A.2.

Acknowledgments. We are very grateful to Pavol Brunovsky who kindly pointed out to us that the Nemytskiĭ operator from $L^1\left([-h,0];\mathbf{R}^N\right)$ to $L^1\left([-h,0];\mathbf{R}^N\right)$ generated by a smooth function $f:\mathbf{R}^N\to\mathbf{R}^N$ is Fréchet differentiable if and only if f is affine, and that the same necessary and sufficient conditions hold for the operator taking I in $L^1[0,h]$ to the solution $x\in C[0,h]$ of the differential equation $\dot{x}(t)=f(x(t),I(t))$.

During the past years O. D., M. G., and J. A. J. Metz have had the opportunity to meet regularly for a couple of weeks each year thanks to the hospitality of the Volkswagen-Stiftung (Research in Pairs programme at Oberwolfach), the Universities of Utrecht and Helsinki, as well as the International Institute for Applied Systems Analysis (IIASA). In the spring semester of 2006, M. G. held the F. C. Donders Visiting Chair of Mathematics at the University of Utrecht. The present paper is the result of research done at these *rendezvous*.

REFERENCES

[1] W. C. Allee, A. Emerson, T. Park, and K. Schmidt, *Principles of Animal Ecology*, Saunders, Philadelphia, 1949.

[2] W. Arendt, A. Grabosch, G. Greiner, U. Groh, H. P. Lotz, U. Moustakas, R. Nagel, F. Neubrander, and U. Schlotterbeck, *One-Parameter Semigroups of Positive Operators*, Lecture Notes in Math. 1184, Springer-Verlag, Berlin, 1986.

[3] B. Boldin, *Introducing a population into a steady community: The critical case, the center manifold, and the direction of bifurcation*, SIAM J. Appl. Math., 66 (2006), pp. 1424–1453.

[4] P. L. Butzer and H. Berens, *Semi-Groups of Operators and Approximation*, Springer-Verlag, Berlin, Heidelberg, New York, 1967.

[5] À. Calsina and J. Saldaña, *A model of physiologically structured population dynamics with a nonlinear individual growth rate*, J. Math. Biol., 33 (1995), pp. 335–364.

[6] À. Calsina and J. Saldaña, *Asymptotic behaviour of a model of hierarchically structured population dynamics*, J. Math. Biol., 35 (1997), pp. 967–987.

[7] S.-N. Chow and J. Mallet-Paret, *Integral averaging and bifurcation*, J. Differential Equations, 26 (1977), pp. 112–159.

[8] D. Claessen and A. M. de Roos, *Bistability in a size-structured population model of cannibalistic fish—a continuation study*, Theor. Popul. Biol., 64 (2003), pp. 49–65.

[9] Ph. Clément, O. Diekmann, M. Gyllenberg, H. J. A. M. Heijmans, and H. R. Thieme, *Perturbation theory for dual semigroups.* I. *The sun-reflexive case*, Math. Ann., 277 (1987), pp. 709–725.

[10] Ph. Clément, O. Diekmann, M. Gyllenberg, H. J. A. M. Heijmans, and H. R. Thieme, *Perturbation theory for dual semigroups.* II. *Time-dependent perturbations in the sun-reflexive case*, Proc. Roy. Soc. Edinburgh Sect. A, 109 (1988), pp. 145–172.

[11] Ph. Clément, O. Diekmann, M. Gyllenberg, H. J. A. M. Heijmans, and H. R. Thieme, *Perturbation theory for dual semigroups.* III. *Nonlinear Lipschitz continuous perturbations in the sun-reflexive case*, in Volterra Integrodifferential Equations in Banach Spaces and Applications, Pitman Res. Notes Math. Ser. 190, G. da Prato and M. Iannelli, eds., Longman Scientific and Technical, Harlow, UK, 1989, pp. 67–89.

[12] Ph. Clément, O. Diekmann, M. Gyllenberg, H. J. A. M. Heijmans, and H. R. Thieme, *Perturbation theory for dual semigroups.* IV. *The intertwining formula and the canonical pairing*, in Trends in Semigroup Theory and Applications, Ph. Clément, S. Invernizzi, E. Mitidieri, and I. I. Vrabie, eds., Dekker, New York, pp. 95–116.

[13] Ph. Clément, H. J. A. M. Heijmans, S. Angenent, C. J. van Duijn, and B. de Pagter, *One-Parameter Semigroups*, North–Holland, Amsterdam, 1987.

[14] R. Courant and D. Hilbert, *Methoden der Mathematischen Physik.* I, Springer-Verlag, Berlin, 1924.

[15] M. G. Crandall and P. H. Rabinowitz, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.

[16] M. G. Crandall and P. H. Rabinowitz, *The principle of exchange of stability*, in Dynamical Systems, A. R. Bednarek and L. Cesari, eds., Academic Press, New York, London, 1977, pp. 27–41.

[17] J. M. Cushing, *An Introduction to Structured Population Dynamics*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 71, SIAM, Philadelphia, 1998.
[18] B. Dennis, *Allee effects: Population growth, critical density, and the chance of extinction*, Natur. Resource Modeling, 3 (1989), pp. 1–58.
[19] W. Desch and W. Schappacher, *Linearized stability for nonlinear semigroups*, in Differential Equations in Banach Spaces, Lecture Notes in Math. 1223, A. Favini and E. Obrecht, eds., Springer-Verlag, Berlin, 1986, pp. 61–73.
[20] O. Diekmann, *Perturbed dual semigroups and delay equation*, in Dynamics of Infinite Dimensional Systems, S.-N. Chow and J. K. Hale, eds., Springer-Verlag, New York, 1987, pp. 67–74.
[21] O. Diekmann and S. A. van Gils, *Invariant manifolds for Volterra integral equations of convolution type*, J. Differential Equations, 54 (1984), pp. 139–180.
[22] O. Diekmann and S. M. Verduyn Lunel, *A new short proof of an old folk theorem in functional-differential equations*, in Semigroup Theory and Evolution Equations, Lect. Notes in Pure and Appl. Math. 135, Dekker, New York, 1991, pp. 101–106.
[23] O. Diekmann, S. A. van Gils, S. M. Verduyn Lunel, and H.-O. Walther, *Delay Equations. Functional, Complex, and Nonlinear Analysis*, Springer-Verlag, New York, 1995.
[24] O. Diekmann, M. Gyllenberg, H. Huang, M. Kirkilionis, J. A. J. Metz, and H. R. Thieme, *On the formulation and analysis of general deterministic structured population models.* II. *Nonlinear theory*, J. Math. Biol., 43 (2001), pp. 157–189.
[25] O. Diekmann, M. Gyllenberg, and J. A. J. Metz, *Steady-state analysis of structured population models*, Theor. Popul. Biol., 65 (2003), pp. 309–338.
[26] K.-J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations*, Springer-Verlag, New York, 2000.
[27] Ph. Getto, *On Some Quasilinear Structured Population Models*, Ph.D thesis, Utrecht University, Utrecht, The Netherlands, 2005; available online from http://igitur-archive.library.uu.nl/dissertations/2005-0816-200002/index.htm.
[28] Ph. Getto, O. Diekmann, and A. M. de Roos, *On the (dis)advantages of cannibalism*, J. Math. Biol., 51 (2005), pp. 695–712.
[29] G. Gripenberg, S.-O. Londen, and O. Staffans, *Volterra Integral and Functional Equations*, Cambridge University Press, Cambridge, UK, 1990.
[30] M. E. Gurtin and R. C. MacCamy, *Non-linear age-dependent population dynamics*, Arch. Rational Mech. Anal., 54 (1974), pp. 281–300.
[31] M. Gyllenberg, *Nonlinear age-dependent population dynamics in continuously propagated bacterial cultures*, Math. Biosci., 62 (1982), pp. 45–74.
[32] M. Gyllenberg, *Stability of a nonlinear age-dependent population model containing a control variable*, SIAM J. Appl. Math., 43 (1983), pp. 1418–1438.
[33] M. Gyllenberg and I. Hanski, *Single-species metapopulation dynamics: A structured model*, Theor. Popul. Biol., 42 (1992), pp. 35–62.
[34] M. Gyllenberg and I. Hanski, *Habitat deterioration, habitat destruction and metapopulation persistence in a heterogeneous landscape*, Theor. Popul. Biol., 52 (1997), pp. 198–215.
[35] M. Gyllenberg and J. A. J. Metz, *On fitness in structured metapopulations*, J. Math. Biol., 43 (2001), pp. 545–560.
[36] M. Gyllenberg and K. Parvinen, *Necessary and sufficient conditions for evolutionary suicide*, Bull. Math. Biol., 63 (2001), pp. 981–993.
[37] M. Gyllenberg, I. Hanski, and A. Hastings, *Structured metapopulation models*, in Metapopulation Dynamics: Ecology, Genetics and Evolution, I. Hanski and M. Gilpin, eds., Academic Press, London, 1997, pp. 93–122.
[38] M. Gyllenberg, J. Hemminki, and T. Tammaru, *Allee effects can both conserve and create spatial heterogeneity in population densities*, Theor. Popul. Biol., 56 (1999), pp. 231–242.
[39] M. Gyllenberg, K. Parvinen, and U. Dieckmann, *Evolutionary suicide and evolution of dispersal in structured metapopulations*, J. Math. Biol., 45 (2002), pp. 79–105.
[40] J. K. Hale and S. M. Verduyn Lunel, *Introduction to Functional-Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1991.
[41] I. Hanski and M. Gyllenberg, *Two general metapopulation models and the core-satellite species hypothesis*, Am. Naturalist, 142 (1993), pp. 17–41.
[42] E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups*, AMS, Providence, RI, 1957.
[43] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, New York, 1996.
[44] H. Kielhöfer, *Bifurcation Theory. An Introduction with Applications to PDEs*, Springer-Verlag, New York, 2004.
[45] J. A. J. Metz and O. Diekmann, *The Dynamics of Physiologically Structured Populations*, Lecture Notes in Biomath. 68, Springer-Verlag, Berlin, 1986.

[46] I.P. Natanson, *Theorie der Funktionen einer reellen Veränderlichen*, Akademie-Verlag, Berlin, 1954.

[47] J. van Neerven, *The Adjoint of a Semigroup of Linear Operators*, Lecture Notes in Math. 1529, Springer-Verlag, Berlin, 1992.

[48] R. E. A. C. Paley and N. Wiener, *Fourier Transforms in the Complex Domain*, Amer. Math. Soc. Colloq. Publ. 19, AMS, Providence, RI, 1934.

[49] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[50] J. Prüß, *Stability analysis for equilibria in age-specific population dynamics*, Nonlinear Anal., 7 (1983), pp. 1291–1313.

[51] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1973.

[52] W. Rudin, *Real and Complex Analysis*, 2nd ed., McGraw-Hill, New York, 1974.

[53] H. R. Thieme, *Semiflows generated by Lipschitz perturbations of non-densely defined operators*, Differential Integral Equations, 3 (1990), pp. 1035–1066.

[54] H. R. Thieme, *Quasi-compact semigroups via bounded perturbation*, in Advances in Mathematical Population Dynamics: Molecules, Cells and Man, O. Arino, D. Axelrod, and M. Kimmel, eds., World Scientific, River Edge, NJ, 1997, pp. 691–711.

[55] S. L. Tucker and S. O. Zimmerman, *A nonlinear model of population dynamics containing an arbitrary number of continuous structure variables*, SIAM J. Appl. Math., 48 (1988), pp. 549–591.

[56] G. F. Webb, *Nonlinear Age-Dependent Population Dynamics*, Dekker, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 39, No. 4, pp. 1070–1111

SPECTRA OF LINEARIZED OPERATORS FOR NLS SOLITARY WAVES*

SHU-MING CHANG†, STEPHEN GUSTAFSON‡, KENJI NAKANISHI§, AND TAI-PENG TSAI‡

Abstract. Nonlinear Schrödinger equations (NLSs) with focusing power nonlinearities have solitary wave solutions. The spectra of the linearized operators around these solitary waves are intimately connected to stability properties of the solitary waves and to the long-time dynamics of solutions of NLSs. We study these spectra in detail, both analytically and numerically.

Key words. spectrum, linearized operator, nonlinear Schrödinger equation, solitary waves, stability

AMS subject classifications. 35Q55, 35P15

DOI. 10.1137/050648389

1. Introduction.

Consider the nonlinear Schrödinger equation (NLS) with focusing power nonlinearity,

$$i\partial_t\psi = -\Delta\psi - |\psi|^{p-1}\psi, \tag{1.1}$$

where $\psi(t,x) : \mathbb{R}\times\mathbb{R}^n \to \mathbb{C}$ and $1 < p < \infty$. Such equations arise in many physical settings, including nonlinear optics, water waves, and quantum physics. Mathematically, nonlinear Schrödinger equations with various nonlinearities are studied as basic models of nonlinear dispersive phenomena. In this paper, we stick to the case of a pure power nonlinearity for the sake of simplicity.

For a certain range of the power p (see below), the NLS (1.1) has special solutions of the form $\psi(t,x) = Q(x)\,e^{it}$. These are called *solitary waves.* The aim of this paper is to study the spectra of the *linearized operators* which arise when (1.1) is linearized around solitary waves. The main motivation for this study is that properties of these spectra are intimately related to the problem of the stability (orbital and asymptotic) of these solitary waves and to the long-time dynamics of solutions of NLSs.

Let us begin by recalling some well-known facts about (1.1). Standard references include [5, 33, 34]. Many basic results on the linearized operators we study here were proved by Weinstein [38, 39]. The Cauchy (initial value) problem for (1.1) is locally (in time) well-posed in $H^1(\mathbb{R}^n)$ if $1 < p < p_{\max}$, where

$$p_{\max} := 1 + 4/(n-2) \quad \text{if } n \ge 3; \quad p_{\max} := \infty \quad \text{if } n = 1, 2.$$

Moreover, if $1 < p < p_c$, where

$$p_c := 1 + 4/n,$$

*Received by the editors December 25, 2005; accepted for publication (in revised form) April 5, 2007; published electronically October 24, 2007.
http://www.siam.org/journals/sima/39-4/64838.html

†Department of Mathematics, National Tsing Hua University, Hsinchu, Taiwan (schang@mx.nthu.edu.tw). This author was partly supported by an NSC grant of Taiwan.

‡Department of Mathematics, University of British Columbia, Vancouver, BC, V6T1Z2, Canada (gustaf@math.ubc.ca, ttsai@math.ubc.ca). The research of these authors was partly supported by NSERC grants.

§Department of Mathematics, Kyoto University, Kyoto 606-8502, Japan (n-kenji@math.kyoto-u.ac.jp). This author was partly supported by JSPS grant 15740086 of Japan.

the problem is globally well-posed. For $p \geq p_c$, there exist solutions whose H^1-norms go to ∞ (*blow up*) in finite time. In this paper, the cases $p < p_c$, $p = p_c$, and $p > p_c$ are called *subcritical*, *critical*, and *supercritical*, respectively.

The set of all solutions of (1.1) is invariant under the symmetries of translation, rotation, phase, Galilean transform, and scaling: if $\psi(t,x)$ is a solution, then so is

$$\widetilde{\psi}(t,x) := \lambda^{2/(p-1)}\psi\left(\lambda^2 t,\ \lambda Rx - \lambda^2 tv - x_0\right) \exp\left\{i\left[\frac{\lambda Rx\cdot v}{2} - \frac{\lambda^2 tv^2}{4} + \gamma_0\right]\right\}$$

for any constant $x_0, v \in \mathbb{R}^n$, $\lambda > 0$, $\gamma_0 \in \mathbb{R}$, and $R \in O(n)$. When $p = p_c$, there is an additional symmetry called the "pseudoconformal transform" (see [34, p. 35]).

We are interested here in solutions of (1.1) of the form

$$\psi(t,x) = Q(x)\, e^{it}, \tag{1.2}$$

where $Q(x)$ must therefore satisfy the nonlinear elliptic equation

$$-\Delta Q - |Q|^{p-1}Q = -Q. \tag{1.3}$$

Any such solution generates a family of solutions by the above-mentioned symmetries, called *solitary waves*. Solitary waves are special examples of *nonlinear bound states*, which, roughly speaking, are solutions that are spatially localized for all time. More precisely, one could define nonlinear bound states to be solutions $\psi(t,x)$ which are *nondispersive* in the sense that

$$\sup_{t\in\mathbb{R}} \inf_{x_0\in\mathbb{R}^n} \||x|\psi(t,x-x_0)\|_{L^2_x(\mathbb{R}^n)} < \infty.$$

Testing (1.3) with $\bar{Q}$ and $x.\nabla\bar{Q}$ and taking real parts, one arrives at the *Pohozaev identity* [28]

$$\frac{1}{2}\int |Q|^2 = b\frac{1}{p+1}\int |Q|^{p+1}, \quad \frac{1}{2}\int |\nabla Q|^2 = a\frac{1}{p+1}\int |Q|^{p+1}, \tag{1.4}$$

where

$$a = \frac{n(p-1)}{4}, \quad b = \frac{n+2-(n-2)p}{4}.$$

The coefficients a and b must be positive, and hence a necessary condition for existence of nontrivial solutions is $p \in (1, p_{\max})$.

For $p \in (1, p_{\max})$, and for all space dimensions, there exists at least one nontrivial radial solution $Q(x) = Q(|x|)$ of (1.3) (existence goes back to [28]). This solution, called a *nonlinear ground state*, is smooth, decreases monotonically as a function of $|x|$, decays exponentially at infinity, and can be taken to be positive: $Q(x) > 0$. It is the unique positive solution. (See [34] for references for the various existence and uniqueness results for various nonlinearities.) The ground state can be obtained as the minimizer of several different variational problems. One such result we shall briefly use later is that, for all $n \geq 1$ and $p \in (1, p_{\max})$, the ground state minimizes the Gagliardo–Nirenberg quotient

$$J[u] := \frac{\left(\int |\nabla u|^2\right)^a \left(\int u^2\right)^b}{\int u^{p+1}} \tag{1.5}$$

among nonzero $H^1(\mathbb{R}^n)$ radial functions (see Weinstein [38]).

For $n=1$, the ground state is the unique $H^1(\mathbb{R})$-solution of (1.3) up to translation and phase [5, Theorem 8.1.6, p. 259]. For $n \geq 2$, this is not the case: there are countably infinitely many radial solutions (still real-valued), denoted in this paper by $Q_{0,k,p}(x)$, $k = 0, 1, 2, 3, \ldots$, each with exactly k positive zeros as a function of $|x|$ (Strauss [32]; see also [2, 3]). In this notation, $Q_{0,0,p}$ is the ground state.

There are also nonradial (and complex-valued) solutions, for example, those suggested by Lions [22] with nonzero angular momenta,

$$n = 2, \quad Q = \phi(r)\, e^{im\theta} \quad \text{in polar coordinates } r, \theta;$$

$$n = 3, \quad Q = \phi(r, x_3)\, e^{im\theta} \quad \text{in cylindrical coordinates } r, \theta, x_3;$$

and similar definitions for $n \geq 4$. When $n = 2$, some of these solutions are denoted here by $Q_{m,k,p}$, with $p \in (1, p_{\max})$ and $k = 0, 1, 2, \ldots$ denoting their numbers of positive zeros. See section 4 for more details.

We will refer to all the solitary waves generated by $Q_{0,0,p}$ as *nonlinear ground states* and all others as *nonlinear excited states*. We are not aware of a complete characterization of all solutions of (1.3) or of (1.1). For example, the uniqueness of $Q_{m,k,p}$ with $m, k \geq 1$ is apparently open. Also, we do not know if there are "breather" solutions, analogous to those of the generalized KdV equations. In this paper we will mainly study radial solutions (and in particular the ground state), but we will also briefly consider nonradial solutions numerically in section 4.

To study the stability of a solitary wave solution (1.2), one considers solutions of NLSs of the form

$$\psi(t, x) = [Q(x) + h(t, x)]\, e^{it}. \tag{1.6}$$

For simplicity, let $Q = Q_{0,0,p}$ be the ground state for the remainder of this introduction (see section 4 for the general case). The perturbation $h(t, x)$ satisfies an equation

$$\partial_t h = \mathcal{L}h + (\text{nonlinear terms}), \tag{1.7}$$

where $\mathcal{L}$ is the *linearized operator* around Q:

$$\mathcal{L}h = -i\left\{(-\Delta + 1 - Q^{p-1})h - \tfrac{p-1}{2}\, Q^{p-1}(h + \bar{h})\right\}. \tag{1.8}$$

It is convenient to write $\mathcal{L}$ as a matrix operator acting on $\left[\begin{smallmatrix}\operatorname{Re} h \\ \operatorname{Im} h\end{smallmatrix}\right]$,

$$\mathcal{L} = \begin{bmatrix} 0 & L_- \\ -L_+ & 0 \end{bmatrix}, \tag{1.9}$$

where

$$L_+ = -\Delta + 1 - pQ^{p-1}, \quad L_- = -\Delta + 1 - Q^{p-1}. \tag{1.10}$$

Clearly the operators L_- and L_+ play a central role in the stability theory. They are self-adjoint Schrödinger operators with continuous spectrum $[1, \infty)$ and with finitely many eigenvalues below 1. In fact, when Q is the ground state, it is easy to see that L_- is a nonnegative operator, while L_+ has exactly one negative eigenvalue (these facts follow from Lemma 2.2 below).

Because of its connection to the stability problem, the object of interest to us in this paper is the spectrum of the non-self-adjoint operator $\mathcal{L}$. The simplest properties of this spectrum are the following:

1. For all $p \in (1, p_{\max})$, 0 is an eigenvalue of $\mathcal{L}$.
2. The set $\Sigma_c := \{ir : r \in \mathbb{R}, |r| \geq 1\}$ is the continuous spectrum of $\mathcal{L}$.

(See the next section for the first statement. The second is easily checked.)

It is well known that the exponent $p = p_c$ is critical for stability of the ground state solitary wave (as well as for blow-up of solutions). For $p < p_c$ the ground state is *orbitally stable*, while for $p \geq p_c$ it is unstable (see [40, 14]). These facts have immediate spectral counterparts: for $p \in (1, p_c]$, all eigenvalues of $\mathcal{L}$ are purely imaginary, while for $p \in (p_c, p_{\max})$, $\mathcal{L}$ has at least one eigenvalue with positive real part.

The goal of this paper is to get a more detailed understanding of the spectrum of $\mathcal{L}$ using both analytical and numerical techniques. See [13, 7, 8, 9] for related work. This finer information is essential for understanding the long-time dynamics of solutions of NLSs : (i) To prove asymptotic (rather than simply orbital) stability, one often assumes either the linearized operator $\mathcal{L}$ has no nonzero eigenvalue, or its nonzero eigenvalues are $\pm ri$ with $0.5 < r < 1$. These assumptions need to be verified. (ii) To determine the rate of relaxation to stable solitary waves when there is a unique pair of nonzero eigenvalues $\pm ri$ with $0 < r < 1$, heuristic arguments suggest that $[1/r]$, the smallest integer no larger than $1/r$, may decide the rate. (iii) To construct stable manifolds of unstable solitary waves, one needs to know if there are eigenvalues which are not purely imaginary and to find their locations. These are highly active areas of current research; see, e.g., [15, 18, 30] and the references therein.

Interesting questions with direct relevance to these stability-type problems include the following:

(i) Can one determine (or estimate) the number and locations of the eigenvalues of $\mathcal{L}$ lying on the segment between 0 and i?
(ii) Can $\pm i$, the thresholds of the continuous spectrum Σ_c, be eigenvalues or resonances?
(iii) Can eigenvalues be embedded inside the continuous spectrum?
(iv) Can the linearized operator have eigenvalues with nonzero real *and* imaginary parts (this is already known not to happen for the ground state—see the next section—and so we pose this question with excited states in mind).
(v) Are there bifurcations, as p varies, of pairs of purely imaginary eigenvalues into pairs of eigenvalues with nonzero real part (a stability/instability transition)?

The detailed discussion of the numerical methods is postponed to the appendix. Roughly speaking, we first compute the nonlinear ground state by iteration and renormalization and then compute the spectra of various suitably discretized linear operators.

Let us now summarize the main results and observations of this paper.

1. *Numerics for spectra.* When Q is the ground state, we compute numerically the spectra of $\mathcal{L}$, L_+, and L_- as functions of p; see Figures 1–5. In these figures, the horizontal axis is the logarithm of $p-1$. The vertical axis is comprised of the following: Solid lines are purely imaginary eigenvalues of $\mathcal{L}$ (without i) for $p \in (1, p_c)$; dashed lines are real eigenvalues of $\mathcal{L}$; dotted lines are eigenvalues of L_+; dash-dot lines are eigenvalues of L_-. We have ignored imaginary eigenvalues of the discretized operators with modulus greater than one, which correspond to the continuous spectra of the original operators. Figure 1 is the one-dimensional case. Figures 2 and 3 are the spectra of these operators restricted to radial functions for space dimensions $n = 2$ and 3.

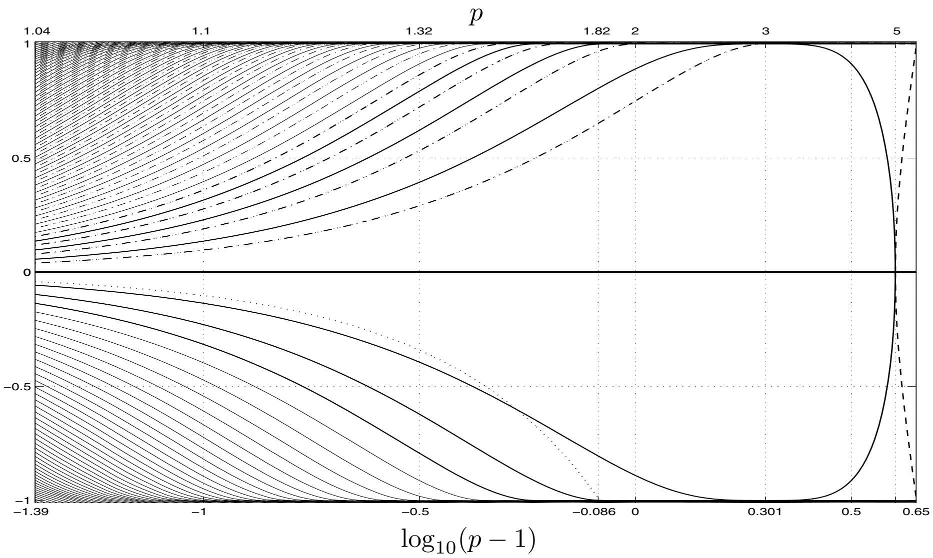


FIG. 1. *Spectra of $\mathcal{L}$, L_+, and L_- for $n = 1$ with logarithmic axis for the values of $p-1$. (Solid line: purely imaginary eigenvalues of $\mathcal{L}$; dashed line: real eigenvalues of $\mathcal{L}$; dotted line: eigenvalues of L_+; dash-dot line: eigenvalues of L_-).*

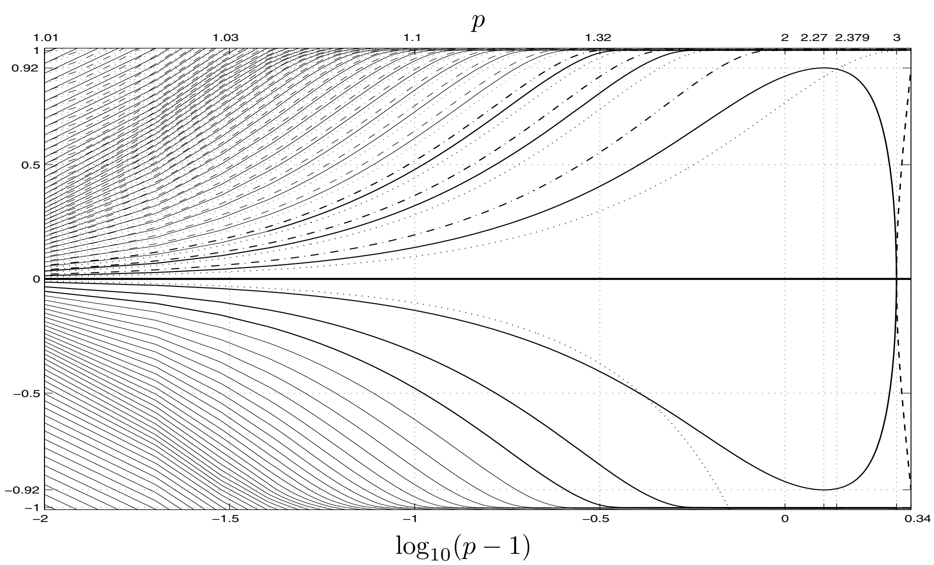


FIG. 2. *Spectra of $\mathcal{L}$, L_+, and L_- restricted to radial functions in the two-dimensional space with logarithmic axis for the values of $p-1$. (Solid line: purely imaginary eigenvalues of $\mathcal{L}$; dashed line: real eigenvalues of $\mathcal{L}$; dotted line: eigenvalues of L_+; dash-dot line: eigenvalues of L_-).*

Figures 4 and 5 are for $n = 2$ and are the spectra restricted to functions of the form $\phi(r)e^{i\theta}$ and $\phi(r)e^{i2\theta}$, respectively. These pictures shed some light on questions (i), (iv), and (v) above and to a certain extent on question (ii). Figures 10–15 are concerned with the spectra of excited states; see the discussion below.

2. *One-dimensional phenomena.* The case $n = 1$ is the easiest case to handle analytically. In section 3, we undertake a detailed study of the one-dimensional

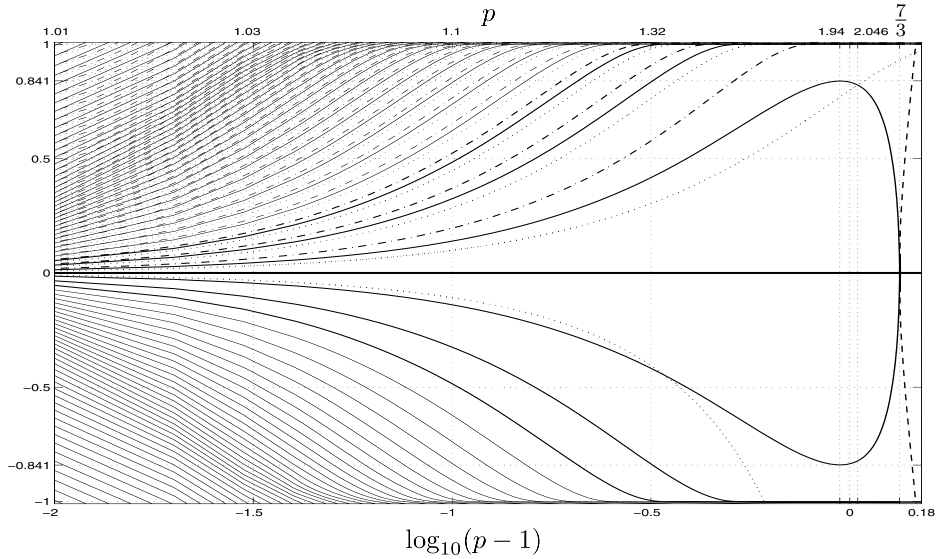


FIG. 3. *Spectra of $\mathcal{L}$, L_+, and L_- restricted to radial functions in the three-dimensional space.*

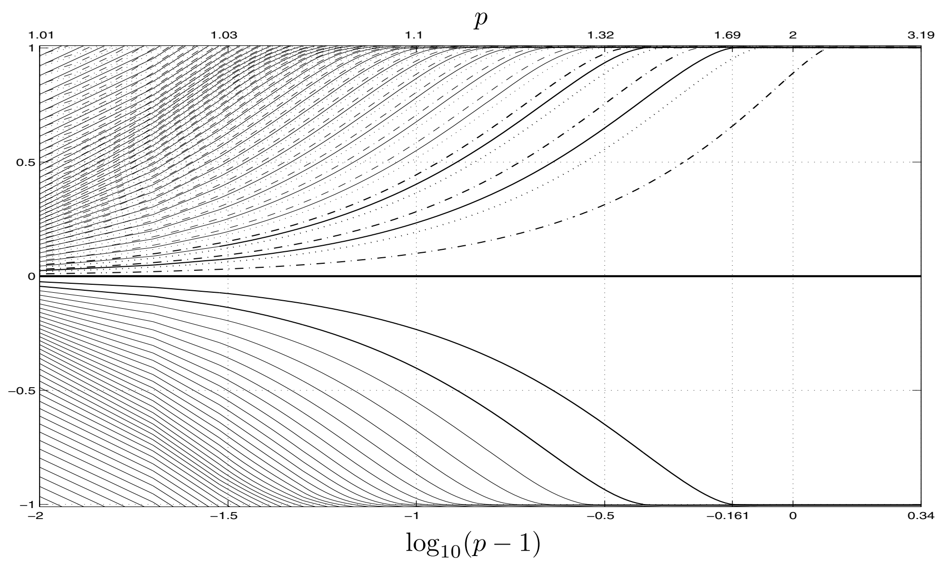


FIG. 4. *Spectra of $\mathcal{L}$, L_+, and L_- restricted to functions of the form $\phi(r)e^{i\theta}$ in the two-dimensional space.*

problem, giving rigorous proofs of a number of phenomena observed in Figure 1. One simple such phenomenon is the (actually classical) fact that the eigenvalues of L_+ and L_- exactly coincide, with the exception of the first (negative) eigenvalue of L_+ (note that this appears to be a strictly one-dimensional phenomenon: the eigenvalues of L_+ and L_- are different for $n \geq 2$, as Figures 2–5 indicate). In fact, we are able to prove sufficiently precise upper and lower bounds on the eigenvalues of $\mathcal{L}$ (lying outside the continuous spectrum) to determine their number, and estimate their positions, as functions of p (see Theorem 3.8). We use two basic techniques: an embedding of L_+ and L_- into a hierarchy of related operators, and a novel

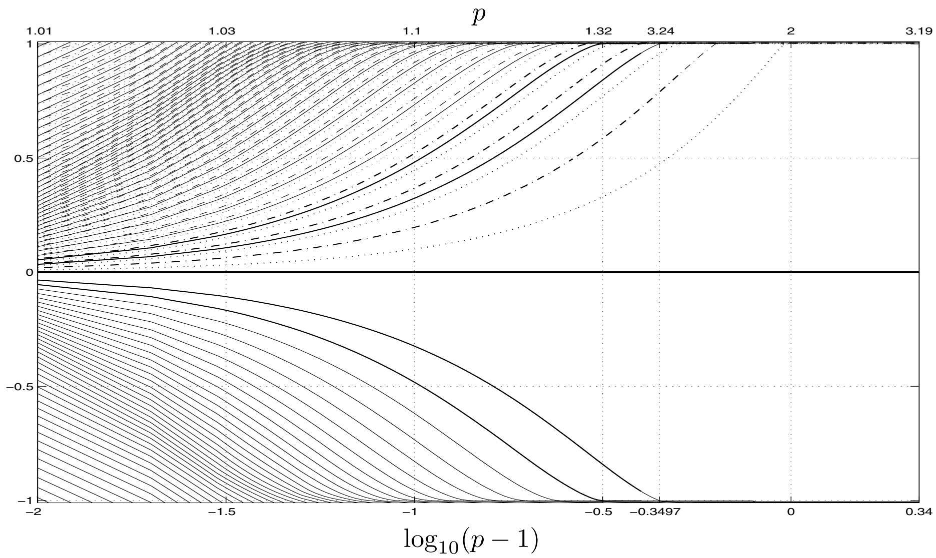


FIG. 5. *Spectra of $\mathcal{L}$, L_+, and L_- restricted to functions of the form $\phi(r)e^{i2\theta}$ in the two-dimensional space.*

variational problem for the eigenvalues, in terms of a fourth-order self-adjoint differential operator (see Theorem 3.6). In this way, we get a fairly complete answer to question (i) above for $n = 1$.

3. *Variational characterization of eigenvalues*. We present self-adjoint variational formulations of the eigenvalue problem for $\mathcal{L}$ in any dimension (see Summary 2.5), including the novel $n = 1$ formulation mentioned above. In principle, these provide a means of counting/estimating the eigenvalues of $\mathcal{L}$ (and hence addressing question (i) above in higher dimensions), though we obtain only such detailed information for $n = 1$.
4. *Bifurcation at $p = p_c$*. In each of Figures 1–3, a pair of purely imaginary eigenvalues for $p < p_c$ appears to collide at 0 at $p = p_c$ and becomes a pair of real eigenvalues for $p > p_c$. This is exactly the stability/instability transition for the ground state. We rigorously verify this picture, determining analytically the spectrum near 0 for p near p_c and making concrete a bifurcation picture suggested by M. I. Weinstein (personal communication); see Theorem 2.6. This gives a partial answer to question (v) above. It is worth pointing out that for $n = 1$, the imaginary part of the (purely imaginary) eigenvalue bifurcating for $p < p_c$ is always larger than the third eigenvalue of L_+ (the first is negative and the second is zero)—this is proved analytically in Theorem 3.8. For $n \geq 2$, however, they intersect at $p \approx 2.379$ for two dimensions and $p \approx 2.046$ for three dimensions, (see Figures 1–3).
5. *Interlacing property.* A numerical observation is that in all the figures, the adjacent eigenvalues of $\mathcal{L}$ each seem to bound an eigenvalue of L_+ and one of L_- (at least for p small enough). We are able to establish this "interlacing" property analytically in dimension one (see Theorem 3.8).
6. *Threshold resonance*. An interesting fact observed numerically (Figure 1) is that, in the one-dimensional case, as $p \to 3$, one eigenvalue curve converges to $\pm i$, the threshold of the continuous spectrum. One might suspect that, at $p = 3$, $\pm i$ corresponds to a resonance or embedded eigenvalue. It is indeed

a resonance: we find an explicit nonspatially-decaying "eigenfunction" and show numerically in section 3.7 that the corresponding eigenfunctions converges, as $p \to 3$, to this function. This observation addresses question (ii) above for $n = 1$.

7. *Excited states.* In section 4 we consider the spectra of linearized operators around excited states with nonzero angular momenta. We observe that, in addition to the bifurcation mentioned above at $p = p_c$, there are complex eigenvalues which are neither real nor purely imaginary (addressing question (iv) above; see Figures 10–15), symmetric w.r.t. both real and purely imaginary axes. These complex eigenvalues also come from bifurcation: as p decreases, a quadruple of complex eigenvalues will collide into the imaginary axis away from 0 and then split into four purely imaginary eigenvalues. It seems that all eigenvalues lie on the imaginary axis for $p \in (1, p_*)$ for some p_* close to 1. In other words, *numerically these excited states are spectrally stable for p close to* 1. It is possible that the numerical error increases enormously as $p \to 1_+$ due to the artificial boundary condition, since the spectrum is approaching the continuous one for $p = 1$. This has to be verified analytically in the future. Even if they are indeed spectrally stable, it is not clear if they are *nonlinearly* stable.

It is worth mentioning some important questions we *cannot* answer:

1. We are so far unable to give precise rigorous estimates on the number and positions of the eigenvalues of $\mathcal{L}$ for $n \geq 2$ (question (i) above).
2. We cannot exclude the possible existence of embedded eigenvalues (question (iii) above).
3. We do not know a nice variational formulation for eigenvalues of $\mathcal{L}$ when Q is an *excited* state (this problem is also linked to question (i) above).
4. We do not have a complete characterization of solitary waves or, more generally, of nonlinear bound states.

We end this introduction by describing some related numerical work. Buslaev and Grikurov [4] and Grikurov [12] study the linearized operators for solitary waves of the following one-dimensional NLS with $p < q$:

$$i\psi_t + \psi_{xx} + |\psi|^p\psi - \alpha|\psi|^q\psi = 0.$$

They draw the bifurcation picture for eigenvalues near zero when the parameter $\alpha > 0$ is near a critical value with the frequency of the solitary wave fixed. This picture is similar to Weinstein's picture, which we study in section 2.3.

Demanet and Schlag [9] consider the same linearization as us and study the supercritical case $n = 3$ and $p \leq 3$ near 3. In this case, it is numerically shown that both L_+ and L_- have no eigenvalues in $(0, 1]$ and no resonance at 1, a condition which implies (see [30]) that $\mathcal{L}$ has no purely imaginary eigenvalues in $[-i, 0) \cup (0, i]$ and no resonance at $\pm i$.

We outline the rest of the paper: in section 2 we consider general results for all dimensions for ground states. In section 3 we consider one-dimensional theory. In section 4 we discuss the spectra for excited states with angular momenta. In the appendix we discuss the numerical methods.

Notation. For an operator A, $N(A) = \{\phi \in L^2 |\ A\phi = 0\}$ denotes the nullspace of A. $N_g(A) = \cup_{k=1}^{\infty} N(A^k)$ denotes the generalized nullspace of A. The L^2-inner product in $\mathbb{R}^n$ is $(f, g) = \int_{\mathbb{R}^n} \bar{f} g \, dx$.

2. Revisiting the general theory for ground states. In this section we review mostly well-known results which are valid for all dimensions and for the ground state $Q(x) = Q_{0,0,p}(x)$, and we give new proofs of some statements.

We begin by recalling some well-known results for the linearized operator $\mathcal{L}$ defined by (1.8). As is well known for linearized Hamiltonian systems (and can be checked directly), if λ is an eigenvalue, then so are $-\lambda$ and $\pm\bar{\lambda}$. Hence if $\lambda \neq 0$ is real or purely imaginary, it comes in a pair. If it is complex with nonzero real and imaginary parts, it comes in a quadruple. It follows from nonlinear stability and instability results [40, 14] that all eigenvalues are purely imaginary if $p \in (1, p_c)$ and that there is at least one eigenvalue with positive real part when $p \in (p_c, p_{\max})$. It is also known (see, e.g., [8]) that the set of isolated and embedded eigenvalues is finite, and the dimensions of the corresponding generalized eigenspaces are finite.

2.1. L_+, L_-, and the generalized nullspace of $\mathcal{L}$. Here we recall the makeup of the generalized nullspace $N_g(\mathcal{L})$ of $\mathcal{L}$. Easy computations give

$$L_+ Q_1 = -2Q, \quad L_- Q = 0, \quad \text{where } Q_1 := \left(\tfrac{2}{p-1} + x \cdot \nabla\right)Q, \tag{2.1}$$

and

$$L_- xQ = -2\nabla Q, \quad L_+ \nabla Q = 0. \tag{2.2}$$

In the critical case $p = p_c$, we also have

$$L_-(|x|^2 Q) = -4Q_1, \quad L_+ \rho = |x|^2 Q \tag{2.3}$$

for some radial function $\rho(x)$ (for which we do not know an explicit formula in terms of Q). Denote

$$\delta^p_{p_c} = \begin{cases} 1, & p = p_c, \\ 0, & p \neq p_c. \end{cases} \tag{2.4}$$

For $1 < p < p_{max}$, the generalized nullspace of $\mathcal{L}$ is given by (see [39])

$$N_g(\mathcal{L}) = \operatorname{span}\left\{ \begin{bmatrix} 0 \\ Q \end{bmatrix}, \begin{bmatrix} 0 \\ xQ \end{bmatrix}, \delta^p_{p_c} \begin{bmatrix} 0 \\ |x|^2 Q \end{bmatrix}, \begin{bmatrix} \nabla Q \\ 0 \end{bmatrix}, \begin{bmatrix} Q_1 \\ 0 \end{bmatrix}, \delta^p_{p_c} \begin{bmatrix} \rho \\ 0 \end{bmatrix} \right\}. \tag{2.5}$$

In particular

$$\dim N_g(\mathcal{L}) = 2n + 2 + 2\delta^p_{p_c}.$$

The fact that the vectors on the right-hand side of (2.5) lie in $N_g(\mathcal{L})$ follows immediately from the computations (2.1)–(2.3). That these vectors span $N_g(\mathcal{L})$ is established in [39, Theorems B.2 and B.3]. These theorems rely on the nondegeneracy of the kernel of L_+.

LEMMA 2.1. *For all $n \geq 1$ and $p \in (1, p_{\max})$,*

$$N(L_+) = \operatorname{span}\{\nabla Q\}.$$

This lemma is proved in [39] for certain n and p ($n = 1$ and $1 < p < \infty$, or $n = 3$ and $1 < p \leq 3$) and is completely proved later by a general result of [19]. We present here a direct proof of this lemma, without referring to [19], relying only on oscillation properties of Sturm–Liouville ODE eigenvalue problems. A similar argument (which in the present case, however, applies only for $p \le 3$) appears in [10, Appendix C]. For completeness, we also include some arguments of [39].

A new proof. We begin with the cases $n \ge 2$. Since the potential in L_+ is radial, any solution of $L_+ v = 0$ can be decomposed as $v = \sum_{k \ge 0} \sum_{\mathbf{j} \in \Sigma_k} v_{k,\mathbf{j}}(r) Y_{k,\mathbf{j}}(\hat{x})$, where $r = |x|$, $\hat{x} = \frac{x}{r}$ is the spherical variable, and $Y_{k,\mathbf{j}}$ denote spherical harmonics: $-\Delta_{S^{n-1}} Y_{k,\mathbf{j}} = \lambda_k Y_{k,\mathbf{j}}$ (a secondary multi-index $\mathbf{j}$, appropriate to the dimension, runs over a finite set Σ_k for each k). Then $L_+ v = 0$ can be written as $A_k v_{k,\mathbf{j}} = 0$, where, for $k = 0, 1, 2, 3, \dots,$

$$A_0 = -\partial_r^2 - \frac{n-1}{r}\partial_r + 1 - pQ^{p-1}(r), \quad A_k = A_0 + \lambda_k r^{-2}, \quad \lambda_k = k(k+n-2).$$

Case 1. $k = 1$. Note that $\nabla Q = Q'(r)\hat{x}$. Since $A_1 Q' = 0$ and $Q'(r) < 0$ (monotonicity of the ground state) for $r \in (0, \infty)$, $Q'(r)$ is the unique ground state of A_1 (up to a factor), and so $A_1 \ge 0$, $A_1|_{\{Q'\}^\perp} > 0$.

Case 2. $k \ge 2$. Since $A_k = A_1 + (\lambda_k - \lambda_1) r^{-2}$ and $\lambda_k > \lambda_1$, we have $A_k > 0$, and hence $A_k v_k = 0$ has no nonzero L^2-solution.

Case 3. $k = 0$. Note that the first eigenvalue of A_0 is negative because $(Q, A_0 Q) = (Q, -(p-1)Q^p) < 0$. The second eigenvalue is nonnegative due to (2.7) and the minimax principle. Hence, if there is a nonzero solution of $A_0 v_0 = 0$, then 0 is the second eigenvalue. By Sturm–Liouville theory, $v_0(r)$ can be taken to have only one positive zero, which we denote by $r_0 > 0$. By (2.1), $A_0 Q = -(p-1)Q^p$ and $A_0 Q_1 = -2Q$. Hence $(Q^p, v_0) = 0 = (Q, v_0)$. Let $\alpha = (Q(r_0))^{p-1}$. Since $Q'(r) < 0$ for $r > 0$, the function $Q^p - \alpha Q = Q(Q^{p-1} - \alpha)$ is positive for $r < r_0$ and negative for $r > r_0$. Thus $v_0(Q^p - \alpha Q)$ does not change sign, contradicting $(v_0, Q^p - \alpha Q) = 0$. Combining all these cases gives Lemma 2.1 for $n \ge 2$.

Finally, consider $n = 1$. Suppose $L_+ v = 0$. Since L_+ preserves oddness and evenness, we may assume v is either odd or even. If it is odd, it vanishes at the origin, and so by linear ODE uniqueness, v is a multiple of Q'. So suppose v is even. As in Case 3 above, since L_+ has precisely one negative eigenvalue and has Q' in its kernel, $v(x)$ can be taken to have two zeros at $x = \pm x_0$, $x_0 \ne 0$. The argument of Case 3 above then applies on $[0, \infty)$ to yield a contradiction. □

We complete this section by summarizing some positivity estimates for the operators L_+ and L_-. These estimates are closely related to the stability/instability of the ground state.

Lemma 2.2.

$$L_- \ge 0, \quad L_-|_{\{Q\}^\perp} > 0 \quad (1 < p < p_{\max}), \tag{2.6}$$

$$(Q, L_+ Q) < 0, \quad L_+|_{\{Q^p\}^\perp} \ge 0 \quad (1 < p < p_{\max}), \tag{2.7}$$

$$L_+|_{\{Q\}^\perp} \ge 0 \quad (1 < p \le p_c), \tag{2.8}$$

$$L_+|_{\{Q, xQ\}^\perp} > 0, \quad L_-|_{\{Q_1\}^\perp} > 0 \quad (1 < p < p_c), \tag{2.9}$$

$$L_+|_{\{Q, xQ, |x|^2 Q\}^\perp} > 0, \quad L_-|_{\{Q_1, \rho\}^\perp} > 0 \quad (p = p_c). \tag{2.10}$$

Proof. Most estimates here are proved in [39] except the second part of (2.7) when $p > p_c$. It can be proved for $p \in (1, p_{\max})$ by modifying the proof of [39, Proposition 2.7] for (2.8) as follows. (It is probably also well known, but we do not know a reference.)

Recall that the ground state Q is obtained by the minimization problem (1.5). If a minimizer $Q(x)$ is rescaled so that

$$\frac{\int |\nabla Q|^2}{2a} = \frac{\int Q^2}{2b} = \frac{\int Q^{p+1}}{p+1} = \text{constant } k > 0,$$

i.e., (1.4) is satisfied, then $Q(x)$ satisfies (1.3). The minimization inequality $\frac{d^2}{d\varepsilon^2}\Big|_{\varepsilon=0} J[Q+\varepsilon\eta] \geq 0$ for all real functions η is equivalent to

$$k(\eta,\, L_+\eta) \geq \frac{1}{a}\left(\int \eta\Delta Q\right)^2 + \frac{1}{b}\left(\int Q\eta\right)^2 - \left(\int Q^p\eta\right)^2. \tag{2.11}$$

Thus $(\eta,\, L_+\eta) \geq 0$ if $\eta \perp Q^p$. Note that, if $\eta \perp Q$, by (1.3) the right-hand side of (2.11) is positive if $a \leq 1$, i.e., $p \leq p_c$. In this way, we recover (2.8). ∎

2.2. Variational formulations of the eigenvalue problem for $\mathcal{L}$. In this subsection we summarize various variational formulations for eigenvalues of $\mathcal{L}$. The generalized nullspace is given by (2.5). Suppose $\lambda \neq 0$ is a (complex) eigenvalue of $\mathcal{L}$ with corresponding eigenfunction $\left[\begin{smallmatrix} u \\ w \end{smallmatrix}\right] \in L^2$:

$$\begin{bmatrix} 0 & L_- \\ -L_+ & 0 \end{bmatrix} \begin{bmatrix} u \\ w \end{bmatrix} = \lambda \begin{bmatrix} u \\ w \end{bmatrix}. \tag{2.12}$$

The functions u and w satisfy

$$L_+ u = -\lambda w, \quad L_- w = \lambda u. \tag{2.13}$$

Therefore

$$L_- L_+ u = \mu u, \quad \mu = -\lambda^2. \tag{2.14}$$

Since $(\mu u, Q) = (L_- L_+ u, Q) = (L_+ u, L_- Q) = 0$ and $\mu \neq 0$, we have $u \perp Q$.

Denote by Π the L^2-orthogonal projection onto $Q^\perp$. We can write $L_+ u = \Pi L_+ u + \alpha Q$. Equation (2.14) implies $L_- \Pi L_+ u = \mu u$, and hence, using $u \perp Q$ and (2.6), $\Pi L_+ u = L_-^{-1} \mu u$. Thus

$$(u, Q) = 0, \quad L_+ u = \mu L_-^{-1} u + \alpha Q. \tag{2.15}$$

Since (2.14) is also implied by (2.15), these two equations are equivalent.

If $Q(x)$ is a general solution of (1.3), then $\mu = -\lambda^2$ may not be real. However, it must be real for the nonlinear ground state $Q = Q_{0,0,p}$. This fact is already known (see [29]). We will give a different proof.

Lemma. For $Q = Q_{0,0,p}$, every eigenvalue μ of (2.14) is real.

A new proof. Multiply (2.13) by $\bar{u}$ and $\bar{w}$, respectively, and integrate. Then we get

$$(u, L_+ u) = -\lambda(u, w), \quad (w, L_- w) = \lambda(w, u) = \lambda\overline{(u, w)}. \tag{2.16}$$

Taking the product, we get

$$(u, L_+ u)(w, L_- w) = -\lambda^2 |(u, w)|^2 = \mu |(u, w)|^2.$$

If $\mu \neq 0$, then w is not a multiple of Q, and so by (2.6), $(w, L_- w) > 0$. Hence $(u, w) \neq 0$ by (2.16). Thus

$$\mu = \frac{(u, L_+ u)(w, L_- w)}{|(u,w)|^2} \in \mathbb{R}. \qquad \square$$

This argument does not work when Q is an excited state, since (u, w) may be zero (see, e.g., [37, equation (2.63)]). The fact that $\mu \in \mathbb{R}$ implies that eigenvalues λ of $\mathcal{L}$ are either real or purely imaginary. Thus $\mathcal{L}$ has no complex eigenvalues with nonzero real and imaginary parts. This is not the case for excited states (see section 4 and [37]).

The proof of reality of μ in [29] uses the following formulation. For the nonlinear ground state Q, L_- is nonnegative and the operator $L_-^{1/2}$ is defined on L^2 and invertible on $Q^\perp$. A nonzero $\mu \in \mathbb{C}$ is an eigenvalue of (2.14) if and only if it is also an eigenvalue of the following problem:

$$L_-^{1/2} L_+ L_-^{1/2} g = \mu g, \tag{2.17}$$

with $g = L_-^{-1/2} u$. The operator $L_-^{1/2} L_+ L_-^{1/2}$ already appeared in [36]. Since it can be realized as a self-adjoint operator, μ must be real.

Furthermore, the eigenvalues of $L_-^{1/2} L_+ L_-^{1/2}$ can be counted using the minimax principle. Note that Q is an eigenfunction with eigenvalue 0. For easy comparison with other formulations, we formulate the principle on $Q^\perp$. Let

$$\mu_j := \inf_{g \perp Q, g_k, k=1,\dots,j-1} \frac{(g, L_-^{1/2} L_+ L_-^{1/2} g)}{(g,g)} \quad (j = 1,2,3,\dots) \tag{2.18}$$

with a suitably normalized minimizer denoted by g_j (if it exists—the definition terminates once a minimizer fails to exist). The corresponding definition for (2.15) is

$$\mu_j := \inf_{u \perp Q,\, (u, L_-^{-1} u_k)=0,\, k=1,\dots,j-1} \frac{(u, L_+ u)}{(u, L_-^{-1} u)} \quad (j = 1,2,3,\dots) \tag{2.19}$$

with a suitably normalized minimizer denoted by u_j (if it exists). In fact, the minimizer u_j satisfies

$$L_+ u_j = \mu_j L_-^{-1} u_j + \alpha_j Q + \beta_1 L_-^{-1} u_1 + \cdots + \beta_{j-1} L_-^{-1} u_{j-1} \tag{2.20}$$

for some Lagrange multipliers $\beta_1, \dots, \beta_{j-1}$. Testing (2.20) with u_k, with $k < j$, we get $(u_k, \beta_k L_-^{-1} u_k) = (u_k, L_+ u_j) = (L_+ u_k, u_j) = 0$ by (2.20) for u_k and the orthogonality conditions. Thus $\beta_k = 0$ and $L_+ u_j = \mu_j L_-^{-1} u_j + \alpha_j Q$, and hence u_j satisfies (2.15).

LEMMA 2.3. *The eigenvalues of* (2.18) *and* (2.19) *are the same, and*

$$\begin{aligned} \textit{if } 1 < p < p_c: &\quad \mu_1 = \cdots = \mu_n = 0, \quad \mu_{n+1} > 0; \\ \textit{if } p = p_c: &\quad \mu_1 = \cdots = \mu_{n+1} = 0, \quad \mu_{n+2} > 0; \\ \textit{if } p_c < p < p_{\max}: &\quad \mu_1 < 0, \quad \mu_2 = \cdots = \mu_{n+1} = 0, \quad \mu_{n+2} > 0. \end{aligned}$$

The 0*-eigenspaces are* span $L_-^{-1/2}\{\nabla Q, \delta_{p_c}^p Q_1\}$ *for* (2.18) *and* span$\{\nabla Q, \delta_{p_c}^p Q_1\}$ *for* (2.19)*, where* $\delta_{p_c}^p$ *is defined in* (2.4).

Proof. The eigenvalues of (2.18) and (2.19) are seen to be the same by taking $g = L_-^{-1/2} u$ up to a factor. By estimate (2.8), $\mu_1 \geq 0$ for $p \in (1, p_c]$. For $p \in (p_c, p_{\max})$, using (1.4), $\Pi Q_1 = Q_1 - \frac{(Q_1,Q)}{(Q,Q)}Q$, and elementary computations (such as (2.22) below), one finds that

$$(\Pi Q_1, L_+\Pi Q_1) = \frac{n^2(p-1)}{4}(p_c - p)\frac{1}{p+1}\int Q^{p+1},$$

which is negative for $p > p_c$. Thus $\mu_1 < 0$. By estimate (2.7), $\mu_2 \geq 0$ for $p \in (1, p_{\max})$.

It is clear that $u = \frac{\partial}{\partial x_j}Q$, $j = 1, \dots, n$, provides n 0-eigenfunctions. For $p = p_c$, another 0-eigenfunction is $u = Q_1$, since $Q_1 \perp Q$ (see again (2.22) below), $L_-^{-1}\nabla Q = -\frac{1}{2}xQ$, and $(Q_1, L_+Q_1) = 0$. It remains to show that $\mu_{n+1} > 0$ for $p \in (1, p_c)$ and $\mu_{n+2} > 0$ for $p \in [p_c, p_{\max})$. If $\mu_{n+2} = 0$ for $p \in (p_c, p_{\max})$, the argument after (2.20) shows the existence of a function $u_{n+2} \neq 0$ satisfying

$$L_+u_{n+2} = \alpha Q \quad \text{for some } \alpha \in \mathbb{R}, \quad u_{n+2} \perp Q,\, L_-^{-1}u_1,\, L_-^{-1}\nabla Q = -\tfrac{1}{2}xQ.$$

By Lemma 2.1, $u_{n+2} + \frac{\alpha}{2}Q_1 = c \cdot \nabla Q$ for some $c \in \mathbb{R}^d$. The orthogonality conditions imply $u_{n+2} = 0$. The cases $p \in (1, p_c]$ are proved similarly. ∎

Remark 2.4. The formulation (2.19) for μ_1 has been used for the stability problem (see, e.g., [34, equation (4.1.9), p. 73]), which can be used to prove that $\mu_1 < 0$ if and only if $p \in (p_c, p_{\max})$ by a different argument. The latter fact also follows from [39, 14] indirectly.

We summarize our previous discussion in the following.

SUMMARY 2.5. *Let $Q(x)$ be the unique positive radial ground state solution of* (1.3)*, and let $\mathcal{L}$, L_+, and L_- be as in* (1.8) *and* (1.10)*. The eigenvalue problems* (2.14), (2.15)*, and* (2.17) *for $\mu \neq 0$ are equivalent, and the eigenvalues μ must be real. These eigenvalues can be counted by either* (2.18) *or* (2.19)*. $\mu_1 < 0$ if and only if $p \in (p_c, p_{\max})$. Furthermore, all eigenvalues of $\mathcal{L}$ are purely imaginary except for an additional real pair when $p \in (p_c, p_{\max})$.*

The last statement follows from the relation $\mu = -\lambda^2$ in (2.14).

2.3. Spectrum near 0 for p near p_c. We now consider eigenvalues of $\mathcal{L}$ near 0 when p is near p_c. It was suggested by Weinstein that as p approaches p_c from below, a pair of purely imaginary eigenvalues will collide at the origin and split into a pair of real eigenvalues for $p > p_c$. In the following theorem and corollary we prove this picture rigorously and identify the leading terms of the eigenvalues and eigenfunctions.

Note that Comech and Pelinovsky [7] consider a different problem, where the equation is fixed and the varying parameter is frequency ω rather than exponent p of the nonlinearity. That problem has only $U(1)$ symmetry and no translation, but its situation is similar to ours, since we consider radial functions only in our proof. It seems one can adapt their approach to give an alternative proof. They use an abstract projection (Riesz projection) onto the discrete spectrum to reduce the problem to a 4×4 matrix problem (and exploit the complex structure), while we are more direct.

THEOREM 2.6. *There are small constants $\mu_* > 0$ and $\varepsilon_* > 0$ so that for every $p \in (p_c - \varepsilon_*, p_c + \varepsilon_*)$, there is a solution of*

$$L_+L_-w = \mu w \tag{2.21}$$

of the form

$$w = w_0 + (p-p_c)^2 g, \quad w_0 = Q + a(p-p_c)|x|^2 Q, \quad g \perp Q,$$

$$\mu = 8a(p - p_c) + (p - p_c)^2\eta, \quad a = a(p) = \frac{n(Q_1, Q^p)}{4(Q_1, x^2 Q)} < 0,$$

with $\|g\|_{L^2}$, $|\eta|$, $|a|$, *and* $1/|a|$ *uniformly bounded in* p. *Moreover, for* $p \neq p_c$, *this is the unique solution of* (2.21) *with* $0 < |\mu| \leq \mu_*$.

Proof. Set $\varepsilon := p - p_c$. Computations yield

$$(Q_1, Q) = \left(\frac{2}{p-1} - \frac{n}{2}\right)(Q, Q) = -\frac{\varepsilon n}{2(p-1)}(Q, Q), \tag{2.22}$$

$$(Q_1, Q^p) = -\frac{1}{p-1}(L_+ Q, Q_1) = -\frac{1}{p-1}(Q, L_+ Q_1) = \frac{2}{p-1}(Q, Q), \tag{2.23}$$

and

$$(Q_1, |x|^2 Q) = \left(\frac{2}{p-1} - \frac{n+2}{2}\right)(Q, |x|^2 Q) = -\left(1 + \frac{\varepsilon n}{2(p-1)}\right)(Q, |x|^2 Q). \tag{2.24}$$

Since by (2.21) with $\mu \neq 0$

$$(Q_1, w) = \mu^{-1}(Q_1, L_+ L_- w) = \mu^{-1}(L_- L_+ Q_1, w) = 0,$$

we require the leading term $(Q_1, w_0) = 0$, which decides the value of a using (2.22) and (2.24). Thus we also need $(Q_1, g) = 0$. That $a < 0$ (at least for ε sufficiently small) follows from (2.23) and (2.24).

Using the computations

$$L_- |x|^2 Q = [L_-, |x|^2] Q = -4x \cdot \nabla Q - 2nQ = -4Q_1 - \frac{2n}{p-1}\varepsilon Q \tag{2.25}$$

and

$$L_+ Q = [L_- - (p-1)Q^{p-1}]Q = -(p-1)Q^p,$$

we find that

$$L_+ L_- w_0 = a\varepsilon L_+ \left[-4Q_1 - \frac{2n}{p-1}\varepsilon Q\right] = a\varepsilon[8Q + 2n\varepsilon Q^p].$$

Thus $\mu = 8a\varepsilon + o(\varepsilon)$ and we need to solve

$$0 = [L_+ L_- - 8a\varepsilon - \varepsilon^2\eta][w_0 + \varepsilon^2 g],$$

which yields our main equation for g and η:

$$L_+ L_- g = 8a^2(|x|^2 Q) - 2an(Q^p) + \eta w_0 + (8a\varepsilon + \varepsilon^2\eta)g. \tag{2.26}$$

Recall that on radial functions (we will work only on radial functions here)

$$\ker[(L_+ L_-)^*] = \ker[L_- L_+] = \operatorname{span}\{Q_1\}.$$

Let P denote the L^2-orthogonal projection onto Q_1 and $\bar{P} := \mathbf{1} - P$. It is necessary that

$$P[8a^2(|x|^2 Q) - 2an(Q^p) + \eta w_0 + (8a\varepsilon + \varepsilon^2\eta)g] = 0$$

for (2.26) to be solvable. This solvability condition holds since $(Q_1, g) = (Q_1, w_0) = 0$, and, using the relations (2.24) and (2.23), $(Q_1, 8a^2(|x|^2 Q) - 2an(Q^p)) = 0$.

Consider the restriction (on radial functions)

$$T = L_+ L_- : \ [\ker L_-]^\perp = Q^\perp \longrightarrow Ran(\bar{P}) = Q_1^\perp.$$

Its inverse $T^{-1} = (L_-)^{-1}(L_+)^{-1}$ is bounded because $(L_+)^{-1} : Q_1^\perp \to Q^\perp$ and $(L_-)^{-1} : Q^\perp \to Q^\perp$ are bounded. So our strategy is to solve (2.26) as

$$g = T^{-1}\bar{P}[8a^2(|x|^2 Q) - 2an(Q^p) + \eta w_0 + (8a\varepsilon + \varepsilon^2\eta)g] \tag{2.27}$$

by a contraction mapping argument, with η chosen so that $(Q_1, g) = 0$. Specifically, we define a sequence $g_0 = 0$, $\eta_0 = 0$, and

$$\begin{aligned} g_{k+1} &= \bar{P}T^{-1}\bar{P}[8a^2(|x|^2 Q) - 2an(Q^p) + \eta_k w_0 + (8a\varepsilon + \varepsilon^2\eta_k)g_k], \\ \eta_{k+1} &= -\frac{1}{(Q_1, T^{-1}w_0)}\,(Q_1, T^{-1}\bar{P}[8a^2(|x|^2 Q) - 2an(Q^p) + (8a\varepsilon + \varepsilon^2\eta_k)g_k]). \end{aligned}$$

We need to check that$(Q_1, T^{-1}w_0)$ is of order one. Since $w_0 = Q + O(\varepsilon)$ and $L_+ Q_1 = -2Q$, we have $(L_+)^{-1}w_0 = -\frac{1}{2}\Pi Q_1 + O(\varepsilon)$, where Π denotes the orthogonal projection onto $Q^\perp$. Thus, using (2.25) and (2.22),

$$\begin{aligned} (Q_1, T^{-1}w_0) &= -\frac{1}{2}(Q_1, (L_-)^{-1}\Pi Q_1) + O(\varepsilon) \\ &= \frac{1}{8}(Q_1, \Pi|x|^2 Q) + O(\varepsilon) = \frac{1}{8}(Q_1, |x|^2 Q) + O(\varepsilon), \end{aligned}$$

which is of order one because of (2.24). One may then check that $N_k := \|g_{k+1} - g_k\|_{L^2} + \varepsilon^{1/2}|\eta_{k+1} - \eta_k|$ satisfies $N_{k+1} \le C\varepsilon^{1/2}N_k$, and hence (g_k, η_k) is indeed a Cauchy sequence.

Finally, the uniqueness follows from the invariance of the total dimension of generalized eigenspaces near 0 under perturbations. □

Remark 2.7. To understand heuristically the leading terms in w and μ, consider the following analogy. Let $A_\varepsilon = \left[\begin{smallmatrix} 0 & 1 \\ 0 & \varepsilon \end{smallmatrix}\right]$, which corresponds to L_+L_-. One has $A_\varepsilon \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$, $A_\varepsilon \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 1 \\ \varepsilon \end{smallmatrix}\right]$, and $A_\varepsilon \left[\begin{smallmatrix} 1 \\ \varepsilon \end{smallmatrix}\right] = \varepsilon \left[\begin{smallmatrix} 1 \\ \varepsilon \end{smallmatrix}\right]$. The vectors $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$, $\left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$, and $\left[\begin{smallmatrix} 1 \\ \varepsilon \end{smallmatrix}\right]$ correspond to Q, $|x|^2 Q$, and w, respectively.

The theorem yields an eigenvalue μ with the same sign as $p_c - p$. Since the eigenvalues of $\mathcal{L}$ are given by $\lambda = \pm\sqrt{-\mu}$, we have the following corollary.

COROLLARY 2.8. *With notation as in Theorem* 2.6*, $\mathcal{L}$ has a pair of eigenvalues $\lambda = \pm\sqrt{-\mu} = \pm\sqrt{8|a|(p - p_c) - (p - p_c)^2\eta}$ with corresponding eigenvectors $\left[\begin{smallmatrix} u \\ w \end{smallmatrix}\right]$ solving* (2.12) *and*

$$u = \lambda^{-1}L_- w = \mp\sqrt{2|a|(p - p_c)}\,Q_1 + O((p - p_c)^{3/2}).$$

When $p \in (p_c - \varepsilon_*,\ p_c)$ (stable case), λ and u are purely imaginary.

When $p \in (p_c,\ p_c + \varepsilon_*)$ (unstable case), λ and u are real.

In deriving the leading term of u we have used (2.25). We solved for w before u simply because w is larger than u.

3. One-dimensional theory. In this section we focus on the one-dimensional theory. For $n = 1$, the ground state $Q(x)$ has an explicit formula for all $p \in (1, \infty)$:

$$Q(x) = c_p \cosh^{-\beta}\left(\frac{x}{\beta}\right), \quad c_p := \left(\frac{p+1}{2}\right)^{\frac{1}{p-1}}, \quad \beta := \frac{2}{p-1}. \tag{3.1}$$

The function $Q(x)$ satisfies (1.3) and is the unique $H^1(\mathbb{R})$-solution of (1.3) up to translation and phase [5, Theorem 8.1.6, p. 259].

3.1. Eigenfunctions of L_+ and L_-. We first consider eigenvalues and eigenfunctions of L_+ and L_-. For $n = 1$,

$$L_+ = -\partial_{xx} + 1 - pQ^{p-1}, \quad L_- = -\partial_{xx} + 1 - Q^{p-1}. \tag{3.2}$$

By (3.1), these operators are both of the form

$$-\partial_{xx} + 1 - C\,\mathrm{sech}^2(x/\beta).$$

Such operators have essential spectrum $[1, \infty)$ and finitely many eigenvalues below 1. A lot of information about such operators is available in the classical book [35, p. 103]:

- all eigenvalues are simple and can be computed explicitly as zeros and poles of an explicit meromorphic function;
- all eigenfunctions can be expressed in terms of the hypergeometric function.

We begin by presenting another way to derive the eigenvalues as well as different formulas for the eigenfunctions. We will not prove right here that this set contains all of the eigenvalues/eigenfunctions. This fact is a consequence of the more general theorem, Theorem 3.4, proved later (and see also [35]).

Define

$$\begin{aligned} &\lambda_m := 1 - k_m^2, \quad k_m := \frac{p+1}{2} - \frac{m(p-1)}{2}, \\ &p_m := \frac{m+1}{m-1} \quad \text{for } m > 1, \quad p_1 = \infty. \end{aligned} \tag{3.3}$$

The following theorem agrees with the numerical observation Figure 1.

Theorem 3.1. *For $n = 1$ and $1 < p < \infty$, let $Q(x)$ be defined by (3.1), L_+ and L_- be defined by (3.2), and λ_m, k_m, p_m be defined by (3.3). Suppose for $M \in \mathbb{Z}^+$,*

$$p_{M+1} \le p < p_M. \tag{3.4}$$

Then the operator L_+ has eigenvalues λ_m, $0 \le m \le M$, with eigenfunctions of the form

$$\varphi_{2\ell} = \sum_{j=0}^{\ell} c_{2j}^{2\ell} Q^{k_{2j}}, \quad \varphi_{2\ell-1} = \sum_{j=1}^{\ell} c_{2j-1}^{2\ell-1} (Q^{k_{2j-1}})_x,$$

and the operator L_- has eigenvalues λ_m, $1 \le m \le M$, with eigenfunctions of the form

$$\psi_{2\ell-1} = \sum_{j=1}^{\ell} d_{2j-1}^{2\ell-1} Q^{k_{2j-1}}, \quad \psi_{2\ell} = \sum_{j=1}^{\ell} d_{2j}^{2\ell} (Q^{k_{2j}})_x.$$

In particular, all eigenvalues of L_- are eigenvalues of L_+, and L_+ always has one more eigenvalue ($\lambda_0 < 0$) than L_-.

Proof. It can be proved by induction, using

$$Q^{p-1} = \frac{p+1}{2}\cosh^{-2}\left(\frac{x}{\beta}\right), \quad Q_x = -Q\tanh\left(\frac{x}{\beta}\right), \quad Q_x^2 = Q^2\left(1 - \frac{2}{p+1}Q^{p-1}\right),$$

and

$$Q^{-k}L_+Q^k = \frac{(k+p)\,(2k-p-1)}{p+1}\,Q^{p-1} + (1-k^2), \tag{3.5}$$

$$[(Q^k)_x]^{-1}L_+(Q^k)_x = \frac{(k-1)(2k+3p-1)}{p+1}\,Q^{p-1} + (1-k^2),$$

$$Q^{-k}L_-Q^k = \frac{(k-1)(2k+p+1)}{p+1}\,Q^{p-1} + (1-k^2), \tag{3.6}$$

$$[(Q^k)_x]^{-1}L_-(Q^k)_x = \frac{(k+p)(2k+p-3)}{p+1}\,Q^{p-1} + (1-k^2).$$

The coefficients of Q^{p-1} vanish when $k = \frac{p+1}{2}, 1, 1, \frac{3-p}{2}$, respectively. It is why the highest power of Q is $Q^{\frac{p+1}{2}}$ in $\varphi_{2\ell}$, Q_x in $\phi_{2\ell-1}$, Q in $\psi_{2\ell-1}$, and $(Q^{\frac{3-p}{2}})_x$ in $\psi_{2\ell}$. ◻

3.2. Connection between L_+ and L_- and their factorizations. In light of Theorem 3.1, it is natural to ask *why* all eigenvalues of L_- are also eigenvalues of L_+. Is there a simple connection between their eigenfunctions? In this section we prove this is indeed so.

We first look for an operator U of the form

$$U = \partial_x + R(x) \quad (\text{ so } U^* = -\partial_x + R(x))$$

such that

$$L_-U = UL_+ \quad (\text{ so } U^*L_- = L_+U^*). \tag{3.7}$$

It turns out that there is a unique choice of $R(x)$:

$$R(x) = -\frac{p+1}{2}\frac{Q_x}{Q} = \frac{p+1}{2}\tanh\left(\frac{(p-1)x}{2}\right).$$

In fact, with this choice of $R(x)$,

$$U = \varphi_0\partial_x\varphi_0^{-1} \quad (\text{ so } U^* = -\varphi_0^{-1}\partial_x\varphi_0), \tag{3.8}$$

where $\varphi_0 = Q^{\frac{p+1}{2}}$ is the ground state of L_+ and is considered here as a multiplication operator: $Uf = \varphi_0\partial_x(\varphi_0^{-1}f)$.

Suppose now ψ is an eigenfunction of L_- with eigenvalue λ: $L_-\psi = \lambda\psi$. By (3.7),

$$0 = U^*(L_- - \lambda)\psi = (L_+ - \lambda)U^*\psi.$$

Thus $U^*\psi$ is an eigenfunction of L_+ with the same eigenvalue λ (provided $U^*\psi \in L^2$). Therefore the map

$$\psi \mapsto U^*\psi$$

sends an eigenfunction of L_- to an eigenfunction of L_+ with the same eigenvalue. This map is not onto because U^* is not invertible. Specifically, the ground state φ_0 is not in the range. In fact, $U\varphi_0 = \varphi_0\partial_x\varphi_0^{-1}\varphi_0 = 0$. If $\varphi_0 = U^*\psi$, then $(\varphi_0,\varphi_0) = (\varphi_0, U^*\psi) = (U\varphi_0, \psi) = 0$, a contradiction. We summarize our finding as the following proposition.

PROPOSITION 3.2. *Under the same assumptions and notation as Theorem* 3.1, *the eigenfunctions* φ_m *and* ψ_m *of* L_+ *and* L_- *satisfy*

$$\varphi_m = U^* \psi_m \quad (m = 1, \dots, M)$$

up to constant factors. Note that U^* *sends even functions to odd functions and vice versa.*

Proof. We need only verify that $U^*\psi_m \in L^2$. This is the case, since $U^* = -\partial_x + \frac{p+1}{2}\tanh(x/\beta)$, $\psi_m(x)$ are sums of powers of Q and Q_x and since $\tanh(x/\beta)$, Q_x/Q, and Q_{xx}/Q_x are bounded. ◻

Analogous to the definition of U, we define

$$S := Q\partial_x Q^{-1} = \partial_x - \frac{Q_x}{Q} \quad (\text{ so } S^* = -Q^{-1}\partial_x Q). \tag{3.9}$$

Clearly $SQ = 0$. Recall that λ_0 is the first eigenvalue of L_+ with eigenfunction φ_0. Hence $L_+ - \lambda_0$ is a nonnegative operator. In fact we have the following factorizations.

LEMMA 3.3. *Let* U *and* S *be defined by* (3.8) *and* (3.9), *respectively. One has*

$$L_+ - \lambda_0 = U^*U, \quad L_- - \lambda_0 = UU^*. \tag{3.10}$$

$$L_- = S^*S, \quad SS^* = -\partial_x^2 + 1 + \frac{p-3}{p+1}Q^{p-1}. \tag{3.11}$$

Moreover, $SS^* = L_- + \frac{2(p-1)}{p+1}Q^{p-1} > 0$.

The formula $L_- = S^*S$ was known; see, e.g., [34, equation (4.1.8), p. 73]. It is an example of the Darboux transformations; see, e.g., [23]. Factorization of Schrödinger operators into first-order operators has been known since the times of Darboux (1840s).

3.3. Hierarchy of operators. In this subsection we generalize Theorem 3.1 and Lemma 3.3 to a family of operators containing L_+ and L_-. As a reminder, we have

$$\begin{aligned} &\frac{Q''}{Q} = 1 - Q^{p-1}, \quad \left(\frac{Q'}{Q}\right)^2 = 1 - \frac{2}{p+1}Q^{p-1}, \\ &\left(\frac{Q'}{Q}\right)' = \frac{Q''}{Q} - \left(\frac{Q'}{Q}\right)^2 = -\frac{p-1}{p+1}Q^{p-1}. \end{aligned} \tag{3.12}$$

Let $S(a) := Q^a\partial_x Q^{-a}$. We have

$$\begin{aligned} &S(a) = \partial_x - \frac{aQ'}{Q}, \quad S(a)^* = -\partial_x - \frac{aQ'}{Q}, \\ &S(a)^*S(a) = -\partial_x^2 + a^2 - a\left\{a + \frac{p-1}{2}\right\}\frac{2}{p+1}Q^{p-1}. \end{aligned} \tag{3.13}$$

Define the following hierarchy of operators:

$$\begin{aligned} &S_j := S(k_j), \quad \text{where recall} \quad k_j = 1 - (j-1)\frac{p-1}{2}, \\ &L_j := S_{j-1}S_{j-1}^* + \lambda_{j-1} = S_j^*S_j + \lambda_j, \quad \text{where recall} \quad \lambda_j = 1 - k_j^2. \end{aligned} \tag{3.14}$$

Then we have

$$\begin{aligned} &S_0 = U, \quad S_1 = S, \dots, \\ &L_0 = L_+, \quad L_1 = L_-, \quad L_2 = SS^*, \dots, \\ &S_j L_j = L_{j+1} S_j, \quad L_j S_j^* = S_j^* L_{j+1}. \end{aligned} \tag{3.15}$$

More explicitly,

$$L_j = -\partial_x^2 + 1 - k_{j-1} k_j \frac{2}{p+1} Q^{p-1}. \tag{3.16}$$

Note that j here can be any real number.

Recall the definition $p_j := 1 + 2/(j-1)$ for $j > 1$, and set $p_j = \infty$ for $j \le 1$. Then p_j is a monotone decreasing function of j, $k_j > 0$ for $p < p_j$, $k_j = 0$ for $p = p_j$, and $k_j < 0$ for $p > p_j$. Let

$$\lambda_j' := \begin{cases} \lambda_j & (1 < p \le p_j), \\ 1 & (p_j < p < p_{j-1}), \\ \lambda_{j-1} & (p_{j-1} \le p). \end{cases} \tag{3.17}$$

By the second identity of (3.14), and (3.16) together with the fact $k_{j-1}k_j < 0$ for $p_j < p < p_{j-1}$, we have the lower bound

$$L_j \ge \lambda_j'. \tag{3.18}$$

In fact, this estimate is sharp: for $p \in (1, p_j) \cup (p_{j-1}, \infty)$, the ground state is obvious from the second identity of (3.14):

$$\begin{cases} L_j Q_j = \lambda_j Q_j & (1 < p < p_j), \\ L_j Q_{j-1}^* = \lambda_{j-1} Q_{j-1}^* & (p_{j-1} < p), \end{cases} \tag{3.19}$$

where we denote

$$Q_j := Q^{k_j}, \quad Q_j^* := Q^{-k_j}. \tag{3.20}$$

For $p \in [p_j, p_{j-1}]$, there is no ground state. Thus we have completely determined the ground state of L_j for all $p > 1$. The complete spectrum, together with explicit eigenfunctions, is derived using the third identity of (3.15) as follows.

THEOREM 3.4. *For any $j \in \mathbb{R}$ and $p > 1$, the point spectrum of L_j consists of simple eigenvalues*

$$\begin{aligned} spec_p(L_j) = &\{\lambda_k \mid p < p_k,\ k \in \{j, j+1, j+2, \dots\}\} \\ &\cup \{\lambda_k \mid p > p_k,\ k \in \{j-1, j-2, j-3, \dots\}\}, \end{aligned} \tag{3.21}$$

and the eigenfunction for the eigenvalue λ_k is given uniquely up to constant multiple by

$$\begin{cases} S_j^* \cdots S_{k-1}^* Q_k & (k \in \{j, j+1, \dots\}), \\ S_{j-1} \cdots S_{k+1} Q_k^* & (k \in \{j-1, j-2, \dots\}), \end{cases} \tag{3.22}$$

each of which is a linear combination of

$$\begin{cases} Q_j, Q_{j+2}, \dots Q_k & (k \in \{j, j+2, \dots\}), \\ Q_{j+1}R, Q_{j+3}R, \dots Q_k R & (k \in \{j+1, j+3, \dots\}), \\ Q^*_{j-1}, Q^*_{j-3}, \dots Q^*_k & (k \in \{j-1, j-3, \dots\}), \\ Q^*_{j-2}R, Q^*_{j-4}R, \dots Q^*_k R & (k \in \{j-2, j-4, \dots\}), \end{cases} \tag{3.23}$$

where $R := Q'/Q$.

Proof. The ground states have been determined. The third identity of (3.15) implies that (3.22) belongs to the eigenspace of L_j with eigenvalue λ_k. Moreover, each function is nonzero because S_k^* is injective for $p < p_k$, and so is S_k for $p_k < p$. Since S_j annihilates only the ground state Q_j for $p < p_j$ and S^*_{j-1} annihilates only the ground state Q^*_{j-1} for $p > p_j$, all the excited states of L_j for $p < p_j$ are mapped injectively by S_j to bound states of L_{j+1} and for $p > p_j$ by S^*_{j-1} to those of L_{j-1}. Hence we have (3.21), and all the eigenvalues are simple because so too are the ground states. Equation (3.23) follows from the fact that S_j and S_j^* act on Q^a like $C(a,j)R$, while S_jS_{j-1} and $S^*_{j-1}S^*_j$ act on Q^a like $C_1(a,j) + C_2(a,j)Q^{p-1}$. ∎

3.4. Mirror conjugate identity. The following remarkable identity has application to estimating eigenvalues of $\mathcal{L}$ (see section 3.6):

$$S_j(L_{j-1} - \lambda_j)S_j^* = S_j^*(L_{j+2} - \lambda_j)S_j. \tag{3.24}$$

To prove this, start with the formula

$$\begin{aligned}(\partial_x + R)(\partial_x^2 + V)(\partial_x - R) &= \partial_x^4 + (-3R' - R^2 + V)\partial_x^2 + (-3R' - R^2 + V)'\partial_x \\ &\qquad - R''' - (VR)' - RR'' - R^2V,\end{aligned} \tag{3.25}$$

which implies that $(\partial_x + R)(\partial_x^2 + V_+)(\partial_x - R) = (\partial_x - R)(\partial_x^2 + V_-)(\partial_x + R)$ is equivalent to

$$V_\pm = -R''/R \pm 3R' - R^2 + C/R. \tag{3.26}$$

Now set $R := aQ'/Q$. Plugging the identities

$$\begin{aligned} &R^2 = a^2\left(1 - \frac{2}{p+1}Q^{p-1}\right), \quad R' = -a\frac{p-1}{p+1}Q^{p-1}, \\ &\frac{R''}{R} = -\frac{(p-1)^2}{p+1}Q^{p-1} \end{aligned} \tag{3.27}$$

into (3.26), we get, for $C = 0$,

$$V_\pm = -a^2 + \frac{2}{p+1}(a \pm (p-1))\left(a \pm \frac{(p-1)}{2}\right)Q^{p-1}. \tag{3.28}$$

Hence for $a = k_j$ we have

$$V_\pm = -k_j^2 + \frac{2}{p+1}k_{j\pm2}k_{j\pm1}, \tag{3.29}$$

which gives the desired identity (3.24). The above proof also shows that L_{j-1} and L_{j+2} are the unique choices for the identity to hold with S_j (modulo a constant multiple of Q/Q_x, which is singular).

3.5. Variational formulations for eigenvalues of $\mathcal{L}$. We considered two variational formulations for nonzero eigenvalues of $\mathcal{L}$ in general dimensions in section 2.2. Here we present a new variational formulation for one dimension. Define the self-adjoint operator

$$H := SL_+S^*. \tag{3.30}$$

This is a fourth-order differential operator with essential spectrum $[1, \infty)$. By a direct check, we have

$$HQ = SL_+S^*Q = SL_+(-2Q_x) = 0.$$

Thus Q is an eigenfunction with eigenvalue 0. Since $(Q, S^*f) = (SQ, f) = 0$ for any f, we have

$$\text{Range } S^* \perp Q. \tag{3.31}$$

In particular, since $L_+|_{Q^\perp}$ is nonnegative for $p \le 5$ by Lemma 2.2, so is H.

LEMMA 3.5. *The nullspace of H is*

$$N(H) = \text{span}\left\{Q, \delta^p_{p_c} xQ\right\},$$

where, recall, $\delta^p_{p_c}$ is 0 if $p \neq p_c$ and 1 if $p = p_c$.

Remark. Note that $\dim N(H) = 1+\delta^p_{p_c}$, which is different from $\dim N(L_-^{1/2}L_+L_-^{1/2}) = 2 + \delta^p_{p_c}$. We will show below that H and $L_-^{1/2}L_+L_-^{1/2}$ have the same *nonzero* eigenvalues.

Proof. If $Hf = 0$, then $L_+S^*f = -2cQ$ and $S^*f = cQ_1 + dQ_x$ for some $c, d \in \mathbb{R}$ by Lemma 2.1. We have $Q_1 \perp Q$ if and only if $p = p_c = 5$. Thus, if $p \neq 5$, $c = 0$ by (3.31), and $S^*(f + \frac{d}{2}Q) = 0$. We conclude that $f = -\frac{d}{2}Q$.

When $p = 5$, we have $S^*xQ = -Q^{-1}\partial_x(QxQ) = -2Q_1$. Thus $S^*(f+\frac{c}{2}Q_1+\frac{d}{2}Q) = 0$ and $f = -\frac{c}{2}Q_1 - \frac{d}{2}Q$. $\square$

Define eigenvalues of H as follows:

$$\tilde{\mu}_j := \inf_{f\perp f_k, k<j} \frac{(f, Hf)}{(f, f)}, \qquad (j = 1, 2, 3, \dots) \tag{3.32}$$

with a suitably normalized minimizer denoted by f_j if it exists. By standard variational arguments, if $\tilde{\mu}_j < 1$, then a minimizer f_j exists. By convention, if μ_k is the first of the μ_j's to hit 1 (and so f_k may not be defined), we set $\mu_j := 1$ for all $j > k$.

We can expand Summary 2.5 to the following.

THEOREM 3.6 (equivalence). *Let $n = 1$. Let μ_j be defined as in Summary 2.5 and $\tilde{\mu}_j$ be defined by (3.32). Then $\mu_j = \tilde{\mu}_j$. When $\mu_j \neq 0$ and $\mu_j < 1$, the eigenfunctions of (2.19) and (3.32) can be chosen to satisfy*

$$u_j = S^*f_j, \quad f_j = \frac{1}{\mu_j}SL_+u_j.$$

Proof. First, we establish the equivalence of nonzero eigenvalues. Suppose $f = f_j$ is an eigenfunction of (3.32) with eigenvalue $\tilde{\mu} \neq 0$; then $SL_+S^*f = \tilde{\mu}f$. Let $u := S^*f \neq 0$, and apply S^* on both sides. By $L_- = S^*S$ we get $L_-L_+u = \tilde{\mu}u$. Thus u is an eigenfunction satisfying (2.14) with $\mu = \tilde{\mu}$. On the other hand, suppose u satisfies

$L_-L_+u = \mu u$ with $\mu \neq 0$. Applying SL_+ on both sides and using $L_- = S^*S$, we get $SL_+S^*SL_+u = \mu SL_+u$, i.e., $Hf = \mu f$ for $f = \mu^{-1}SL_+u$.

Now use Lemmas 2.3 and 3.5. If $p \in (1,5)$, then $\mu_1 = \tilde{\mu}_1 = 0$, corresponding to Q_x and Q, and $\mu_2 = \tilde{\mu}_2 > 0$. If $p = 5$, then $\mu_1 = \mu_2 = \tilde{\mu}_1 = \tilde{\mu}_2 = 0$, corresponding to Q_x, Q_1, and Q, xQ, and $\mu_3 = \tilde{\mu}_3 = 1$. If $p \in (5,\infty)$, then $\mu_1 = \tilde{\mu}_1 < 0$, $\mu_2 = \tilde{\mu}_2 = 0$, corresponding to Q_x and Q, and $\mu_3 = \mu_3 = 1$. We have shown that $\tilde{\mu}_j = \mu_j$. $\square$

In the following we will make no distinction between μ_j and $\tilde{\mu}_j$. By the minimax principle, (3.32) has the following equivalent formulations:

$$\mu_j = \inf_{\dim M = j} \sup_{f \in M} \frac{(f,Hf)}{(f,f)} = \sup_{\dim M = j-1} \inf_{f \perp M} \frac{(f,Hf)}{(f,f)}. \tag{3.33}$$

Here M runs over all linear subspaces of $L^2(\mathbb{R})$ with the specified dimension.

3.6. Estimates of eigenvalues of $\mathcal{L}$. In this subsection we prove lower and upper bounds for eigenvalues of $\mathcal{L}$, confirming some aspects of the numerical computations shown in Figures 1 and 6. Recall that, by Lemma 2.3, the first positive μ_j is μ_2 for $p \in (1, p_c)$ and μ_3 for $p \in [p_c, p_{\max})$. The first theorem concerns upper bounds for μ_1 and μ_2.

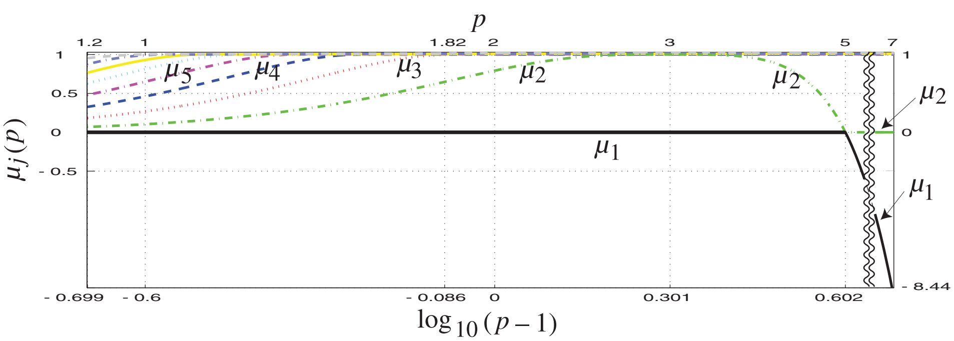


FIG. 6. *p vs. μ_j.*

THEOREM 3.7. *Suppose $n = 1$ and $1 < p < \infty$.*

(a) *If $p \neq 3$, then $\mu_2 \le C_p$ for some explicitly computable $C_p < 1$. In particular, f_2 exists.*

(b) *$\mu_1 < 0$ if and only if $p > 5$. For any $C > 0$, we have $\mu_1(p) \le -Cp^3$ for p sufficiently large.*

Proof. For part (a), we already know $\mu_2 = 0$ for $p \ge 5$. Assume $p \in (1,5)$. Consider test functions of the form $f = SQ^k$ with $k > 0$. f is odd, and hence $f \perp Q$, the 0-eigenfunction of H. Since $H = SL_+S^*$ and $S^*S = L_-$, we have

$$\mu_2 \le \frac{(f,Hf)}{(f,f)} = \frac{(L_-Q^k, L_+L_-Q^k)}{(Q^k, L_-Q^k)}.$$

By formulas (3.5) and (3.6),

$$L_-Q^k = aQ^{k+p-1} + bQ^k, \quad a = \frac{1}{p+1}(k-1)(2k+p+1), \quad b = 1 - k^2,$$

$$L_+Q^{k+p-1} = \sigma Q^{k+2p-2} + dQ^{k+p-1},$$

$$\sigma = \frac{1}{p+1}(k+2p-1)(2k+p-3), \quad d = 1-(k+p-1)^2,$$

$$L_+Q^k = cQ^{k+p-1} + bQ^k, \quad c = \frac{1}{p+1}(k+p)(2k-p-1).$$

Thus

$$\frac{(f,Hf)}{(f,f)} = \frac{a^2\sigma J_3 + a(ad+bc+b\sigma)J_2 + b(ad+ab+bc)J_1 + b^3 J_0}{aJ_1 + bJ_0}, \tag{3.34}$$

where

$$J_m = \int_{\mathbb{R}} Q^{2k+m(p-1)}(x)\,dx \quad (m = 0,1,2,3),$$

which are always positive. If $k \to 0^+$, then J_m converges to $\int_{\mathbb{R}} Q^{m(p-1)}\,dx$ for $m > 0$, and $J_0 = O(k^{-1})$. The above quotient can be written as

$$(3.34) = b^2 + \frac{J}{aJ_1 + bJ_0},$$

where

$$J = a^2\sigma J_3 + a(ad+bc+b\sigma)J_2 + b(ad+bc)J_1.$$

Note that $J_m|_{k=0} = (\frac{p+1}{2})^m \frac{2}{p-1} \int_{\mathbb{R}} \operatorname{sech}^{2m}(y)\,dy$ with $\int_{\mathbb{R}} \operatorname{sech}^{2m}(y)\,dy = 2, \frac{4}{3}, \frac{16}{15}$ for $m = 1,2,3$, respectively. Also, as $k \to 0^+$, $a \to -1$, $b \to 1$, $c \to -p$, $\sigma \to \frac{(2p-1)(p-3)}{p+1}$, and $d \to 1-(p-1)^2$. Direct calculation shows that

$$\lim_{k\to 0^+} J = -\frac{2}{15(p-1)}\,(p+1)^2(p-3)^2.$$

Also note that $b^2 < 1$ for $k > 0$. Thus, if $1 < p < \infty$ and $p \neq 3$, then $J < 0$ and the quotient (3.34) is less than 1 for k sufficiently small. (If $p = 3$, the sign of J is unclear and (3.34) may not be less than 1.) This proves $\mu_2 < 1$ and provides an upper bound less than 1 for μ_2. It also implies the existence of f_2. This establishes statement (a).

For statement (b), the fact that $\mu_1 < 0$ if and only if $p > 5$ is part of Lemma 2.3. We now consider the behavior of μ_1 for p large. Fix $k > 1$ to be chosen later. As $p \to \infty$,

$$J_m = \left(\frac{p+1}{2}\right)^{\frac{2k}{p-1}+m} \cdot \frac{2}{p-1} \cdot \int_{\mathbb{R}} (\operatorname{sech} x)^{\frac{4k}{p-1}+2m}\,dx \sim C_m p^{m-1},$$

with $C_m = 2^{1-m}\int_{\mathbb{R}}(\operatorname{sech} x)^{2m}dx = 2, \frac{2}{3}, \frac{4}{15}$ for $m = 1,2,3$, respectively, and

$$a \sim k-1, \quad b = 1-k^2, \quad c \sim -p, \quad \sigma \sim 2p, \quad d \sim -p^2.$$

Thus, by (3.34),

$$\frac{(f,Hf)}{(f,f)} \sim \frac{a\sigma J_3 + adJ_2}{J_1} \sim \frac{1-k}{15}\,p^3 \quad \text{as } p \to \infty.$$

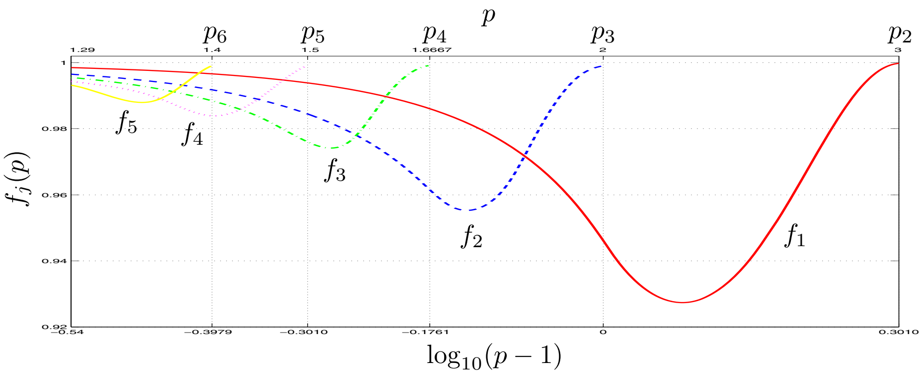


FIG. 7. *p vs. f_j for $j = 1, \ldots, 5$, where $f_j(p) = \frac{\lambda_{j+1}\lambda_{j+2}}{\mu_{j+1}}$ for $1 < p < p_{j+2}$ and $f_j(p) = \frac{\lambda_{j+1}}{\mu_{j+1}}$ for $p_{j+2} \le p < p_{j+1}$.*

By choosing $k > 1$ sufficiently large, we have shown that for any C, $\mu_1 \le -Cp^3$ for p sufficiently large. □

The next theorem bounds eigenvalues of $\mathcal{L}$ by eigenvalues of L_+ and L_-. Recall that p_j and $\lambda_j(p)$ are defined in (3.4) and (3.3).

THEOREM 3.8 (interlacing of eigenvalues). *Fix $k \ge 1$ and $p \in [p_{k+2},\ p_{k+1})$, where, recall, $p_j = \frac{j+1}{j-1}$. Let $\lambda_j(p) = 1 - \frac{1}{4}[(p+1) - j(p-1)]^2$ be as in (3.3), and so $\lambda_{k+1} < 1 \le \lambda_{k+2}$. For the eigenvalues μ_j defined by (3.32), we have*

$$\lambda_{j+1}^2(p) < \mu_{j+1}(p) < \lambda_{j+2}^2(p) \quad (1 \le j < k); \qquad \lambda_{k+1}^2(p) < \mu_{k+1}(p) \le 1. \tag{3.35}$$

In particular, there are K simple eigenvalues $\mu_2, \ldots, \mu_{K+1}$ in $(0, 1)$, where $K = k$ if $\mu_{k+1} < 1$ and $K = k - 1$ if $\mu_{k+1} = 1$. Moreover, K is always 1 when $k = 1$. Finally,

$$\mu_2 \ge \begin{cases} \lambda_2\lambda_3 & (1 < p \le 2), \\ \lambda_2 & (2 < p < 5), \end{cases} \qquad \mu_3 \ge \begin{cases} \lambda_3\lambda_4 & (1 < p \le 5/3), \\ \lambda_3 & (5/3 < p \le 2), \\ 1 & (2 < p < \infty), \end{cases}$$

$$\mu_1 \ge -\frac{1}{16}(p-1)^3(p-5) \quad (5 \le p < \infty).$$

Remark 3.9. In view of the above lower bounds for μ_2 and μ_3, we conjecture that

$$\mu_{j+1} \ge \lambda_{j+1}\lambda_{j+2} \quad (1 < p < p_{j+2}); \qquad \mu_{j+1} \ge \lambda_{j+1} \quad (p_{j+2} \le p < p_{j+1}). \tag{3.36}$$

This is further confirmed numerically for $j = 3, 4, 5$ (see Figure 7). Note that $\lim_{p \to p_{j+1}-} \frac{\lambda_{j+1}}{\mu_{j+1}} = 1$ because both λ_{j+1} and μ_{j+1} converge to 1. It also seems that $\frac{\lambda_{j+1}\lambda_{j+2}}{\mu_{j+1}}$ has a limit as $p \to 1+$, but it is not clear even though we have (3.35) and $\lambda_j = (j-1)(p-1) + O((p-1)^2)$ as $p \to 1+$.

Proof. We first prove the upper bound: For $j < k$, use the test functions

$$S\psi_2,\ S\psi_3, \ldots,\ S\psi_{j+2}$$

(we cannot use $S\psi_1$, since it is zero). Recall $L_-\psi_m = \lambda_m\psi_m$. Let $a = (a_2, \ldots, a_{j+2})$ vary over $\mathbb{C}^{j+1} - \{0\}$. By equivalent definition (3.33), $H = SL_+S^*$, $L_- = S^*S$, and the orthogonality between the ψ_m's, we have

$$\begin{aligned}\mu_{j+1} &\le \sup_a \frac{(\sum_m a_m S\psi_m, H\sum_\ell a_\ell S\psi_\ell)}{(\sum_m a_m S\psi_m, \sum_\ell a_\ell S\psi_\ell)} = \sup_a \frac{(\sum_m a_m\psi_m, L_-L_+L_-\sum_\ell a_\ell\psi_\ell)}{(\sum_m a_m\psi_m, L_-\sum_\ell a_\ell\psi_\ell)} \\ &\le \sup_a \frac{(\sum_m a_m\psi_m, L_-L_-L_-\sum_\ell a_\ell\psi_\ell)}{(\sum_m a_m\psi_m, L_-\sum_\ell a_\ell\psi_\ell)} = \sup_a \frac{\sum_m |a_m|^2\lambda_m^3}{\sum_m |a_m|^2\lambda_m} \\ &\le \max_{m=2,\dots,j+2} \lambda_m^2 = \lambda_{j+2}^2.\end{aligned}$$

Since $\mu_{j+1} \le \lambda_{j+2}^2 < 1$, it is attained at some function for which the second inequality above cannot be replaced by an equality sign. Thus $\mu_{j+1} < \lambda_{j+2}^2$.

For the lower bound of eigenvalues, we use only the special case $j = 1$ of (3.24):

$$H = SL_+S^* = SL_0S^* = S^*L_3S. \tag{3.37}$$

In particular, we have for $1 < p < 3$,

$$H \ge S^*L_2S = S^*SS^*S = L_1^2 = L_-^2, \tag{3.38}$$

which implies that

$$\lambda_{j+1}^2 \le \mu_{j+1} \quad (1 < p < 3) \tag{3.39}$$

(and again, equality is impossible).

For the second eigenvalue μ_2, we can get a more precise estimate by using (3.18) for $L_3 \ge \lambda_3'$ together with

$$L_1|_{Q^\perp} \ge \lambda_2', \tag{3.40}$$

which follows from *spec*(L_1). Combining these estimates, we have for any $f \perp Q$ and $p < 5$,

$$(Hf, f) \ge \lambda_3'(Sf, Sf) \ge \lambda_3'\lambda_2'(f, f), \tag{3.41}$$

which implies that $\mu_2 \ge \lambda_3'\lambda_2'$, i.e.,

$$\mu_2 \ge \begin{cases}\lambda_2\lambda_3 & (1 < p \le 2), \\ \lambda_2 & (2 < p < 5).\end{cases} \tag{3.42}$$

For $p > 3$, we have $L_3 \ge \lambda_2 = -(p-1)(p-5)/4$ and

$$L_3 - L_2 \ge -(p-1)(p-3)/2 =: -a. \tag{3.43}$$

Hence for any $t \in [0, 1]$, we have

$$L_3 \ge tL_2 - at + (1-t)\lambda_2, \tag{3.44}$$

and so for $b > 0$, we have

$$\begin{aligned}(Hf, f) + b(f, f) &\ge (S^*(tL_2 - at + (1-t)\lambda_2)Sf, f) + b(f, f) \\ &= t\|L_1f\|^2 - (at - (1-t)\lambda_2)(L_1f, f) + b\|f\|^2,\end{aligned} \tag{3.45}$$

which is nonnegative if

$$b \geq (at - (1-t)\lambda_2)^2/(4t), \tag{3.46}$$

whose infimum is attained at $t = -\lambda_2/(a+\lambda_2) = (p-5)/(p-1) \in (0,1)$ for $p > 5$. Plugging this back in, we obtain the lower bound

$$\mu_1 \geq \lambda_2(a+\lambda_2) = -\frac{1}{16}(p-1)^3(p-5) \quad (p > 5). \tag{3.47}$$

We have a similar bound on μ_3 by using the even-odd decomposition $L^2(\mathbb{R}) = L^2_{ev}(\mathbb{R}) \oplus L^2_{od}(\mathbb{R})$. Let ψ_j, ξ_j be the eigenfunction of L_1 and L_3 such that

$$L_1\psi_j = \lambda_j\psi_j, \quad L_3\xi_j = \lambda_j\xi_j. \tag{3.48}$$

ψ_j starts from $j = 1$, and ξ_j starts with $j = 3$. They are even for odd j and odd for even j. For any even function $f \perp Q = \psi_1$, Sf is odd, and so we have $f \perp \psi_1 = Q, \psi_2$ and $Sf \perp \xi_3$. Hence by *spec*(L_3) and *spec*(L_1), we have

$$(Hf,f) = (L_3Sf, Sf) \geq \tilde{\lambda}_4(Sf,Sf) = \tilde{\lambda}_4(L_1f,f) \geq \tilde{\lambda}_4\tilde{\lambda}_3(f,f), \tag{3.49}$$

where we denote

$$\tilde{\lambda}_j := \begin{cases} \lambda_j & (1 < p < p_j), \\ 1 & (p_j < p). \end{cases} \tag{3.50}$$

Thus the second eigenvalue of H on L^2_{ev} is $\geq \tilde{\lambda}_4\tilde{\lambda}_3$. Next for any odd function $f \perp \psi_2$, we have $f \perp \psi_1, \psi_2, \psi_3$. Hence we have

$$(Hf,f) = (L_3Sf, Sf) \geq \lambda_3'(Sf,Sf) = \lambda_3'(L_1f,f) \geq \lambda_3'\tilde{\lambda}_4(f,f). \tag{3.51}$$

Similarly, every odd function $f \perp S^*\xi_3$ satisfies $f \perp \psi_1$ and $Sf \perp \xi_3, \xi_4$, and so

$$(Hf,f) \geq \tilde{\lambda}_5\tilde{\lambda}_2(f,f). \tag{3.52}$$

Hence the second eigenfunction on L^2_{od} is $\geq \max(\tilde{\lambda}_4\lambda_3', \tilde{\lambda}_5\tilde{\lambda}_2) \geq \tilde{\lambda}_4\tilde{\lambda}_3$. Therefore we have $\mu_3 \geq \tilde{\lambda}_3\tilde{\lambda}_4$, i.e.,

$$\mu_3 \geq \begin{cases} \lambda_3\lambda_4 & (1 < p < 5/3), \\ \lambda_3 & (5/3 < p < 2), \\ 1 & (2 < p). \end{cases} \tag{3.53}$$

This argument, however, does not yield any useful estimates for the higher μ_j. □

3.7. Resonance for $p = 3$. In the theory of dispersive estimates for the linear Schrödinger evolution, it is important to know whether or not the endpoints of the continuous spectrum of the linear operator are eigenvalues or *resonances*. For our $\mathcal{L}$, the endpoints are $\lambda = \pm i$. Resonance here refers to a function ϕ which satisfies the eigenvalue problem locally in space with eigenvalue i or $-i$ but which does not belong to $L^2(\mathbb{R}^n)$. For dimension $n = 1$, one requires $\phi \in L^\infty(\mathbb{R})$. (Note for comparison's sake that in one dimension, the operator $-d^2/dx^2$ has a resonance—corresponding to the constant function—at the endpoint 0 of its continuous spectrum.)

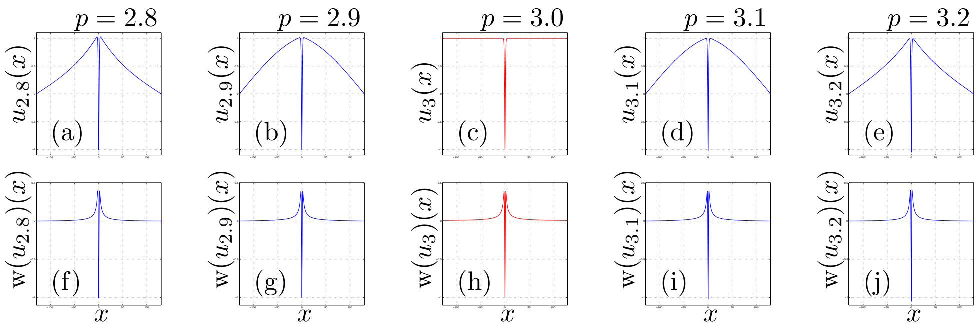


FIG. 8. *$u_p(x)$ and $\mathrm{w}(u_p)(x)$ for $p = 2.8, 2.9, 3, 3.1,$ and* 3.2.

Before we made the numerical calculation, we did not expect to see any resonance. However, from Figure 1, one sees that $\kappa = \sqrt{\mu_2}$ converges to 1 as $p \to 3$. What does the point $\kappa = 1$ at $p = 3$ correspond to? A natural conjecture is that it is a resonance or an eigenvalue, since the $p = 3$ case is well known to be completely integrable and special phenomena may occur.

This is indeed the case, since we have the following solution to the eigenvalue problem (2.12) when $p = 3$:

$$\phi = \begin{bmatrix} 1 - Q^2 \\ i \end{bmatrix}, \quad \lambda = i. \tag{3.54}$$

It is clear that $\phi \in L^\infty(\mathbb{R})$ but $\phi \notin L^q(\mathbb{R})$ for any $q < \infty$.

Let $u_p(x)$ denote the real-valued (and suitably normalized) solution of (2.14) corresponding to $\mu = \mu_2$. It is the first component of the eigenfunction of (2.12). A natural question is: does $u_p(x)$ converge in some sense to $u_3(x) := 1 - Q^2(x)$ as $p \to 3$? Since $u_p - u_3$ is not in $L^q(\mathbb{R})$ for all $q \in [1, \infty)$, it seems natural to measure the convergence in the following weighted norm,

$$\|f\|_{\mathrm{w}} := \int_{\mathbb{R}} \mathrm{w}(f)^2(x)\, dx,$$

where a weighting operator w is defined by $\mathrm{w}(f)(x) := f(x)\frac{1}{\sqrt{1+x^2}}$. This deemphasizes the value of $u_p - u_3$ for x large, and so it should converge to 0 as p goes to 3. This is confirmed numerically as follows.

Let $u_3 := 1 - Q^2$ and $\delta := \|u_3\|_{\mathrm{w}}$. In the appendix we will propose a numerical method to solve for the eigenpair $\{\lambda, [u_p(x),\ w_p(x)]^\top\}$ of (2.12) corresponding to $\mu_2 = -\lambda^2$. Renormalize $u_p(x)$ for $p \neq 3$ so that it is real-valued, $u_p(0) < 0$, and $\|u_p\|_{\mathrm{w}} = \delta$. In Figure 8(c) we plot u_3 in a large interval $|x| < 130$ with $\delta = 1.3588$. According to the numerical method in the appendix, we get $u_{2.8}, u_{2.9}, u_{3.1}$, and $u_{3.2}$ plotted in Figure 8(a), (b), (d), and (e), respectively. The vertical range is roughly $[-1, 1]$. In Figure 8(f)–(j) we plot $\mathrm{w}(u_p)$ for $p = 2.8, 2.9, 3, 3.1$, and 3.2, for $|x| < 130$, and with vertical range $[-1, 0.5]$.

In Figure 9 we plot p vs. $\|u_p - u_3\|_{\mathrm{w}}$ and observe that $u_p(x)$ converge to $u_3(x)$ in the weighted norm $\|\cdot\|_{\mathrm{w}}$ as $p \to 3$. In the numerical calculation for Figure 9, our increment for p is 0.01.

Remark 3.10. For the operators L_+L_- and L_-L_+, and in general fourth-order operators, it seems difficult to exclude the possibility that $\mu = 1$ is an eigenvalue. Consider the following example. Let $\tilde{H} := (L_+)^2$ with $p = \sqrt{8} - 1$. Note that -1 is

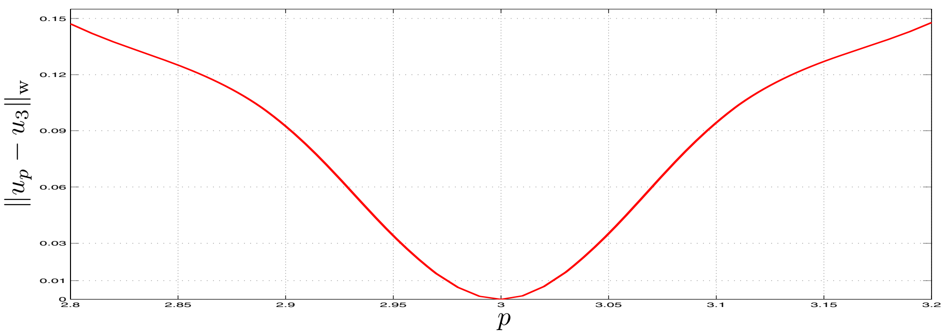


FIG. 9. p *vs.* $\|u_p - u_3\|_{\mathrm{w}}$.

an eigenvalue of L_+ when $p = \sqrt{8} - 1$. Hence 1 is an eigenvalue of $\tilde{H}$ at the endpoint of its continuous spectrum.

It would be interesting to prove the above convergence analytically and characterize the leading-order behavior near $p = 3$ as we did in Theorem 2.6.

4. Excited states with angular momenta. In this section we consider excited states with angular momenta in $\mathbb{R}^n$, $n \geq 2$. Let $k = [n/2]$, the largest integer no larger than $n/2$. For $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, use polar coordinates r_j and θ_j for each pair x_{2j-1} and x_{2j}, $j = 1, \dots, k$. Lions [22] considers solutions of the form

$$Q(x) = \phi(r_1, r_2, \dots r_k, x_n)\, e^{i(m_1\theta_1 + \cdots + m_k\theta_k)}, \quad m_j \in \mathbb{Z}.$$

The dependence of ϕ in x_n is dropped if n is even. He proves the existence of energy minimizing solutions in each such class.

For the simplest case $n = 2$, $Q(x) = \phi(r)\, e^{im\theta}$, and, by (1.3), $\phi = \phi(r)$ satisfies

$$-\phi'' - \frac{1}{r}\phi' + \frac{m^2}{r^2}\phi + \phi - |\phi|^{p-1}\phi = 0 \quad (r > 0). \tag{4.1}$$

The natural boundary conditions are

$$\lim_{r\to 0} r^{-m}\phi(r) = \alpha, \quad \lim_{r\to 0} r^{-m+1}\phi'(r) = m\alpha, \quad \lim_{r\to\infty} \phi(r) = 0 \tag{4.2}$$

for some $\alpha \geq 0$. One can choose $\phi(r)$ real-valued. It is shown by Iaia and Warchall [17] that (4.1)–(4.2) has countably infinite many solutions, denoted by $\phi_{m,k,p}(r)$, each having exactly k positive zeros. They correspond to "m-equivariant" nonlinear bound states of the form

$$Q_{m,k,p} = \phi_{m,k,p}(r)\, e^{im\theta} \quad (k = 0, 1, 2, \dots). \tag{4.3}$$

Note that $Q_{m,k,p}$ are radial if and only if $m = 0$, and the ground state $Q = Q_{0,0,p}$ is considered in the previous sections. The uniqueness question of $\phi_{m,k,p}(r)$ is not addressed in [17]. It is proved for the case $k = 0$ in [24].

Mizumachi [24], [25], [26], and [27] considered the stability problem for these solutions. He showed the following:

1. under m-equivariant perturbations of the form $\varepsilon(r)e^{im\theta}$, $Q_{m,0,p}$ are stable for $1 < p < 3$ and unstable for $p > 3$;
2. under general perturbations, $Q_{m,k,p}$ are unstable for $p > 3$ for any k;

3. linear (spectral) instability implies nonlinear instability (it can also be obtained by extending the results in [31] to higher dimensions using the method of [1]);
4. for fixed $p > 1$, if $m > M(p)$ is sufficiently large, $Q_{m,k,p}$ are linearly unstable and its linearized operator has a positive eigenvalue.

We are most interested in the last result. Intuitively, for $1 < p < \infty$, $Q_{m,k,p}$ should be unstable for all $(m, k) \neq (0, 0)$, since they are excited states. Can this be observed numerically? It turns out to be true for p away from 1 but false for p near 1.

In the following, we first describe our numerical methods for $k = 0$ and then discuss their relations. We will discuss our numerical results in the end. We compute only $m = 1, 2$, but the same methods work for other m.

Remark 4.1. Our numerical methods do not apply when $k > 0$. Indeed, for $m \geq 0$ and $k > 0$, the radial functions $\phi_{m,k,p}(r)$ are sign-changing and cannot be numerically calculated using the method described in the appendix. In fact, it is an open question whether they are unique. Assuming the uniqueness, one needs to develop a new algorithm to compute them before one can compute the spectra of $\mathcal{L}$ for $Q_{m,k,p}$.

4.1. Numerical algorithms. There are two steps in our numerical method: First, compute $\phi_{m,0,p}(r)$. Second, compute the spectra of the discretized linearized operator around $Q_{m,0,p}$. The second step is more involved, and we will present three algorithms.

Step 1. Compute $\phi(r) = \phi_{m,0,p}(r)$. It is energy minimizing among all solutions of (4.1)–(4.2) for fixed m, p, and it is positive for $r > 0$. Since our algorithm in the appendix is applicable to all positive (ground state) solutions, we can use it to calculate the discretized vector of $\phi(r)$ with a small change of the code.

Step 2. Compute the spectra of the discretized linearized operator. The linearized operator $\mathcal{L}$ has a slightly different form than (1.9) because $Q = \phi_{m,0,p}(r)e^{im\theta}$ is no longer real. With the same ansatz (1.6)–(1.7), the linearized operator $\mathcal{L}$ has the form

$$\mathcal{L}h = i\left(\Delta h - h + \frac{p+1}{2}|Q|^{p-1}h + \frac{p-1}{2}|Q|^{p-3}Q^2\bar{h}\right). \tag{4.4}$$

We have developed three algorithms for computing the spectrum of $\mathcal{L}$.

Algorithm 1. Write $Q = \phi(r)\, e^{im\theta} = \phi(r)\cos(m\theta) + i\phi(r)\sin(m\theta)$. In vector form with $\mathcal{L}$ acting on $[\operatorname{Re} h, \operatorname{Im} h]^\top$, we have

$$\mathcal{L} \sim \begin{bmatrix} 0 & -\Delta + 1 \\ \Delta - 1 & 0 \end{bmatrix} + |\phi(r)|^{p-1} \begin{bmatrix} -(p-1)\cos\sin & -\cos^2 - p\sin^2 \\ p\cos^2 + \sin^2 & (p-1)\cos\sin \end{bmatrix}(m\theta). \tag{4.5}$$

It is convenient to use polar coordinates to discretize the operator. We use a two-dimensional mesh:

$$\text{2d mesh:} \quad r = 0 : \delta_r : r_{\max}, \quad \theta = 0 : \delta_\theta : 2\pi. \tag{4.6}$$

The discretized matrix has size NT by NT with $N = r_{\max}/\delta_r$ and $T = 2\pi/\delta_\theta$. We use zero boundary condition with $r_{\max} = 15$, $\delta_r = 0.04$, and $T = 160$.

Although the matrix operator (4.5) is slightly more complicated than (1.9) and the mesh is two-dimensional, the same numerical routine can be applied to compute the spectrum of the discretized matrix of (4.5). The only difference is that the matrix size is much larger.

ALGORITHM 2. By restricting the problem to some invariant subspaces of $\mathcal{L}$, as we did for the computation of Figures 4–5, we will reduce the problem to one dimension.

Observe that functions of the form $a(r)e^{ij\theta}$ with a fixed integer j are not preserved by $\mathcal{L}$ unless $j = m$, but the following L^2-subspaces are invariant under $\mathcal{L}$:

$$X_k = X_k^{(m)} = \left\{h(r,\theta) : h = a(r)e^{i(m+k)\theta} + b(r)e^{i(m-k)\theta}\right\}, \quad 0 \le j \in \mathbb{N}.$$

If $k = 0$, we drop $b(r)$ and $X_0 = \left\{h(r,\theta) : h = a(r)e^{im\theta}\right\}$. We will compute the spectra of $\mathcal{L}$ limited to each subspace X_k. Define

$$V = \frac{p-1}{2}\phi^{p-1}, \quad H_k = -\Delta_r + 1 + \frac{(m+k)^2}{r^2} - \frac{p+1}{2}\phi^{p-1}.$$

For $k = 0$, with $a = a_1 + ia_2$, and $a_1, a_2 \in \mathbb{R}$, we have

$$\mathcal{L}[a(r)e^{im\theta}] = -i\left[H_0 a - V\bar{a}\right]e^{im\theta} = [H_0(a_2 - ia_1) + V(a_2 + ia_1)]e^{im\theta}.$$

Thus, acting on $[a_1, a_2]^\top$, $\mathcal{L}|_{X_0}$ has the matrix form

$$L_{X_0} = \begin{bmatrix} 0 & H_0 + V \\ -H_0 + V & 0 \end{bmatrix}.$$

For $k > 0$, with $a = a_1 + ia_2$, $b = b_1 + ib_2$, and $a_1, a_2, b_1, b_2 \in \mathbb{R}$, we have

$$\begin{aligned} &\mathcal{L}[a(r)e^{i(m+k)\theta} + b(r)e^{i(m-k)\theta}] \\ &= [H_k(a_2 - ia_1) + V(b_2 + ib_1)]e^{i(m+k)\theta} + [H_{-k}(b_2 - ib_1) + V(a_2 + ia_1)]e^{i(m-k)\theta}. \end{aligned}$$

Thus, acting on $[a_1, a_2, b_1, b_2]^\top$, $\mathcal{L}|_{X_k}$ has the matrix form

$$L_{X_k} = \begin{bmatrix} 0 & H_k & 0 & V \\ -H_k & 0 & V & 0 \\ 0 & V & 0 & H_{-k} \\ V & 0 & -H_{-k} & 0 \end{bmatrix}.$$

To discretize the operator, we use the one-dimensional mesh

$$\text{1d mesh:} \quad r = 0 : \delta_r : r_{\max}, \quad N = r_{\max}/\delta_r. \tag{4.7}$$

The matrix corresponding to X_0 has size $2N$ by $2N$. The matrix for X_k with $k > 0$ has size $4N$ by $4N$. We use zero boundary condition with $r_{\max} = 30$ and $\delta_r = 0.01$.

Counting multiplicity, the set of eigenvalues of $\mathcal{L}$ is the union of eigenvalues of $\mathcal{L}|_{X_k}$ with $k = 0, 1, 2, \ldots$.

ALGORITHM 3. Instead of the form (1.6), include the phase $e^{im\theta}$ in the linearization: $\psi = (\phi + h)e^{im\theta + it}$. Then the linearized operator acting on $[\operatorname{Re} h, \operatorname{Im} h]^\top$ is

$$\mathcal{L}' = \begin{bmatrix} -2m/r^2\partial_\theta & -\Delta + 1 + m^2/r^2 - \phi^{p-1} \\ -(-\Delta + 1 + m^2/r^2 - p\phi^{p-1}) & -2m/r^2\partial_\theta \end{bmatrix},$$

which is invariant on subspaces $Z_k = \left\{[a_1(r), a_2(r)]^\top e^{ik\theta}\right\}$ with integers k. We have

$$\mathcal{L}' \begin{bmatrix} a_1(r) \\ a_2(r) \end{bmatrix} e^{ik\theta} = e^{ik\theta} L_{m,k} \begin{bmatrix} a_1(r) \\ a_2(r) \end{bmatrix},$$

where

$$L_{m,k} := \begin{bmatrix} -\frac{2imk}{r^2} & -\Delta_r + 1 + \frac{m^2+k^2}{r^2} - \phi^{p-1} \\ -(-\Delta_r + 1 + \frac{m^2+k^2}{r^2} - p\phi^{p-1}) & -\frac{2imk}{r^2} \end{bmatrix}$$

acting on radial functions. We use the same one-dimensional mesh (4.7) as in Algorithm 2. For every k, the matrix size is $2N$ by $2N$. We then compute the spectra of $L_{m,k}$ for each k.

Counting multiplicity, the set of eigenvalues of $\mathcal{L}$ is the union of eigenvalues of $L_{m,k}$ with $k = 0, \pm 1, \pm 2, \dots$.

4.2. Properties of these algorithms. We now address the relation between these algorithms. First, note that X_k is essentially the sum of Z_k and Z_{-k}. Let us make it more precise, and suppose $k > 0$. The case $k = 0$ is easier. A function $h = (a_1 + ia_2)(r)e^{i(m+k)\theta} + (b_1 + ib_2)(r)e^{i(m-k)\theta}$ in $X_k \subset L^2(\mathbb{R}^2)$ can be identified with $[a_1, a_2, b_1, b_2] \in \tilde{X}_k = L^2_{rad}(\mathbb{R}^2; \mathbb{R}^4)$. The space $\tilde{X}_k$ is a subspace of $L^2_{rad}(\mathbb{R}^2; \mathbb{C}^4)$ on which we compute the spectrum. The function h can be also identified with

$$\begin{bmatrix} a_1(r) \\ a_2(r) \end{bmatrix} e^{ik\theta} + \begin{bmatrix} b_1(r) \\ b_2(r) \end{bmatrix} e^{-ik\theta},$$

the collection of which form a subspace of $Z_k \oplus Z_{-k}$ with real components.

Nullspace of $\mathcal{L}$. The nullspace of $\mathcal{L}$ gives a good test of the correctness of our numerical results. For $k = 0$, the 0-eigenfunction iQ of $\mathcal{L}$ corresponds to $[0, \phi]^\top e^{im\theta}$ in X_0 and $[0, \phi]^\top$ in Z_0. The generalized eigenfunction $Q_1 = \frac{2}{p-1}Q + x \cdot \nabla Q$ corresponds to $[Q_1, 0]^\top e^{im\theta}$ in X_0 and $[\frac{2}{p-1}\phi + r\phi', 0]^\top$ in Z_0. Since $X_0 \subset L^2(\mathbb{R}^2, \mathbb{C})$, they also provide two (generalized) eigenvectors for Algorithm 1.

For $k = \pm 1$, the 0-eigenfunctions

$$2Q_{x_1} = 2(\phi' \cos\theta - i\psi \sin\theta)e^{im\theta} = (\phi' - \psi)e^{i(m+1)\theta} + (\phi' + \psi)e^{i(m-1)\theta},$$

$$2Q_{x_2} = 2(\phi' \sin\theta + i\psi \cos\theta)e^{im\theta} = i(-\phi' + \psi)e^{i(m+1)\theta} + i(\phi' + \psi)e^{i(m-1)\theta},$$

where $\psi = m\phi/r$, belong to X_1, and correspond to 0-eigenvectors $[\phi' - \psi, 0, \phi' + \psi, 0]^\top$ and $[0, -\phi' + \psi, 0, \phi' + \psi]^\top$ of L_{X_1}. For Algorithm 3, they correspond to the following vectors in $Z_1 \oplus Z_{-1}$:

$$2\begin{bmatrix} \phi' \cos\theta \\ -\psi \sin\theta \end{bmatrix} = W_+ e^{i\theta} + W_- e^{-i\theta}, \quad 2\begin{bmatrix} \phi' \sin\theta \\ \psi \cos\theta \end{bmatrix} = -iW_+ e^{i\theta} + iW_- e^{-i\theta},$$

where

$$W_\pm = \begin{bmatrix} \phi' \\ \pm i\psi \end{bmatrix}, \quad L_{m,\pm 1} W_\pm = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Thus $W_+ e^{i\theta}$ is a 0-eigenvector of $\mathcal{L}'$ in Z_1, and $W_- e^{-i\theta}$ is a 0-eigenvector of $\mathcal{L}'$ in Z_{-1}.

The generalized eigenfunctions

$$\begin{aligned} ix_1 Q &= ir\phi \cos\theta e^{im\theta} = ir\phi e^{i(m+1)\theta} + ir\phi e^{i(m-1)\theta}, \\ ix_2 Q &= ir\phi \sin\theta e^{im\theta} = r\phi e^{i(m+1)\theta} - r\phi e^{i(m-1)\theta} \end{aligned}$$

also lie in X_1 and correspond to generalized 0-eigenvectors $[0, r\phi, 0, r\phi]^\top$ and $[r\phi, 0, -r\phi, 0]^\top$ of L_{X_1}. For Algorithm 3, they correspond to $[0, r\phi\cos\theta]^\top$ and $[0, r\phi\sin\theta]^\top$ in $Z_1 \oplus Z_{-1}$. By the same consideration as for Q_{x_1} and Q_{x_2}, their span over $\mathbb{C}$ is the same as the span of $[0, r\phi]^\top e^{i\theta} \in Z_1$ and $[0, r\phi]^\top e^{-i\theta} \in Z_{-1}$. One can check that

$$L_{m,\pm 1}\begin{bmatrix} 0 \\ r\phi \end{bmatrix} = -2\begin{bmatrix} \phi' \\ \pm i\psi \end{bmatrix}. \tag{4.8}$$

Thus, the multiplicity of 0-eigenvalue in each of X_0, Z_{-1}, Z_0, and Z_1 is at least 2. The multiplicity of 0-eigenvalue on X_1 is at least 4.

Symmetry of spectra. If

$$L_{m,k}\begin{bmatrix} A \\ B \end{bmatrix} = \lambda\begin{bmatrix} A \\ B \end{bmatrix},$$

then

$$L_{m,-k}\begin{bmatrix} \bar{A} \\ \bar{B} \end{bmatrix} = \bar{\lambda}\begin{bmatrix} \bar{A} \\ \bar{B} \end{bmatrix}, \quad L_{m,-k}\begin{bmatrix} A \\ -B \end{bmatrix} = -\lambda\begin{bmatrix} A \\ -B \end{bmatrix}, \quad L_{m,k}\begin{bmatrix} \bar{A} \\ -\bar{B} \end{bmatrix} = -\bar{\lambda}\begin{bmatrix} \bar{A} \\ -\bar{B} \end{bmatrix}.$$

In particular, if $\lambda \in \sigma(L_{m,k})$, then $-\bar{\lambda} \in \sigma(L_{m,k})$, and $\bar{\lambda}, -\lambda \in \sigma(L_{m,-k})$. Thus $\sigma(L_{m,k})$ itself is symmetric w.r.t. the imaginary axis, and $\sigma(L_{m,k})$ and $\sigma(L_{m,-k})$ are symmetric w.r.t. the real axis.

Similarly, one can show that the spectra of L_{X_k} are symmetric w.r.t. both real and imaginary axes.

Equivalence of Algorithms 2 and 3. In Algorithm 2, for $k > 0$, we can write

$$L_{X_k} = \begin{bmatrix} H_k J & VU \\ VU & H_{-k} J \end{bmatrix},$$

where

$$J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Let

$$M = \begin{bmatrix} I & -J \\ I & J \end{bmatrix}, \quad M^{-1} = \frac{1}{2}\begin{bmatrix} I & I \\ J & -J \end{bmatrix}, \quad P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad P^{-1} = P.$$

Noting that $JU = -UJ$, we have

$$M^{-1} L_{X_k} M = \begin{bmatrix} \alpha J + cU & -\beta \\ -\beta & \alpha J + cU \end{bmatrix} = \begin{bmatrix} 0 & \alpha + c & \beta & 0 \\ -\alpha + c & 0 & 0 & \beta \\ -\beta & 0 & 0 & \alpha + c \\ 0 & -\beta & -\alpha + c & 0 \end{bmatrix},$$

where

$$\alpha = \frac{1}{2}(H_k + H_{-k}) = H_0 + \frac{k^2}{r^2}, \quad \beta = \frac{1}{2}(H_k - H_{-k}) = \frac{2mk}{r^2}, \quad c = V.$$

Let

$$L' := P^{-1}M^{-1}L_{X_k}MP = \begin{bmatrix} 0 & \beta & \alpha+c & 0 \\ -\beta & 0 & 0 & \alpha+c \\ -\alpha+c & 0 & 0 & \beta \\ 0 & -\alpha+c & -\beta & 0 \end{bmatrix}.$$

In Algorithm 3, $L_{m,k}$ acts on $[A(r), B(r)]^\top$. If we write the enlarged matrix of $L_{m,k}$ acting on $[\operatorname{Re} A, \operatorname{Im} A, \operatorname{Re} B, \operatorname{Im} B]^\top$, the matrix is exactly L'. The matrix for $L_{m,-k}$ will be also L' if it acts on $[\operatorname{Re} A, -\operatorname{Im} A, \operatorname{Re} B, -\operatorname{Im} B]^\top$. This amounts to a choice of assigning J or $-J$ to the complexification of i.

More precisely, if $L_{m,k}u = \lambda u$ with $u = [A, B]^\top$, then $L_{m,k}iu = \lambda iu$. Write $A = A_1 + iA_2$ and $B = B_1 + iB_2$, and suppose $k > 0$. These two equations are equivalent to

$$L' \begin{bmatrix} A_1 \\ A_2 \\ B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} \operatorname{Re} \lambda A \\ \operatorname{Im} \lambda A \\ \operatorname{Re} \lambda B \\ \operatorname{Im} \lambda B \end{bmatrix}, \quad L' \begin{bmatrix} -A_2 \\ A_1 \\ -B_2 \\ B_1 \end{bmatrix} = \begin{bmatrix} -\operatorname{Im} \lambda A \\ \operatorname{Re} \lambda A \\ -\operatorname{Im} \lambda B \\ \operatorname{Re} \lambda B \end{bmatrix}.$$

Adding the second equation multiplied by $-i$ to the first equation, we get

$$L'w = \lambda w, \quad w = [A, -iA, B, -iB]^\top.$$

Taking conjugation we get $L'\bar{w} = \bar{\lambda}\bar{w}$. Thus λ and $\bar{\lambda}$ are eigenvalues of L' and hence of L_{X_k}. Since $L_{m,k}u = \lambda u$ if and only if $L_{m,-k}\bar{u} = \bar{\lambda}\bar{u}$, eigenvalues of $L_{m,-k}$ also correspond to eigenvalues of L_{X_k}.

Counting eigenvalues. L_{X_k} acts on $L^2_{rad}(\mathbb{R}^2, \mathbb{C}^4)$ and $L_{m,\pm k}$ on $L^2_{rad}(\mathbb{R}^2, \mathbb{C}^2)$. The set of eigenvalues of L_{X_k} is the union of eigenvalues of $L_{m,k}$ and $L_{m,-k}$. For any ball B_R on the complex plane disjoint from the continuous spectrum $\Sigma_c = \{ir : r \in \mathbb{R}, |r| \geq 1\}$,

$$\#(\sigma(L_{X_k}) \cap B_R) = \#(\sigma(L_{m,k}) \cap B_R) + \#(\sigma(L_{m,-k})) \cap B_R,$$

which is equal to $2\#(\sigma(L_{m,k}) \cap B_R)$ if the center of B_R is on the real axis.

Numerical efficiency. Algorithm 1 is two-dimensional and thus more expensive to compute and less accurate. Both Algorithms 2 and 3 are one-dimensional and more accurate.

The benefit of Algorithm 3 rather than Algorithm 2 is that it further decomposes the subspace of $L^2(\mathbb{R}^2, \mathbb{C}^4)$ corresponding to X_k to two subspaces. Although its matrix size is only half that of Algorithm 2, its components are complex and hence require more storage space. Numerically these two algorithms are not very different.

4.3. Numerical results. The results of our numerical computations of the spectra of $\mathcal{L}$ for $m = 1, 2$ and various k and p are shown in Figures 10–15. As before, we focus on eigenvalues in the square $\{a + bi : |a| < 1, |b| < 1\}$. Purely imaginary eigenvalues with modulus greater than 1 correspond to the continuous spectrum of $\mathcal{L}$ and are discrete due to discretization.

Let us first describe some simple observations:

1. The distribution of eigenvalues (see Figures 10–11) is more complicated and interesting than Figures 1–5. There are not only purely imaginary eigenvalues and real eigenvalues but also complex eigenvalues, whose existence implies instability.

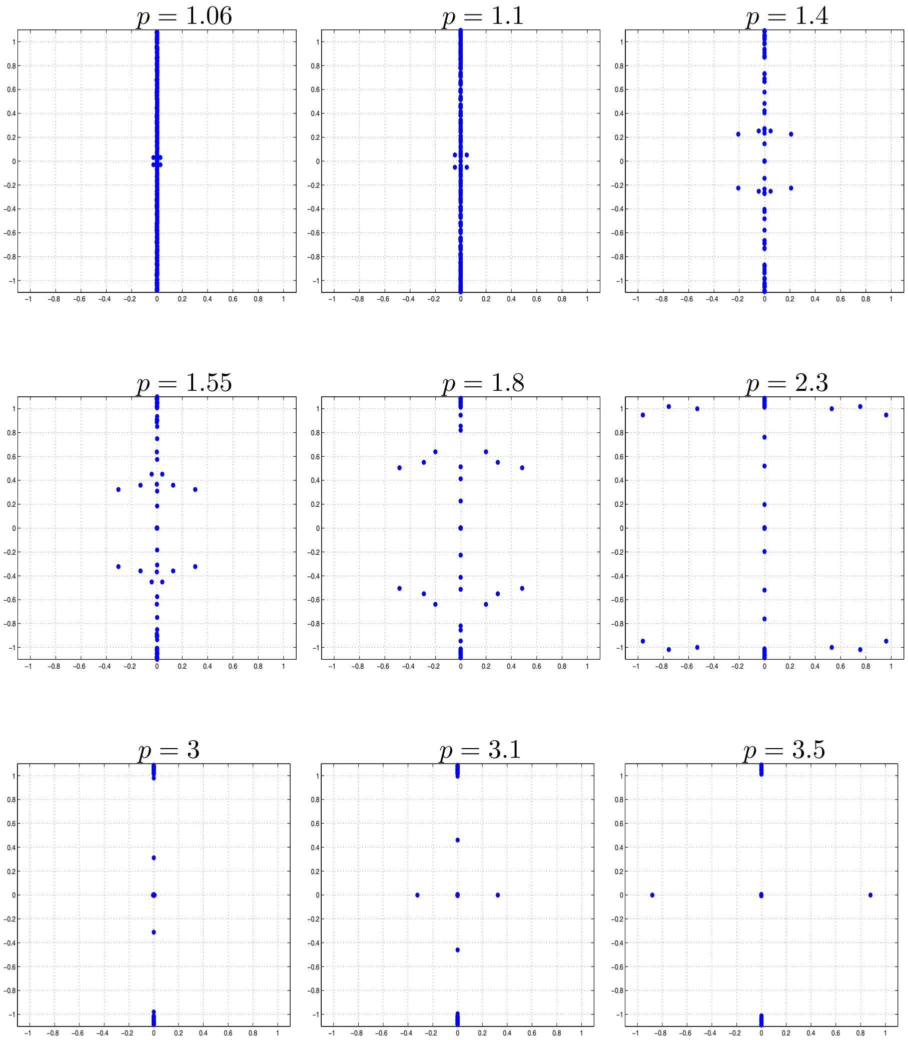


FIG. 10. *Spectra of* $\mathcal{L}$ *in* $\mathbb{R}^2$ *for* $m = 1$ *and* $0 \le k \le 24$ *as* $p = 1.06, 1.1, 1.4, 1.55, 1.8, 2.3, 3, 3.1, 3.5$ *computed by Algorithm* 3.

2. In Figures 12–13, we compare the results obtained from three algorithms. For Algorithms 2 and 3 the parameter k ranges from 0 to 9. Results from these algorithms have a high degree of agreement, except when p is near 3 and eigenvalues are near 0. We will discuss this exceptional case in the end.
3. For Algorithms 2 and 3, the numerical 0-eigenvalue occurs only when $k = 0$ and $k = \pm 1$. It agrees with our discussion in the previous subsection. Their multiplicities also match, and there is no unaccounted eigenvector. In particular, $N_g(\mathcal{L})$ has dimension 6 if $p \neq 3$ and 8 if $p = 3$, the same as the case of ground states. We also numerically verified the nullspace; for example, the discrete version of (4.8) is correct.
4. As p increases, two pairs of purely imaginary eigenvalues may collide away from 0 and then split into a quadruple of complex eigenvalues which are neither real nor purely imaginary. For $m = 1$, this bifurcation phenomenon

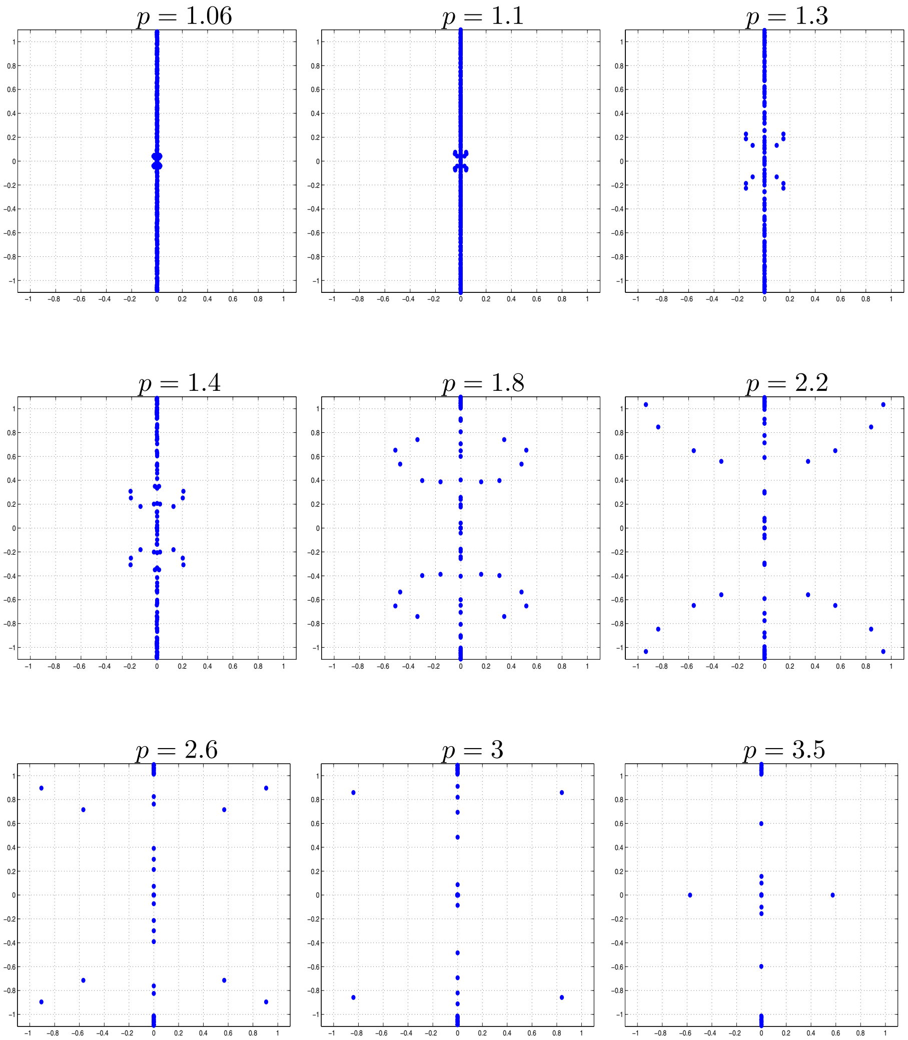


FIG. 11. *Spectra of* $\mathcal{L}$ *in* $\mathbb{R}^2$ *for* $m = 2$ *and* $0 \leq k \leq 28$ *as* $p = 1.06, 1.1, 1.3, 1.4, 1.8, 2.2, 2.6, 3, 3.5$ *computed by Algorithm* 3.

appears three times before $p = 1.55$, and there are three complex quadruples for $p > 1.55$. For $m = 2$, it occurs five times before $p = 1.5$, and there are five quadruples for $p > 1.5$. These complex eigenvalues seem to move away from the imaginary axis as p increases further.

5. As p increases to 3 (by Algorithms 2 and 3), a pair of purely imaginary eigenvalues from the 0th subspace collides at 0 and then splits into a pair of real eigenvalues as p increases further. This is the same picture as in the ground state case in section 3. Indeed, Mizumachi [24] proves that $Q_{m,0,p}$ are stable in the 0th subspace if $p < 3$ and unstable if $p > 3$. Thus $p = 3$ is a bifurcation point. Also note that when $p = 3$ the NLS (1.1) has conformal invariance, and explicit blow-up solutions can be found as in the ground state case.
6. In Figures 14–15 we observe the bifurcation more closely. For $m = 1$, the

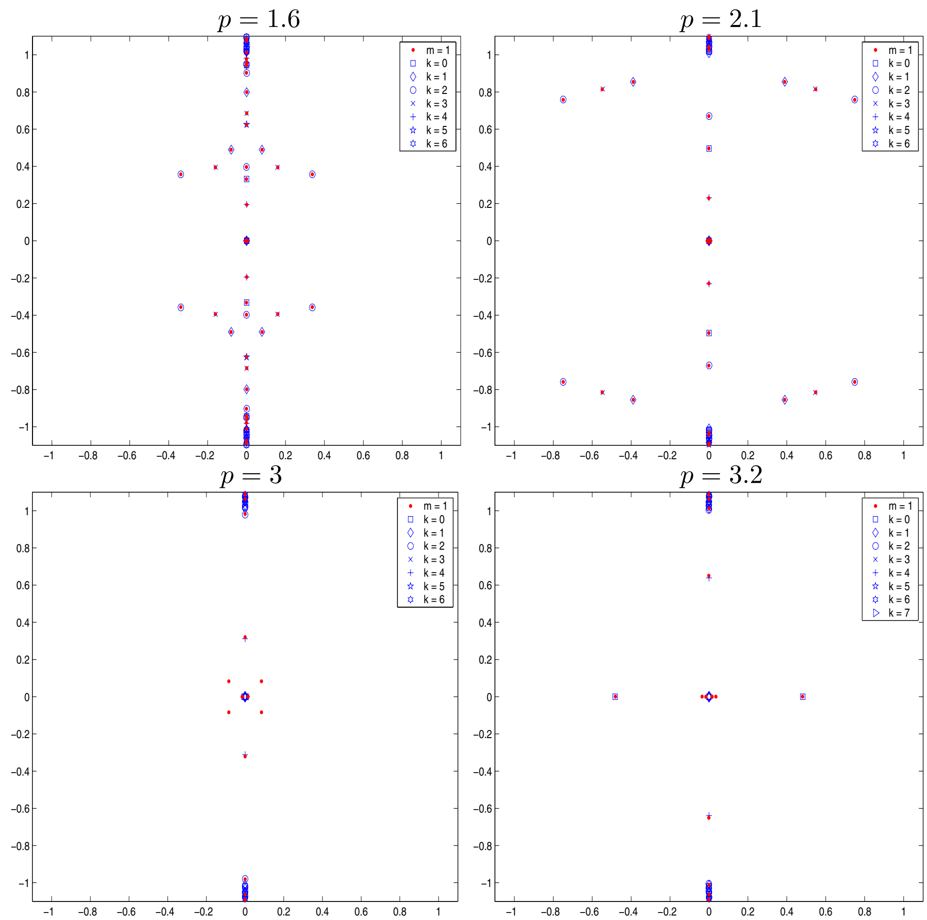


FIG. 12. *Spectra of $\mathcal{L}$ in $\mathbb{R}^2$ for $m = 1$ and various $p = 1.6, 2.1, 3, 3.2$. Point "·" denotes the spectra computed by Algorithm* 1*, and the other symbols denote the spectra computed by Algorithms* 2 *and* 3*.*

bifurcation occurs when (k, p, λ) equal

$$(0, 3, 0), \quad (1, 1.5276, -0.436i), \quad (2, 1.0165, -0.016i), \quad (3, 1.3495, -0.219i).$$

For $m = 2$, the bifurcation occurs when (k, p, λ) equal

$$(0,\ 3,\ 0), \quad (1,\ 1.357,\ -0.180i), \quad (2,\ 1.007,\ -0.027i), \quad (3,\ 1.0245,\ -0.035i)$$

and

$$(4,\ 1.0455,\ -0.045i), \quad (5,\ 1.3955,\ -0.347i).$$

7. Due to the existence of complex eigenvalues for $m = 1, 2$ and $p \geq 1.02$, $Q_{m,0,p}$ is spectrally unstable for these parameters. However, all these complex eigenvalues bifurcate from some discrete eigenvalues $\pm bi$ with $|b| < 1$ and $p > 1.008$. Our computation for both $m = 1, 2$ and

$$p = \ell \cdot 0.001, \quad \ell = 1, 2, 3, \ldots, 8 \quad (\text{up to } 15 \text{ if } m = 1),$$

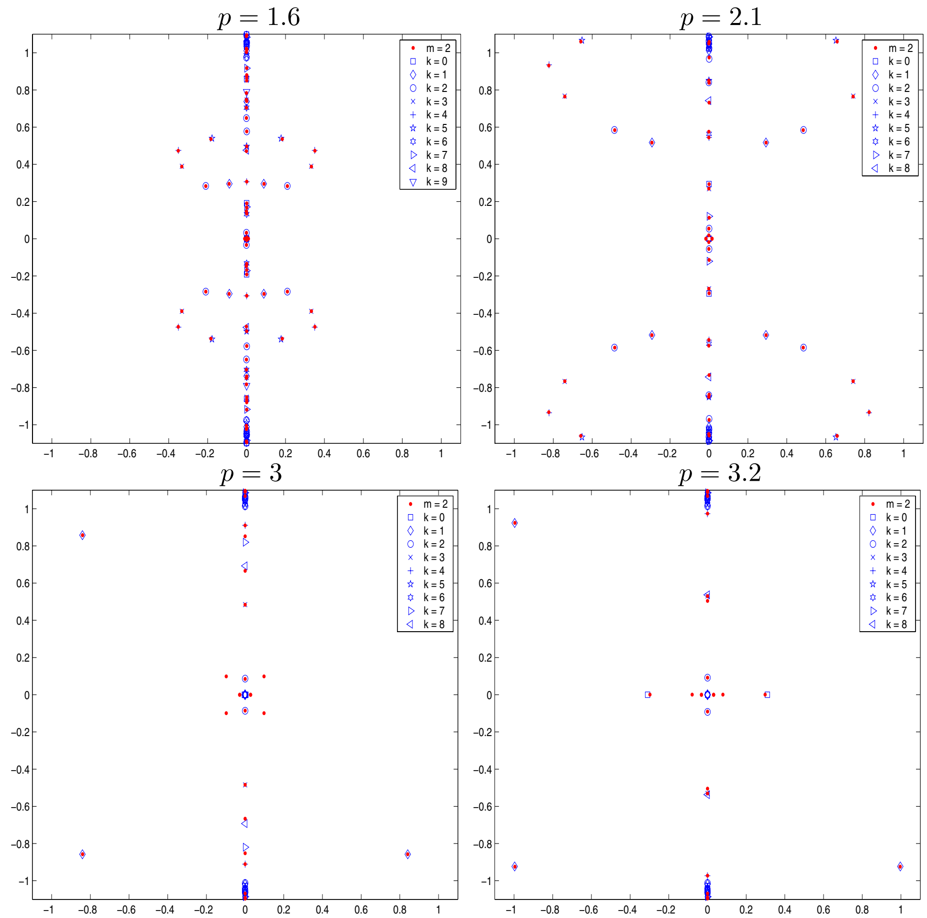


FIG. 13. *Spectra of $\mathcal{L}$ in $\mathbb{R}^2$ for $m = 2$ and various $p = 1.6, 2.1, 3, 3.2$. Point "·" denotes the spectra computed by Algorithm* 1*, and the other symbols denote the spectra computed by Algorithms* 2 *and* 3*.*

does not find any complex eigenvalues. This suggests that the two excited states $\phi_{1,0,p}(r)e^{i\theta}$ and $\phi_{2,0,p}(r)e^{i2\theta}$ are *linearly stable* when p is sufficiently close to 1. It is possible that the numerical error increases enormously as $p \to 1_+$ due to the artificial boundary condition, since the spectrum is approaching the continuous one for $p = 1$. This has to be verified analytically in the future.

We finally discuss the exceptional case when p is near 3 for eigenvalues near 0. In this case Algorithm 1 produces a quadruple of complex eigenvalues $\pm 0.0849 \pm 0.0836i$, and the 0-eigenvalue has multiplicity 4. We expect to see larger errors from Algorithm 1, but the error in this exceptional case is much larger. It is related to the large size of a Jordan block for the 0-eigenvalue. As discussed in the previous subsection, the nullspace is at least six-dimensional. The analysis in section 2.3 suggests that (we do not claim a proof), as p goes to the bifurcation exponent $p_c = 3$ from below, a pair of imaginary eigenvalues merges into the Jordan block containing the eigenfunctions iQ and Q_1, and the Jordan block becomes size 4. As is well known in matrix analysis (see [11, p. 324], [41]), if a matrix contains a Jordan block of size ℓ, the computed

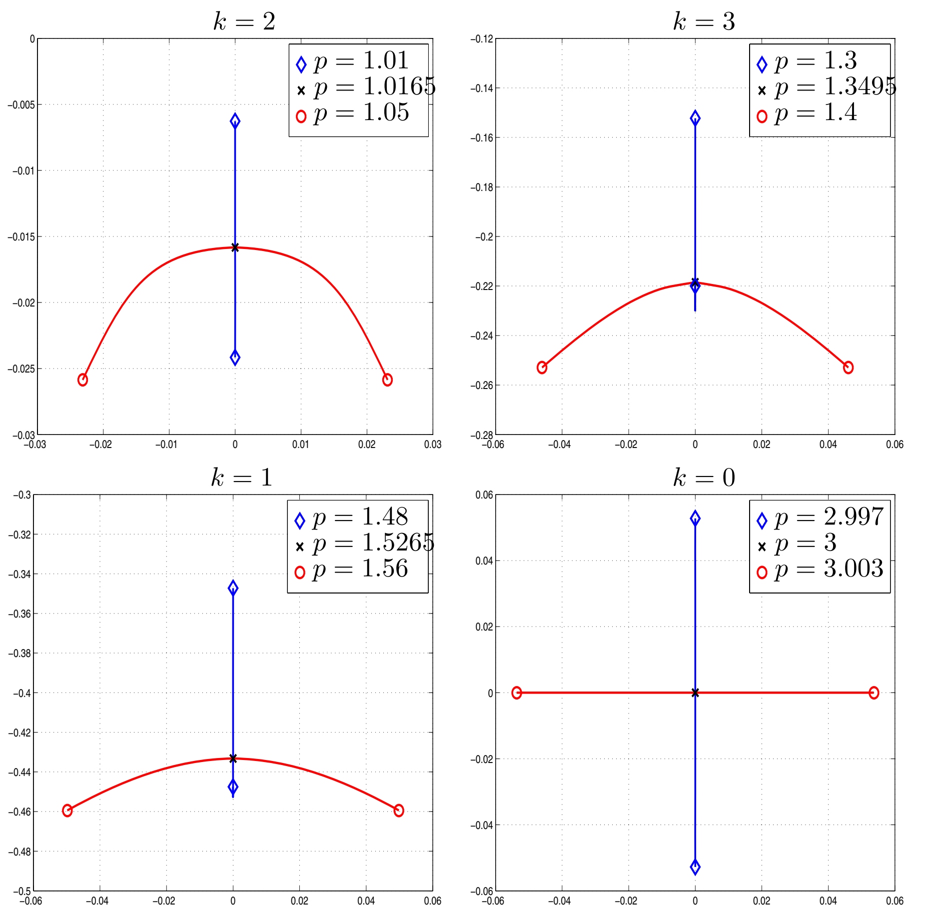


FIG. 14. *Bifurcation diagrams of* $L_{m,k}$ *for* $m = 1$ *and* $0 \leq k \leq 3$.

eigenvalues corresponding to that block have errors of order $\varepsilon^{1/\ell}$, where ε is the sum of the machine zero, the truncation error from discretization, and the perturbation (from varying p). Since $\delta_r = 0.04$ for Algorithm 1 and the truncation error of a central difference scheme for Δ_r has order $O(\delta_r^2)$, the error for the zero eigenvalue near $p = 3$ could be

$$(\delta_r^2)^{1/4} \approx 0.2.$$

In contrast, for other bifurcation points on the imaginary axis, the Jordan block at the bifurcation exponent is of size 2 and the error is of order $(\delta_r^2)^{1/2} = 0.04$. In practice, the error is smaller due to cancellation, and the numerical results by Algorithm 1 do not differ too much from those by Algorithms 2 and 3. Also note that numerically the 0-eigenspace has dimension 4, accounting for Q_{x_j} and ix_jQ. The complex quadruple corresponds to iQ, Q_1, and the joining pair of nonzero eigenvalues.

Appendix: Numerical method. In this appendix we describe a numerical method to compute the spectrum of the linear operator $\mathcal{L}$ defined by (1.8) for $p > 1$ and space dimension $n \geq 1$. There are two main steps in this method. First, we

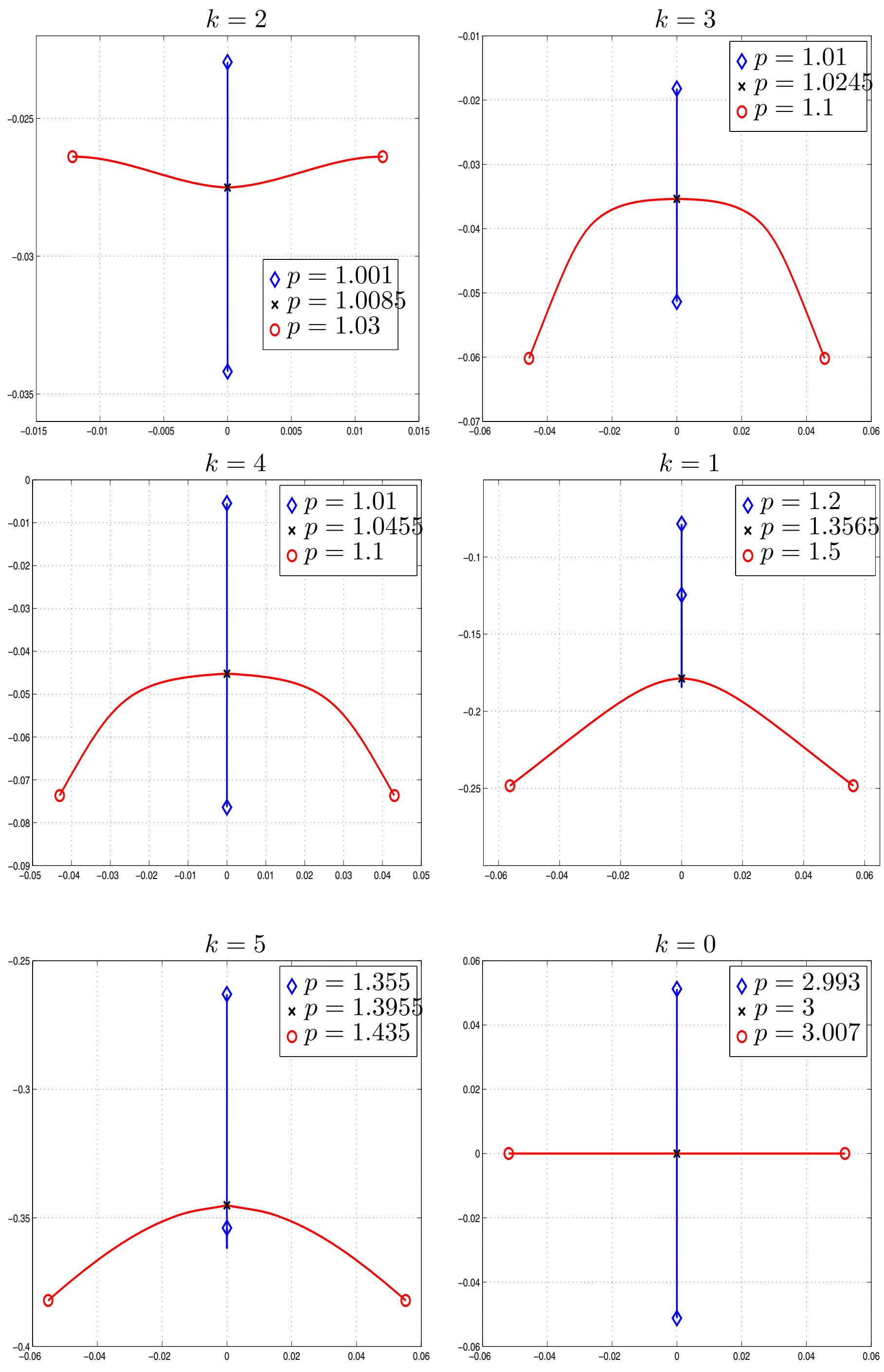


FIG. 15. *Bifurcation diagrams of* $L_{m,k}$ *for* $m = 2$ *and* $0 \le k \le 5$.

will solve the nonlinear problem (1.3) for Q: we will discretize it into a nonlinear algebraic equation and then solve it by an iterative method. Second, we will compute the spectrum of $\mathcal{L}$: we will discretize the operator $\mathcal{L}$ into a large-scale linear algebraic eigenvalue problem and then use implicitly restarted Arnoldi methods to deal with this problem.

Hereafter, we use the boldface letters or symbols to denote a matrix or a vector. For $\mathbf{A} \in \mathbb{R}^{M\times N}$, $\mathbf{q} = (q_1, \ldots, q_N)^\top \in \mathbb{R}^N$, $\mathbf{q}^{ⓟ} = \mathbf{q} \circ \cdots \circ \mathbf{q}$ denotes the p-time Hadamard product of $\mathbf{q}$, and $[\![\mathbf{q}]\!] := \operatorname{diag}(\mathbf{q})$ denotes the diagonal matrix of $\mathbf{q}$.

Step I. We first discretize (1.3) into a nonlinear algebraic equation and consider it on an n-dimensional ball $\Omega = \{x \in \mathbb{R}^n : |x| \le R, R \in \mathbb{R}\}$. We rewrite the Laplace operator $-\Delta$ in the polar coordinate system with a Dirichlet boundary condition. Based on the recently proposed discretization scheme [20], the *standard central finite difference method*, we discretize $-\Delta\mathbf{q}(x)$ into

$$\mathbf{A}\mathbf{q} = \mathbf{A}[q_1, \ldots, q_N]^\top, \ \mathbf{A} \in \mathbb{R}^{N\times N}, \tag{A.1}$$

where $\mathbf{q}$ is an approximation of the function $Q(x)$. The matrix $\mathbf{A}$ is irreducible and diagonally dominant with positive diagonal entries. The discretization of the nonlinear equation (1.3) can now be formulated as the following nonlinear algebraic equation:

$$\mathbf{A}\mathbf{q} + \mathbf{q} - \mathbf{q}^{ⓟ} = 0. \tag{A.2}$$

We introduce an iterative algorithm [16] to solve (A.2):

$$\mathbf{A}\widetilde{\mathbf{q}}_{j+1} + \widetilde{\mathbf{q}}_{j+1} = \mathbf{q}_j^{ⓟ}, \tag{A.3}$$

where $\widetilde{\mathbf{q}}_{j+1}$ and $\mathbf{q}_j$ are the unknown and known discrete values of the function $Q(\mathbf{x})$, respectively. The iterative algorithm is shown below.

Iterative Algorithm for Solving $Q(\mathbf{x})$.

Step 0 Let $j = 0$.
  Choose an initial solution $\widetilde{\mathbf{q}}_0 > 0$, and let $\mathbf{q}_0 = \frac{\widetilde{\mathbf{q}}_0}{\|\widetilde{\mathbf{q}}_0\|_2}$.
Step 1 Solve (A.3); then obtain $\widetilde{\mathbf{q}}_{j+1}$.
Step 2 Let $\alpha_{j+1} = \frac{1}{\|\widetilde{\mathbf{q}}_{j+1}\|_2}$ and normalize $\widetilde{\mathbf{q}}_{j+1}$ to obtain $\mathbf{q}_{j+1} = \alpha_{j+1}\widetilde{\mathbf{q}}_{j+1}$.
Step 3 If (convergent), then
    Output the scaled solution $(\alpha_{j+1})^{\frac{1}{p-1}}\mathbf{q}_{j+1}$. Stop.
  else
    Let $j := j + 1$.
    Goto Step 1.
  end

If the components of $\mathbf{q}_0$ are nonnegative, this property is preserved by each iteration $\mathbf{q}_j$ and hence also by the limit vector if it exists (see [16, Theorem 3.1]). The convergence of a subsequence of this iteration method to a nonzero vector is proved in [16, Theorem 2.1]. Although the convergence of the entire sequence is not proved, it is observed numerically to be very robust. See Chen, Zhou, and Ni [6] for a survey on numerically solving nonlinear elliptic equations.

Step II. Next we discretize the operator $\mathcal{L}$ of (1.10) into a linear algebraic eigenvalue problem:

$$\mathbf{L}\begin{bmatrix}\mathbf{u}\\ \mathbf{w}\end{bmatrix} = \lambda\begin{bmatrix}\mathbf{u}\\ \mathbf{w}\end{bmatrix}, \tag{A.4}$$

where

$$\mathbf{L} = \begin{bmatrix} 0 & \mathbf{A} + \mathrm{I} - [\![\mathbf{q}^{\odot}]\!] \\ -\mathbf{A} - \mathrm{I} + [\![p\,\mathbf{q}^{\odot}]\!] & 0 \end{bmatrix},$$

$\gamma = p - 1$, $\mathbf{u} = (u_1, \ldots, u_N)^\top \in \mathbb{R}^N$, $\mathbf{w} = (w_1, \ldots, w_N)^\top \in \mathbb{R}^N$, and $\mathbf{q}$ is the output of the previous step, and satisfies the equation in (A.2). We use ARPACK [21] in MATLAB version 6.5 to deal with the linear algebraic eigenvalue problem (A.4) and obtain eigenvalues λ of $\mathbf{L}$ near the origin for $p > 1$ and space dimension $n \geq 1$. Furthermore, the eigenvectors of $\mathbf{L}$ can be also produced.

Step II above can in principle be used to compute all eigenfunctions in $L^2(\mathbb{R}^n)$. However, in producing Figures 2–5, we look for eigenfunctions of the form $\phi(r)e^{im\theta}$. These problems can be reformulated as one-dimensional eigenvalue problems for $\phi(r)$, which can be computed using the same algorithm and MATLAB code. This dimensional reduction saves a lot of computation time and memory. Even with this dimensional reduction, and applying an algorithm for sparse matrices, the computation is still very heavy, and we cannot compute all eigenvalues in one step. We can compute only a portion of them each time.

Acknowledgments. Special thanks go to Wen-Wei Lin, whose constant advice and interest were indispensable for this project. We thank V.S. Buslaev, Y. Martel, F. Merle, T. Mizumachi, C. Sulem, and W.C. Wang for fruitful discussions and providing many references to us. We also thank the referees for many valuable suggestions and for pointing out [7] to us. Part of the work was done when Nakanishi was visiting the University of British Columbia, Vancouver, and when Tsai was visiting the National Center for Theoretical Sciences, Hsinchu and Taipei. The hospitality of these institutions is gratefully acknowledged.

REFERENCES

[1] L. Almeida and Y. Guo, *Dynamical instability of symmetric vortices*, Rev. Mat. Iberoamericana, 17 (2001), pp. 409–419.

[2] H. Berestycki and P.-L. Lions, *Nonlinear scalar field equations.* I., Arch. Rational Mech. Anal., 82 (1983), pp. 313–345.

[3] H. Berestycki and P.-L. Lions, *Nonlinear scalar field equations.* II., Arch. Rational Mech. Anal., 82 (1983), pp. 347–375.

[4] V. S. Buslaev and V. E. Grikurov, *Simulation of instability of bright solitons for NLS with saturating nonlinearity*, Math. Comput. Simulation, 56 (2001), pp. 539–546.

[5] T. Cazenave, *Semilinear Schrödinger equations*, AMS, Providence, RI, 2003.

[6] G. Chen, J. Zhou, and W.-M. Ni, *Algorithms and visualization for solutions of nonlinear elliptic equations*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 10 (2000), pp. 1565–1612.

[7] A. Comech and D. Pelinovsky, *Purely nonlinear instability of standing waves with minimal energy*, Comm. Pure Appl. Math., 56 (2003), pp. 1565–1607.

[8] S. Cuccagna, D. Pelinovsky, and V. Vougalter, *Spectra of positive and negative energies in the linearized NLS problem*, Comm. Pure Appl. Math., 58 (2005), pp. 1–29.

[9] L. Demanet and W. Schlag, *Numerical verification of a gap condition for a linearized nonlinear Schrödinger equation*, Nonlinearity, 19 (2006), pp. 829–852.

[10] J. Fröhlich, S. Gustafson, L. Jonsson, and I. M. Sigal, *Solitary wave dynamics in an external potential*, Comm. Math. Phys., 250 (2004), pp. 613–642.

[11] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., John Hopkins University Press, Baltimore, MD, 1996.

[12] V. E. Grikurov, *Perturbation of Instable Solitons for Generalized NLS with Saturating Nonlinearity*, http://math.nw.ru/~grikurov/publlist.html/proc_DD97.ps.gz.

[13] M. Grillakis, *Linearized instability for nonlinear Schrödinger and Klein-Gordon equations*, Comm. Pure Appl. Math., 41 (1988), pp. 747–774.

[14] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry* I, J. Funct. Anal., 74 (1987), pp. 160–197.

[15] S. Gustafson, K. Nakanishi, and T.-P. Tsai, *Asymptotic stability and completeness in the energy space for nonlinear Schrödinger equations with small solitary waves*, Int. Math. Res. Not., 2004 (2004), pp. 3559–3584.

[16] T. M. Hwang and W. Wang, *Analyzing and visualizing a discretized semilinear elliptic problem with Neumann boundary conditions*, Numer. Methods Partial Differential Equations, 18 (2002), pp. 261–279.

[17] J. Iaia and H. Warchall, *Nonradial solutions of a semilinear elliptic equation in two dimensions*, J. Differential Equations, 119 (1995), pp. 533–558.

[18] E. Kirr and A. Zarnescu, *On the asymptotic stability of bound states in 2D cubic Schrödinger equation*, Comm. Math. Phys., 272 (2007), pp. 443–468.

[19] M. K. Kwong, *Uniqueness of positive solutions of $\Delta u - u + u^p = 0$ in $\mathbb{R}^n$*, Arch. Rational Mech. Anal., 105 (1989), pp. 243–266.

[20] M.-C. Lai, *A note on finite difference discretizations for Poisson equation on a disk*, Numer. Methods Partial Differential Equations, 17 (2001), pp. 199–203.

[21] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, Philadelphia, 1998.

[22] P. L. Lions, *Solutions complexes d'équations elliptiques semilinéaires dans R^N*, C. R. Acad. Sci. Paris Sér. I Math., 302 (1986), pp. 673–676.

[23] V. B. Matveev and M. A. Salle, *Darboux Transformations and Solitons*, Springer Ser. Nonlinear Dynam., Springer-Verlag, Berlin, 1991.

[24] T. Mizumachi, *Vortex solitons for 2D focusing nonlinear Schrödinger equation*, Differential Integral Equations, 18 (2005), pp. 431–450.

[25] T. Mizumachi, *Instability of bound states for 2D nonlinear Schrödinger equations*, Discrete Contin. Dyn. Syst., 13 (2005), pp. 413–428.

[26] T. Mizumachi, *A remark on linearly unstable standing wave solutions to NLS*, Nonlinear Anal., 64 (2006), pp. 657–676.

[27] T. Mizumachi, *Instability of vortex solitons for 2D focusing NLS*, Adv. Differential Equations, 12 (2007), pp. 241–264.

[28] S. I. Pohozaev, *Eigenfunctions of the equation $\Delta u + \lambda f(u) = 0$*, Soviet Math. Dokl., 5 (1965), pp. 1408–1411.

[29] I. Rodnianski, W. Schlag, and A. Soffer, *Asymptotic Stability of N-Soliton States of NLS*, http://arxiv.org/abs/math.AP/0309114.

[30] W. Schlag, *Stable Manifolds for an Orbitally Unstable NLS*, http://arxiv.org/abs/math.AP/0405435.

[31] J. Shatah and W. Strauss, *Spectral condition for instability*, in Nonlinear PDE's, Dynamics and Continuum Physics (South Hadley, MA, 1998), Contemp. Math. 255, AMS, Providence, RI, 2000, pp. 189–198.

[32] W. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149–162.

[33] W. Strauss, *Nonlinear Wave Equations*, AMS, Providence, RI, 1989.

[34] C. Sulem and P.-L. Sulem, *The Nonlinear Schrödinger Equations: Self-Focusing and Wave Collapse*, Springer-Verlag, Berlin, 1999.

[35] E. C. Titchmarsh, *Eigenfunction Expansions Associated with Second-Order Differential Equations, Part* I, 2nd ed., Clarendon Press, Oxford, UK, 1962.

[36] T.-P. Tsai and H.-T. Yau, *Asymptotic dynamics of nonlinear Schroedinger equations: Resonance dominated and dispersion dominated solutions*, Comm. Pure Appl. Math., 55 (2002), pp. 153–216.

[37] T.-P. Tsai and H.-T. Yau, *Stable directions for excited states of nonlinear Schrödinger equations*, Comm. Partial Differential Equations, 27 (2002), pp. 2363–2402.

[38] M. I. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates*, Comm. Math. Phys., 87 (1982/83), pp. 567–576.

[39] M. I. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal., 16 (1985), pp. 472–491.

[40] M. I. Weinstein, *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–68.

[41] J. H. Wilkinson, *Convergence of the LR, QR, and related algorithms*, Comput. J., 8 (1965), pp. 77–84.

SIAM J. MATH. ANAL.
Vol. 39, No. 4, pp. 1112–1119

EXTREMAL FIRST DIRICHLET EIGENVALUE OF DOUBLY CONNECTED PLANE DOMAINS AND DIHEDRAL SYMMETRY*

AHMAD EL SOUFI† AND ROLA KIWAN†

Abstract. We deal with the following eigenvalue optimization problem: Given a bounded domain $D \subset \mathbf{R}^2$, place an obstacle B of fixed shape within D so as to maximize or minimize the fundamental eigenvalue λ_1 of the Dirichlet Laplacian on $D \setminus B$. This means that we want to extremize the function $\rho \mapsto \lambda_1(D \setminus \rho(B))$, where ρ runs over the set of rigid motions such that $\rho(B) \subset D$. We answer this problem in the case where both D and B are invariant under the action of a dihedral group $\mathbb{D}_n$, $n \geq 2$, and where the distance from the origin to the boundary is monotonous as a function of the argument between two axes of symmetry. The extremal configurations correspond to the cases where the axes of symmetry of B coincide with those of D.

Key words. eigenvalues, Dirichlet Laplacian, Schrödinger operator, extremal eigenvalue, obstacle, dihedral group

AMS subject classifications. 35J10, 35P15, 49R50, 58J50

DOI. 10.1137/060670250

1. Introduction and statement of the main result.

The relations between the shape of a domain and the eigenvalues of its Dirichlet or Neumann Laplacian have been intensively investigated since the 1920s when Faber [5] and Krahn [12] proved independently the famous eigenvalue isoperimetric inequality first conjectured by Rayleigh (1877): the first Dirichlet eigenvalue $\lambda_1(\Omega)$ of any bounded domain $\Omega \subset \mathbf{R^n}$ satisfies

$$\lambda_1(\Omega) \geq \lambda_1(\Omega^*),$$

where Ω^* is a ball having the same volume as Ω. We refer to the review papers of Ashbaugh [1, 2] and Henrot [9] for a survey of recent results on optimization problems involving eigenvalues.

The present work deals with the following eigenvalue optimization problem: Given a bounded domain D, we want to place an obstacle (or a hole) B, of fixed shape, inside D so as to maximize or minimize the fundamental eigenvalue λ_1 of the Laplacian or Schrödinger operator on $D \setminus B$ with zero Dirichlet conditions on the boundary.

In other words, the problem is to optimize the principal eigenvalue function $\rho \mapsto \lambda_1(D \setminus \rho(B))$, where ρ runs over the set of rigid motions such that $\rho(B) \subset D$.

The first result obtained in this direction concerned the case where both D and B are disks of given radii. Indeed, it follows from Hersch's work [10] that the maximum of λ_1 is achieved when the disks are concentric (see also [14]). This result has been extended to any dimension by several authors (see, e.g., Harrell, Kröger, and Kurata [8] and Kesavan [11]). Actually, Harrell, Kröger, and Kurata [8] gave a more general result showing that if the domain D satisfies an interior symmetry property with respect to a hyperplane P passing through the center of the spherical obstacle B

*Received by the editors September 19, 2006; accepted for publication (in revised form) April 23, 2007; published electronically November 7, 2007.
http://www.siam.org/journals/sima/39-4/67025.html

†Laboratoire de Mathématiques et Physique Théorique, UMR CNRS 6083, Université François Rabelais Tours, Parc de Grandmont, F-37200 Tours France (elsoufi@univ-tours.fr, kiwan@lmpt.univ-tours.fr).

(which means that the image by the reflection with respect to P of one component of $D \setminus P$ is contained in D), then the Dirichlet fundamental eigenvalue $\lambda_1(D \setminus B)$ decreases when the center of B moves perpendicularly to P in the direction of the boundary of D. In the particular case where both the domain D and the obstacle B are balls, this implies that the minimum of $\lambda_1(D \setminus B)$ corresponds to the limit case where B touches the boundary of D.

Notice that when the obstacle B is a disk, only translations of B may affect the eigenvalues of $D \setminus B$, and the optimal placement problem reduces to the choice of the center of B inside D.

In the present work we investigate a kind of dual problem in the sense that we consider a *nonspherical* obstacle B whose center of mass is fixed inside D, and we seek the optimal positions while turning B around its center.

It is of course hopeless to expect a universal solution to this problem. In fact, we will restrict our investigation to a class of domains satisfying a dihedral symmetry and monotonicity conditions.

Thus, let D be a simply connected plane domain and assume that the following conditions are satisfied:

(i) ($\mathbb{D}_n$-symmetry.) For an integer $n \geq 2$, D is invariant under the action of the dihedral group $\mathbb{D}_n$ of order $2n$ generated by the rotation $\rho_{\frac{2\pi}{n}}$ of angle $\frac{2\pi}{n}$ and a reflection S. Such a domain admits n axes of symmetry passing through the origin and such that the angle between 2 consecutive axes is $\frac{\pi}{n}$.

(ii) (Monotonicity of the boundary.) The distance $d(O,x)$ from the origin to a point x of the boundary of D is monotonous as a function of the argument of x, in a sector delimited by two consecutive symmetry axes.

Notice that assumption (i) guarantees that the center of mass of D is at the origin. Regular n-gones centered at the origin are the simplest examples of domains satisfying these assumptions. More generally, if g is any positive even $\frac{2\pi}{n}$-periodic continuous function that is monotonous on the interval $(0, \frac{\pi}{n})$, then the domain

$$D = \{re^{\mathbf{i}\theta}; \theta \in [0, 2\pi), 0 \leq r < g(\theta)\}$$

satisfies assumptions (i) and (ii). Actually, up to a rigid motion, any domain satisfying assumptions (i) and (ii) can be parametrized in such a manner.

It is worth noticing that, due to the monotonicity condition, the "distance to the origin" function on the boundary of D achieves its maximum and its minimum alternatively at the intersection points of ∂D with the $2n$ half-axes of symmetry. The n points of ∂D at maximal (resp., minimal) distance from the origin will be called "outer vertices" (resp., "inner vertices") of D.

Our main result is the following.

THEOREM 1. *Let D and B be two plane domains satisfying the assumptions of $\mathbb{D}_n$-symmetry and monotonicity* (i) *and* (ii) *above for an integer $n \geq 2$. Assume furthermore that B has C^2 boundary and that $\rho(B) \subset D$ for all $\rho \in SO(2)$. Then the fundamental Dirichlet eigenvalue $\lambda_1(D \setminus B)$ of $D \setminus B$ is optimized exactly when the axes of symmetry of B coincide with those of D.*

The maximizing configuration corresponds to the case where the outer vertices of B and D lie on the same half-axes of symmetry (we will then say that B occupies the "ON" position in D).

The minimizing configuration corresponds to the case where the outer vertices of B lie on the half-axes of symmetry passing through the inner vertices of D (this is what will be called the "OFF" position).

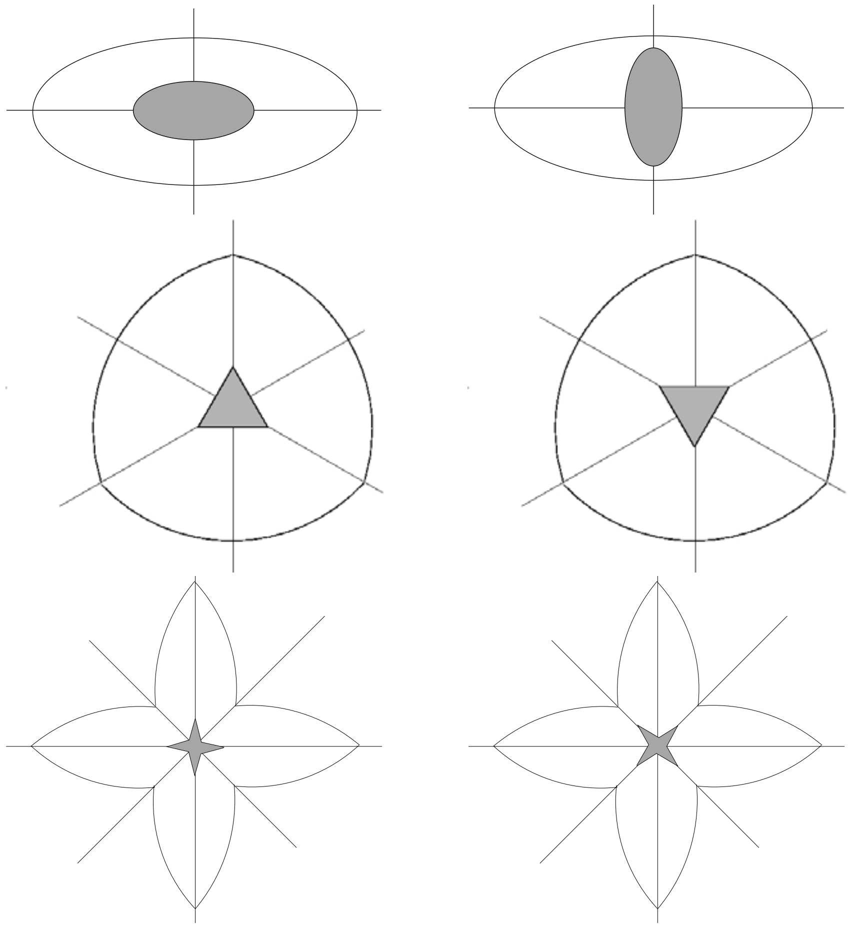


FIG. 1. *Examples of maximal (left) and minimal (right) configurations with* $n = 2$, 3, *and* 4, *respectively.*

Actually, we will prove that, except for the trivial case where D or B is a disk, the fundamental Dirichlet eigenvalue of $D \setminus B$ decreases gradually when B switches from "ON" to "OFF" (see Figure 1 for examples).

The main ingredients of the proof of Theorem 1 are Hadamard's variation formula for λ_1 and the technique of domain reflection initiated by Serrin [17] in a PDE setting.

Extensions of Theorem 1 to the following situations can be obtained up to slight changes in the proof (indeed, only the Hadamard formula should be replaced by the variation formula corresponding to the new functional):

1. *Soft obstacles*. Instead of considering the Dirichlet Laplacian on $D \setminus B$, we consider the Schrödinger-type operator

$$H(\alpha, B) := \Delta - \alpha\chi_B$$

acting on $H_0^1(D)$, where $\alpha > 0$ and χ_B is the indicator function of B. Optimization problems related to the fundamental eigenvalue of operators of this kind have been investigated in particular in [8] and [3]. Under the assumptions of Theorem 1 on D and B, for all $\alpha > 0$, the fundamental eigenvalue of $H(\alpha, B)$ achieves its maximum at the "ON" position and its minimum at the "OFF" position.

2. *Wells.* This case corresponds to the operator $H(\alpha, B)$ with $\alpha < 0$. Under the circumstances of Theorem 1, for all $\alpha < 0$, the first eigenvalue of $H(\alpha, B)$ achieves its maximum at the "OFF" position and its minimum at the "ON" position.
3. *Stationary problem.* The problem now is to optimize the Dirichlet energy $J(D \setminus B) := \int_{D\setminus B} |\nabla u|^2 dx$ of the unique solution u of the problem

$$\begin{cases} \Delta u = -1 & \text{in } D \setminus B, \\ u = 0 & \text{on } \partial(D \setminus B). \end{cases}$$

   This problem was treated in [11, section 2] in the case where both D and B are balls. Under the assumptions of Theorem 1 on D and B, one can prove that $J(D \setminus B)$ achieves its maximum when B is at the "ON" position and its minimum when B is at the "OFF" position.

2. Proof of the main result. Without loss of generality, we may assume that the domain D and the obstacle B are centered at the origin and are both symmetric with respect to the x_1-axis so that they can be parametrized in polar coordinates by

$$D = \{re^{\mathbf{i}\theta}; \theta \in [0, 2\pi), 0 \le r < g(\theta)\},$$

$$B = \{re^{\mathbf{i}\theta}; \theta \in [0, 2\pi), 0 \le r < f(\theta)\},$$

where f and g are two positive even $\frac{2\pi}{n}$-periodic functions which are *nondecreasing* on $(0, \frac{\pi}{n})$. To avoid technicalities, we suppose throughout that g is continuous and f is C^2. Extensions of our result to a wider class of domains would certainly be possible up to some additional technical difficulties.

The condition that the obstacle B can freely rotate around his center inside D, that is, $\rho(\bar{B}) \subset D$ for all $\rho \in SO(2)$, amounts to the following:

$$f\left(\frac{\pi}{n}\right) = \max_{0\le\theta\le 2\pi} f(\theta) < \min_{0\le\theta\le 2\pi} g(\theta) = g(0).$$

Let us denote, for all $t \in \mathbf{R}$, by ρ_t the rotation of angle t, that is, for all $\zeta \in \mathbf{R^2} \cong \mathbf{C}$, $\rho_t(\zeta) = e^{\mathbf{i}t}\zeta$, and set

$$B_t := \rho_t(B) \quad \text{and} \quad \Omega(t) := D \setminus B_t.$$

Let $\lambda(t)$ be the fundamental eigenvalue of the Dirichlet Laplacian on $\Omega(t)$. It is well known that, since it is simple, the first Dirichlet eigenvalue $\lambda(t)$ is a differentiable function of t (see [6, 15]). We denote by $u(t)$ the one parameter family of nonnegative first eigenfunctions satisfying, for all $t \in \mathbf{R}$,

$$\begin{cases} \Delta u(t) = -\lambda(t) u(t) & \text{in } \Omega(t), \\ u(t) = 0 & \text{on } \partial\Omega(t), \\ \int_{\Omega(t)} u^2(t) = 1. \end{cases}$$

The derivative of $\lambda(t)$ is then given by the following so-called Hadamard formula (see [4, 6, 7, 16]):

$$\lambda'(t) = \int_{\partial B_t} \left|\frac{\partial u(t)}{\partial \eta_t}\right|^2 \eta_t \cdot v \, d\sigma, \tag{1}$$

where η_t is the inward unit normal vector field of $\partial\Omega(t)$ (hence, along ∂B_t the vector η_t is outward with respect to B_t) and v denotes the restriction to $\partial\Omega(t) = \partial D \cup \partial B_t$ of the deformation vector field. In our case, the vector v vanishes on ∂D and is given by $v(\zeta) = \mathbf{i}\zeta$ for all $\zeta \in \partial B_t$.

Since both Ω and B are invariant by the dihedral group $\mathbb{D}_n$, it follows that, for all $t \in \mathbf{R}$, $\Omega(t + \frac{2\pi}{n}) = \Omega_t$. Moreover, if we denote by S_0 the reflection with respect to the x_1-axis, then we clearly have $\rho_{-t} = S_0 \circ \rho_t \circ S_0$, which gives $B_{-t} = S_0(B_t)$ and $\Omega_{-t} = S_0(\Omega_t)$. Hence, as a function of t, the first Dirichlet eigenvalue of Ω_t is even and periodic of period $\frac{2\pi}{n}$, that is, for all $t \in \mathbf{R}$,

$$\lambda\left(t + \frac{2\pi}{n}\right) = \lambda(t) \quad \text{and} \quad \lambda(-t) = \lambda(t).$$

Therefore, it suffices to investigate the variations of $\lambda(t)$ on the interval $\left[0, \frac{\pi}{n}\right]$, and Theorem 1 is a consequence of the following.

Theorem 2. *Assume that neither D nor B is a disk.*

(i) *For all $t \in \left(0, \frac{\pi}{n}\right)$, $\lambda'(t) < 0$. Hence, $\lambda(t)$ is strictly decreasing on $\left(0, \frac{\pi}{n}\right)$.*

(ii) *For all $k \in \mathbf{Z}$, $\lambda'(k\frac{\pi}{n}) = 0$ and $k\frac{\pi}{n}$, $k \in \mathbf{Z}$, are the only critical points of λ on $\mathbf{R}$.*

Hence, $\lambda(t)$ achieves its maximum for $t = 0 \mod \frac{2\pi}{n}$, which corresponds to the "ON" position, and its minimum for $t = \frac{\pi}{n} \mod \frac{2\pi}{n}$, which corresponds to the "OFF" position. Of course, if D or B is a disk, then the function $\lambda(t)$ is constant.

In what follows we will denote, for any $\alpha \in \mathbf{R}$, by z_α the $\theta = \alpha$ axis, that is, $z_\alpha := \{re^{\mathbf{i}\alpha};\ r \in \mathbf{R}\}$, and by z_α^+ the half-axis $\{re^{\mathbf{i}\alpha};\ r \geq 0\}$.

We start the proof with the following elementary lemma.

Lemma 1. *Let K be a plane domain defined in polar coordinates by $K = \{re^{\mathbf{i}\theta}; \theta \in [0, 2\pi), 0 \leq r < h(\theta)\}$, where h is a positive 2π-periodic function of class C^1, and let v be a vector field whose restriction to ∂K is given by*

$$v(\theta) := v(h(\theta)e^{\mathbf{i}\theta}) = \mathbf{i}h(\theta)e^{\mathbf{i}\theta} = h(\theta)e^{\mathbf{i}(\theta + \frac{\pi}{2})}.$$

We denote by η the unit outward normal vector field of ∂K. One has, at any point $h(\theta)e^{\mathbf{i}\theta}$ of ∂K where η is defined, the following:

(i) $\eta(\theta) := \eta(h(\theta)e^{\mathbf{i}\theta}) = \frac{h(\theta)e^{\mathbf{i}\theta} - \mathbf{i}h'(\theta)e^{\mathbf{i}\theta}}{\sqrt{h^2(\theta) + h'^2(\theta)}}$.

(ii) $\eta \cdot v(\theta) = \frac{-h(\theta)h'(\theta)}{\sqrt{h^2(\theta) + h'^2(\theta)}}$. *Hence, $\eta . v(\theta)$ has constant sign on an interval I If and only if h is monotonous in I.*

(iii) *If for some $\alpha > 0$, the domain K is symmetric with respect to the axis z_α, then the function $\eta \cdot v$ is antisymmetric with respect to this axis, that is,*

$$\eta \cdot v(\alpha + \theta) = -\eta \cdot v(\alpha - \theta).$$

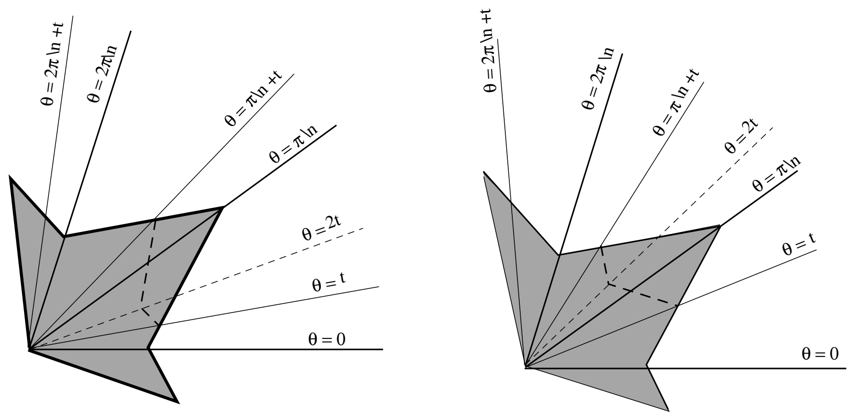
Proof. Assertions (i) and (ii) are direct consequences from the definition of K. The fact that K is symmetric with respect to the axis z_α implies that the function h satisfies $h(\alpha + \theta) = h(\alpha - \theta)$. Therefore, (iii) follows immediately from (ii). ◻

We will denote by S_α the symmetry with respect to the axis z_α. We will also denote, for $\alpha < \beta$, by $\sigma(\alpha, \beta)$ the sector delimited by z_α^+ and z_β^+, that is,

$$\sigma(\alpha, \beta) = \{re^{\mathbf{i}\theta}; r > 0 \text{ and } \alpha < \theta < \beta\}.$$

Lemma 2. *Let D be as above. For all $t \in \left(0, \frac{\pi}{n}\right)$, we have*

$$S_{\frac{\pi}{n}+t}\left(D \cap \sigma\left(\frac{\pi}{n} + t, \frac{2\pi}{n} + t\right)\right) \subseteq D \cap \sigma\left(t, \frac{\pi}{n} + t\right).$$



case $2t < \frac{\pi}{n}$ case $2t > \frac{\pi}{n}$

FIG. 2.

Moreover, if D is not a disk, then

$$S_{\frac{\pi}{n}+t}\left(\partial D \cap \sigma\left(\frac{\pi}{n}+t, \frac{2\pi}{n}+t\right)\right) \cap D \neq \emptyset.$$

Proof. The action of the symmetry $S_{\frac{\pi}{n}+t}$ is given in polar coordinates by $S_{\frac{\pi}{n}+t}(re^{\mathbf{i}\theta}) = re^{\mathbf{i}(2(\frac{\pi}{n}+t)-\theta)}$. Hence,

$$S_{\frac{\pi}{n}+t}\left(D \cap \sigma\left(\frac{\pi}{n}+t, \frac{2\pi}{n}+t\right)\right) = S_{\frac{\pi}{n}+t}(D) \cap \sigma\left(t, \frac{\pi}{n}+t\right).$$

Moreover, the domain D being parametrized by a positive even $\frac{2\pi}{n}$-periodic function $g(\theta)$, that is, $D = \{re^{\mathbf{i}\theta}; \theta \in [0, 2\pi), 0 \leq r < g(\theta)\}$, its image $S_{\frac{\pi}{n}+t}(D)$ can be parametrized in the same manner by the function $g^*(\theta) = g(\theta - 2t)$. Thus

$$S_{\frac{\pi}{n}+t}(D) \cap \sigma\left(t, \frac{\pi}{n}+t\right) = \left\{re^{\mathbf{i}\theta}; \theta \in \left(t, \frac{\pi}{n}+t\right), 0 \leq r < g(\theta - 2t)\right\}.$$

Therefore, we need to prove that $F(\theta) = g(\theta) - g^*(\theta)$ is nonnegative for every θ in the interval $(t, \frac{\pi}{n}+t)$. This will be possible thanks to the assumptions of symmetry (that is, g is even and $\frac{2\pi}{n}$-periodic) and monotonicity (that is, g is nondecreasing on $[0, \frac{\pi}{n}]$). Indeed, these properties imply that on the interval $\left(t, \frac{\pi}{n}+t\right)$,

- g achieves its maximum at $\theta = \frac{\pi}{n}$,
- g^* achieves its minimum at $\theta = 2t$.

Four cases must be considered separately (see Figure 2):

- If $t < \theta \leq \min\{2t, \frac{\pi}{n}\}$, we may write, since g is even, $F(\theta) = g(\theta) - g(2t - \theta)$, with $0 \leq 2t - \theta < \theta \leq \frac{\pi}{n}$. Since g is nondecreasing on $[0, \frac{\pi}{n}]$, we get $F(\theta) \geq 0$.
- If $\max\{2t, \frac{\pi}{n}\} \leq \theta < \frac{\pi}{n} + t$, we may write, since g is even and $\frac{2\pi}{n}$-periodic, $F(\theta) = g(2\frac{\pi}{n} - \theta) - g(\theta - 2t)$ with $0 \leq \theta - 2t < 2\frac{\pi}{n} - \theta \leq \frac{\pi}{n}$. Hence, $F(\theta) \geq 0$.
- If $2t < \frac{\pi}{n}$ and $2t \leq \theta \leq \frac{\pi}{n}$, then $0 \leq \theta - 2t < \theta \leq \frac{\pi}{n}$ and, then, $F(\theta) = g(\theta) - g(\theta - 2t) \geq 0$.
- If $2t > \frac{\pi}{n}$ and $\frac{\pi}{n} \leq \theta \leq 2t$, then $0 \leq 2t - \theta < 2\frac{\pi}{n} - \theta \leq \frac{\pi}{n}$ and, then, $F(\theta) = g(2\frac{\pi}{n} - \theta) - g(2t - \theta) \geq 0$.

Hence, $F(\theta)$ is nonnegative for all θ in $(t, \frac{\pi}{n}+t)$.

Now, if D is not a disk, then g is nonconstant on $[0, \frac{\pi}{n}]$. Following the arguments above, we deduce that the function $F(\theta)$ is positive somewhere on $(t, \frac{\pi}{n}+t)$, which means that $S_{\frac{\pi}{n}+t}\left(\partial D \cap \sigma\left(\frac{\pi}{n}+t, \frac{2\pi}{n}+t\right)\right)$ meets the interior of D. ∎

Proof of Theorem 2. Notice first that, since λ is an even and $\frac{2\pi}{n}$-periodic function of t, one immediately gets, for all $k \in \mathbf{Z}$, $\lambda(k\frac{\pi}{n}-t) = \lambda(k\frac{\pi}{n}+t)$ and, then,

$$\lambda'\left(k\frac{\pi}{n}\right) = 0.$$

Alternatively, one can deduce that $\lambda'\left(k\frac{\pi}{n}\right) = 0$ from Hadamard's variation formula (1) after noticing that the domain $\Omega(k\frac{\pi}{n})$ is symmetric with respect to the x_1-axis and that the first Dirichlet eigenfunction $u(k\frac{\pi}{n})$ satisfies $u \circ S_0 = u$, where S_0 is the symmetry with respect to the x_1-axis.

Let us fix a t in $\left(0, \frac{\pi}{n}\right)$ and denote by u the nonnegative first Dirichlet eigenfunction of $\Omega(t)$ satisfying $\int_{\Omega(t)} u^2 = 1$. The domain $\Omega(t)$ is clearly invariant by the rotation $\rho_{\frac{2\pi}{n}}$ of angle $\frac{2\pi}{n}$, and hence $u \circ \rho_{\frac{2\pi}{n}} = u$. On the other hand, the domain B being parametrized by a positive even $\frac{2\pi}{n}$-periodic function $f(\theta)$, that is, $B = \{re^{\mathbf{i}\theta}; \theta \in [0, 2\pi), 0 \le r < f(\theta)\}$, one has

$$B_t = \{re^{\mathbf{i}\theta}; \theta \in [0, 2\pi), 0 \le r < h(\theta)\},$$

with $h(\theta) = f(\theta - t)$. Hence, the function $\eta_t \cdot v$ is invariant by $\rho_{\frac{2\pi}{n}}$ (Lemma 1) and we have (Hadamard formula (1))

$$\lambda'(t) = \int_{\partial B_t} \left|\frac{\partial u}{\partial \eta_t}\right|^2 \eta_t \cdot v \; d\sigma = n \int_{\partial B_t \cap \sigma(t, \frac{2\pi}{n}+t)} \left|\frac{\partial u}{\partial \eta_t}\right|^2 \eta_t \cdot v \; d\sigma.$$

Since B_t is symmetric with respect to the axis $z_{\frac{\pi}{n}+t}$, we have (Lemma 1) $\eta_t \cdot v(\frac{\pi}{n} + t + \theta) = -\eta_t \cdot v(\frac{\pi}{n} + t - \theta)$ or, equivalently, $\eta_t \cdot v(x) = -\eta_t \cdot v(x^*)$, where x^* denotes the symmetric of x with respect to $z_{\frac{\pi}{n}+t}$. This yields

$$\lambda'(t) = n \int_{\partial B_t \cap \sigma(\frac{\pi}{n}+t, \frac{2\pi}{n}+t)} \left(\left|\frac{\partial u}{\partial \eta_t}(x)\right|^2 - \left|\frac{\partial u}{\partial \eta_t}(x^*)\right|^2\right) \eta_t \cdot v(x) \; d\sigma.$$

Notice that the function $h(\theta)$ is decreasing between $\frac{\pi}{n}+t$ and $\frac{2\pi}{n}+t$ and, then, $\eta_t \cdot v$ is nonnegative on $\partial B_t \cap \sigma(\frac{\pi}{n}+t, \frac{2\pi}{n}+t)$ (Lemma 1).

Let $H(t) := \Omega(t) \cap \sigma(\frac{\pi}{n}+t, \frac{2\pi}{n}+t)$. Applying Lemma 2, and since B_t is symmetric with respect to the axis $z_{\frac{\pi}{n}+t}$, one gets

$$S_{\frac{\pi}{n}+t}(H(t)) \subset \Omega(t) \cap \sigma\left(t, \frac{\pi}{n}+t\right).$$

Hence, the function $w(x) = u(x) - u(x^*)$ is well defined on $H(t)$ and satisfies $w(x) = 0$ for all x in $\partial H(t) \cap \left(\partial B_t \cup z_{\frac{\pi}{n}+t} \cup z_{\frac{2\pi}{n}+t}\right)$. Moreover, since u vanishes on ∂D and is positive inside $\Omega(t)$, $w(x) \le 0$ for all x in $\partial H(t) \cap \partial D$ and $w(x) < 0$ for certain x in $\partial H(t) \cap \partial D$ (recall that D is not a disk and apply the second part of Lemma 2).

Therefore, the nonconstant function w satisfies the following:

$$\begin{cases} \Delta w = -\lambda(t) w & \text{in } H(t), \\ w \le 0 & \text{on } \partial H(t). \end{cases}$$

Hence, w must be nonpositive on the whole of $H(t)$. Otherwise, a nodal domain $V \subset H(t)$ of w would have the same first Dirichlet eigenvalue as $\Omega(t)$. But, due to the invariance of $\Omega(t)$ by $\rho_{\frac{2\pi}{n}}$, the domain $\Omega(t)$ would contain n copies of V leading to a strong contradiction with the domain monotonicity theorem for eigenvalues. Therefore, $\Delta w \geq 0$ in $H(t)$ and w achieves its maximal value (i.e., zero) on $\partial B_t \cap \sigma(\frac{\pi}{n} + t, \frac{2\pi}{n} + t) \subset \partial H(t)$. The Hopf maximum principle (see [13, Theorem 7, Chapter 2]) then implies that, at any regular point x of $\partial B_t \cap \sigma(\frac{\pi}{n} + t, \frac{2\pi}{n} + t)$, one has

$$\frac{\partial w}{\partial \eta_t}(x) = \frac{\partial u}{\partial \eta_t}(x) - \frac{\partial u}{\partial \eta_t}(x^*) < 0.$$

It follows that $\lambda'(t) \leq 0$ and that the equality holds if and only if $\eta_t \cdot v \equiv 0$. By Lemma 1, this last equality occurs if and only if f is constant, which means that B is a disk. $\square$

REFERENCES

[1] M. S. Ashbaugh, *Isoperimetric and universal inequalities for eigenvalues*, in Spectral Theory and Geometry (Edinburgh, 1998), London Math. Soc. Lecture Note Ser. 273, Cambridge University Press, Cambridge, UK, 1999, pp. 95–139.

[2] M. S. Ashbaugh, *Open problems on eigenvalues of the Laplacian*, in Analytic and Geometric Inequalities and Applications, Math. Appl. 478, Kluwer Academic, Dordrecht, 1999, pp. 13–28.

[3] S. Chanillo, D. Grieser, M. Imai, K. Kurata, and I. Ohnishi, *Symmetry breaking and other phenomena in the optimization of eigenvalues for composite membranes*, Comm. Math. Phys., 214 (2000), pp. 315–337.

[4] A. El Soufi and S. Ilias, *Domain deformations and eigenvalues of the Dirichlet Laplacian in a Riemannian manifold*, Illinois J. Math., to appear.

[5] G. Faber, *Beweis, dass unter allen homogenen membranen von gleicher fläche und gleicher spannung die kreisförmige den tiefsten grundton gibt*, Sitz. Ber. Bayer. Akad. Wiss., (1923), pp. 169–172.

[6] P. R. Garabedian and M. Schiffer, *Convexity of domain functionals*, J. Anal. Math., 2 (1953), pp. 281–368.

[7] J. Hadamard, *Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées. Œuvres de J. Hadamard. Tome* II, Éditions du Centre National de la Recherche Scientifique, Paris, 1968, pp. 515–631.

[8] E. M. Harrell II, P. Kröger, and K. Kurata, *On the placement of an obstacle or a well so as to optimize the fundamental eigenvalue*, SIAM J. Math. Anal., 33 (2001), pp. 240–259.

[9] A. Henrot, *Minimization problems for eigenvalues of the Laplacian*, J. Evol. Equ., 3 (2003), pp. 443–461.

[10] J. Hersch, *The method of interior parallels applied to polygonal or multiply connected membranes*, Pacific J. Math., 13 (1963), pp. 1229–1238.

[11] S. Kesavan, *On two functionals connected to the Laplacian in a class of doubly connected domains*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 617–624.

[12] E. Krahn, *Über eine von Rayleigh formulierte Minimaleigenschaft des Kreises*, Math. Ann., 94 (1925), pp. 97–100.

[13] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Springer-Verlag, New York, 1984 (corrected reprint of the 1967 original).

[14] A. G. Ramm and P. N. Shivakumar, *Inequalities for the minimal eigenvalue of the Laplacian in an annulus*, Math. Inequal. Appl., 1 (1998), pp. 559–563.

[15] F. Rellich, *New results in the perturbation theory of eigenvalue problems*, in Simultaneous Linear Equations and the Determination of Eigenvalues, National Bureau of Standards Appl. Math. Ser. 29, U.S. Government Printing Office, Washington, D.C., 1953, pp. 95–99.

[16] M. Schiffer, *Hadamard's formula and variation of domain-functions*, Amer. J. Math., 68 (1946), pp. 417–448.

[17] J. Serrin, *A symmetry problem in potential theory*, Arch. Rational Mech. Anal., 43 (1971), pp. 304–318.

SIAM J. Math. Anal.
Vol. 39, No. 4, pp. 1120–1134

A NEW LOOK AT EQUILIBRIA IN STEFAN-TYPE PROBLEMS IN THE PLANE*

PIOTR B. MUCHA† and PIOTR RYBKA†

Abstract. We study steady states of Stefan-type problems in the plane with the Gibbs–Thomson correction involving a general anisotropic energy density function. By a local analysis we prove the global result showing that the solution is the Wulff shape. The key element is a stability result which enables us to approximate singular models by regular ones.

Key words. Stefan-type problems, singular interfacial energies, Wulff shape, constant weighted mean curvature, steady states, monotone operators

AMS subject classifications. 53C21 (74E15)

DOI. 10.1137/060677124

1. Introduction.

An important aspect of studying dynamical systems is determining steady states and their stability properties. We have in mind a class of models of phase transitions involving the Gibbs–Thomson correction on the interface. They include the Stefan (see [Lu], [AW], [CR], [Ra], [FR], [GR1]) and Hele–Shaw problems (see [DE], [A]). In special cases the modeling system takes the following one-phase quasi-stationary form:

$$\begin{array}{lll} \Delta p = 0 & \text{in} & \Omega(t), \\ p = \kappa_\gamma & \text{on} & \partial\Omega(t), \\ \frac{\partial p}{\partial \mathbf{n}} = -V & \text{on} & \partial\Omega(t). \end{array} \tag{1.1}$$

This is a free boundary problem where we seek the time evolution of an open set $\Omega(t)$ in $\mathbb{R}^n$, $\mathbf{n}$ is the normal to the boundary, and V is the normal velocity of the interface $\partial\Omega(t)$. Here κ_γ is the weighted mean curvature, which is the most important object for us; it will be explained below and in more detail in section 3.

This system is augmented with initial data for Ω. The interpretation of p depends upon the phenomenon we wish to model with (1.1). In the Stefan problem this is the temperature; in the Hele–Shaw problem p is the fluid pressure; in the tumor growth model (see [FR]) p is the internal pressure of the proliferating tissue; in the crystal growth from vapor (see [GR1]) p is the supersaturation of the diffusing water vapor.

We want to stress the important fact that κ_γ in (1.1) is the weighted (anisotropic) mean curvature, i.e.,

$$\kappa_\gamma = \mathrm{div}_{\partial\Omega}(\nabla_\xi \bar{\gamma}|_{\xi=\mathbf{n}}). \tag{1.2}$$

*Received by the editors December 8, 2006; accepted for publication (in revised form) July 24, 2007; published electronically November 7, 2007. The present work has been partially supported by Polish KBN grant No. 1 P03A 037 28.

http://www.siam.org/journals/sima/39-4/67712.html

†Instytut Matematyki Stosowanej i Mechaniki, Uniwersytet Warszawski, ul. Banacha 2, 02-097 Warszawa, Poland (p.mucha@mimuw.edu.pl, p.rybka@mimuw.edu.pl). The first author has been partly supported by EC FP6 Marie Curie ToK program SPADE2, MTKD-CT-2004-014508, and Polish MNiSW SPB-M. The second author thanks the Institute for Mathematics and its Applications at the University of Minnesota for its hospitality, where a part of the research for this paper was performed. The second author was partially supported by a grant of the Office of Naval Research, grant N00014-05-1-4020.

This formula is correct for sufficiently smooth $\bar{\gamma}$. On the other hand, the case of low smoothness $\bar{\gamma}$ and singular $\partial\Omega$ will require special considerations.

Function $\bar{\gamma}$ is related to the anisotropy of the modeled process. Here we assume that function $\bar{\gamma} : \mathbb{R}^n \to \mathbb{R}$ is one-homogeneous, $\bar{\gamma}(x) > 0$ for $x \neq 0$, i.e.,

$$\bar{\gamma}(x) = |x|\gamma\left(\frac{x}{|x|}\right) \quad \text{if } x \neq 0, \qquad \bar{\gamma}(0) = 0, \tag{1.3}$$

while function γ depends essentially on the orientation, i.e., $\gamma : S^{n-1} \to \mathbb{R}$. In the terminology of [Gu, Chapter 7], function γ is called the interfacial energy, while $\bar{\gamma}$ is the extended energy.

Our goal is to study static solutions to (1.1), i.e.,

$$\begin{array}{lll} \Delta p = 0 & \text{in} & \Omega, \\ p = \operatorname{div}_{\partial\Omega}(\nabla_\xi \bar{\gamma}|_{\xi=\mathbf{n}}) & \text{on} & \partial\Omega, \\ \frac{\partial p}{\partial \mathbf{n}} = 0 & \text{on} & \partial\Omega. \end{array} \qquad |\Omega| \text{ is given,} \tag{1.4}$$

Momentarily, we are going to explain why the measure of Ω has to be fixed. By $(1.1)_{1,3}$ we conclude that $\int_{\partial\Omega(t)} V\, d\mathcal{H}^{n-1} = 0$, and hence the measure of $\Omega(t)$ is constant. Additionally from $(1.4)_{1,3}$ we conclude that $p \equiv \text{const}$, and thus, we get

$$\operatorname{div}_{\partial\Omega}(\nabla_\xi \bar{\gamma}|_{\xi=\mathbf{n}}) = \text{const} \qquad \text{and} \qquad |\Omega| \text{ is given.}$$

By a simple rescaling we arrive at a geometric problem expressed as a differential equation

$$\operatorname{div}_{\partial\Omega}(\nabla_\xi \bar{\gamma}|_{\xi=\mathbf{n}}) = 1. \tag{1.5}$$

At this moment we have to specify the assumptions on $\bar{\gamma}$. A minimal hypothesis, besides (1.3), which we impose is the *convexity* of $\bar{\gamma}$. Thus, we immediately conclude that $\bar{\gamma}$ is Lipschitz continuous.

Under such broad assumptions on $\bar{\gamma}$, making sense out of (1.5) requires some work. For example, in [BNP] and [GR2] a separate variational problem was considered for the definition of κ_γ. It is our desire to consider quite a general, as far as the smoothness is concerned, surface energy density function γ.

In our paper we consider a two-dimensional case of (1.5). That is, we look for a curve Γ which locally has a constant weighted mean curvature κ_γ. We show the existence of such a curve, which turns out to be the boundary of a region (in particular Γ is closed).

It is possible to adopt an energy point of view, interpret (1.5) as a critical point of surface energy functional $E(S) = \int_S \gamma(\mathbf{n}(x))\, d\mathcal{H}^{n-1}(x)$, and study its minimizers under the volume constraint. In fact this approach has been carried out; see [T], [FM], [Pa], [Mo], and references therein. In [T] Taylor gives the first rigorous proof of Wulff's theorem, stating that the only minimizer of E under the volume constraint is the Wulff shape. Later, various proofs of this result were found; see [FM] and references therein. Palmer [Pa] shows that the only stable smooth critical point of E is the Wulff shape. Morgan [Mo] studied equilibria of E, and he dropped the stability and smoothness assumptions. He showed that the only equilibrium is the Wulff shape.

An important assumption in the above papers is that they deal only with manifolds without boundary (e.g., closed curves). Here we adopt a different view, which we may call a local one. Namely, we can regard (1.5) as a *locally* defined differential equation. By this we mean that if (1.5) is given some appropriate data, then there exists a solution which is defined in a neighborhood of the data. Our goal, however, is *global*, to show that one can glue up those local solutions, and, despite the apparent freedom in choosing the data for the equation, one can obtain a uniquely defined geometric object: a closed manifold, in this case a closed curve.

We stress that we set a restricted goal to consider only the two-dimensional case. This leads to a significant simplification of (1.5) and the notion of its solution. However, thanks to this simplification we will be able to show interesting qualitative properties of obtained solutions. A solution to this equation will be a curve of class W^1_p, whose curvature satisfies (1.5) in an appropriate sense explained later; see section 3.

We may now state our main result.

THEOREM 1.1. *Let us suppose that $\bar\gamma : \mathbb{R}^2 \to \mathbb{R}$ is a one-homogeneous, convex function. Then there exists a solution to* (1.5). *Moreover, the solution defines a closed curve Γ, and it is unique up to a translation; namely, it is the Wulff shape of $\bar\gamma$.*

Let us stress that compared to the literature we have already mentioned, the closedness of curve Γ is a *conclusion* of our analysis, not its assumption. Moreover, we consider the weakest possible regularity assumptions on γ, i.e., only convexity of $\bar\gamma$. A general notion of the solution is introduced in the definition of section 3 by (3.13).

In order to prove this theorem, we first consider the regular case (we will make clear what this means for us). Careful analysis gives us a hint how to proceed for general $\bar\gamma$. We will see that it is easier to show the existence of local solutions to (1.5) than to show that they can be glued up to form a closed curve. The last task is simpler for regular curves, because we can use the power of the classical differential geometry. Our goal is reached through an appropriate stability theorem, which is our second important result.

THEOREM 1.2. *Let us assume that $1 \le p < \infty$ and $\bar\gamma, \bar\gamma_\epsilon : \mathbb{R}^2 \to \mathbb{R}$ are convex, one-homogeneous and $\gamma_\epsilon : S^1 \to \mathbb{R}$ are defined by* (1.3) *with $\bar\gamma$ replaced by $\bar\gamma_\epsilon$. We also assume that Γ_ϵ is a solution to*

$$\mathrm{div}_{\partial\Omega}(\nabla_\xi \bar\gamma_\epsilon|_{\xi=\mathbf{n}}) = 1$$

and Γ is a solution to (1.5). *If $\gamma_\epsilon \to \gamma$ in $W^1_p(S^1)$ as $\epsilon \to 0$, then possibly after a translation of Γ_ϵ it is true that $\Gamma_\epsilon \to \Gamma$ in W^1_p as submanifolds in $\mathbb{R}^2$.*

Let us first explain that convergence of Γ_ϵ means convergence of curves locally treated as graphs. The $W^1_p(S^1)$-space is the standard Sobolev space of functions, defined on the unit circle $S^1 = \{x \in \mathbb{R}^2 : |x| = 1\}$, integrable with its derivative with the pth power.

The last result enables us to approximate any low regularity problem by its suitable regularization. A crucial point here is the choice of a proper notion of convergence. At first it seems that the measures are the best spaces, because the second derivative of a convex function is in this space, or, equivalently, the first derivative belongs to the TV-space. However, from our point of view this setting does not appear to be appropriate, because we are not able to find any smooth approximations for the general case. Hence, we investigate the L_p-approach which implies a weaker topology, but it provides sufficient information. The method developed here gives us a tool to analyze the behavior of possible singularities appearing in our problem.

Let us draw a corollary from Theorem 1.2. If $\|x\|_p$ is the usual p-norm in $\mathbb{R}^2$, i.e., $\|x\|_p = (|x_1|^p + |x_2|^p)^{1/p}$ for $p < \infty$, and $\|x\|_\infty = \max\{|x_1|, |x_2|\}$, then we define γ_p

by means of (1.3) as $\gamma_p(x) = \|x\|_p$ for $x \in S^1$. It is a well-known fact that Γ_p, which is a solution to (1.5) with γ_p replacing γ, is given as a Wulff shape of $\|\cdot\|_p$; i.e., Γ_p is the unit ball of the norm $\|\cdot\|_q$, where $\frac{1}{p} + \frac{1}{q} = 1$. Thus, $\Gamma_p = \{\|x\|_q = 1\}$. We can easily see that, for any $r < \infty$, if $\gamma_p \to \gamma_\infty$ in W^1_r as $p \to \infty$, then $\Gamma_p \to \Gamma_\infty$ in W^1_r as submanifolds in $\mathbb{R}^2$.

We briefly describe the content of the paper. In section 2, we recall some well-known facts from the classical differential geometry and explain the meaning of (1.5). Moreover, we prove there Theorem 1.1 in the case of regular γ. This proof will be the starting point for further considerations. Next, we reformulate the problem as a differential inclusion in the case of general γ. In section 4 are the main results: Theorems 1.1 and 1.2 are proved. Finally, we analyze the qualitative properties of the obtained solutions.

2. Preliminary analysis. In this section we recall some facts and introduce further notation. Namely, we shall start by writing

$$\bar{\gamma}(x) = r\gamma(\varphi),$$

where $r = |x|$ and φ are defined by the relation $x = r(\cos\varphi, \sin\varphi)$.

Let us suppose that $s \mapsto \mathbf{x}(s)$ is an arc-length parameterization of a given smooth curve Γ, and then $\mathbf{t}(s) = \frac{d\mathbf{x}}{ds}(s)$ is a unit vector tangent to Γ. If Γ is the boundary of Ω, then we rewrite the left-hand side (LHS) of (1.5) to obtain

$$\operatorname{div}_{\partial\Omega}(\nabla_\xi \bar{\gamma}|_{\xi=\mathbf{n}}) = \mathbf{t} \cdot \frac{\partial}{\partial \mathbf{t}}(\nabla\bar{\gamma}(\mathbf{n})) \equiv \mathbf{t}(s) \cdot \frac{d}{ds}[\nabla\bar{\gamma}(\mathbf{n}(s))]. \tag{2.1}$$

While keeping in mind (2.1), we want to rewrite system (1.5) in a suitable coordinate system. For this purpose, we use the normal angle and subsequently the angle parameterization of Γ. Let us suppose first that $\mathbf{n}$ is a unit vector normal to Γ such that moving frame $\{\mathbf{n}, \mathbf{t}\}$ is a positively oriented basis of $\mathbb{R}^2$. Then the *normal angle* $\varphi(s)$ is defined as a smooth function of s through

$$\begin{aligned} \mathbf{n} = \mathbf{n}(s) = (\cos\varphi(s), \sin\varphi(s)), \\ \mathbf{t} = \mathbf{t}(s) = (-\sin\varphi(s), \cos\varphi(s)). \end{aligned} \tag{2.2}$$

We will refer to the range of the function $s \mapsto \varphi(s)$ as the *angle set.* With the help of this function we define a parameterization of our curve by

$$\mathbf{x}(s) = \int_{s_0}^{s} \mathbf{t}(\varphi(t))dt + \mathbf{v_0} = \int_{s_0}^{s} (-\sin\varphi(t), \cos\varphi(t))\, dt + \mathbf{v_0}, \tag{2.3}$$

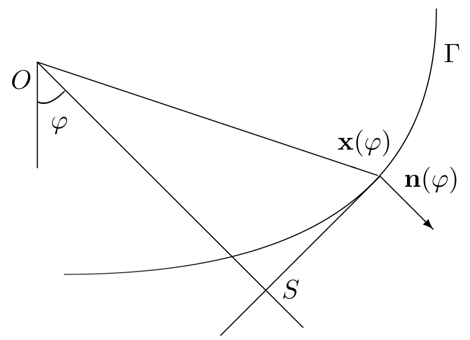
where $\mathbf{v_0}$ is a fixed point in $\mathbb{R}^2$. Let us write $\Gamma = \mathbf{x}(\mathbb{R})$. This formula is a direct result of the fact that s is an arc-length parameter.

The function

$$\frac{d\varphi}{ds}(s) = \kappa(s) \tag{2.4}$$

is the Euclidean curvature of Γ. It obeys the Frenet formulas: $\frac{d\mathbf{n}}{ds} = \kappa\mathbf{t}$, $\frac{d\mathbf{t}}{ds} = -\kappa\mathbf{n}$. In order to proceed, we assume for the moment that γ is at least C^2-smooth. Then (1.5) becomes

$$\left[\frac{d}{ds}\varphi\right][\mathbf{t} \cdot K(\mathbf{n}) \cdot \mathbf{t}] = 1, \tag{2.5}$$



$$P_0(\varphi) = \mathbf{x}(\varphi) \cdot \mathbf{n}(\varphi) = |OS|.$$

FIG. 1.

where

$$K = \begin{bmatrix} \bar{\gamma}_{11} & \bar{\gamma}_{12} \\ \bar{\gamma}_{21} & \bar{\gamma}_{22} \end{bmatrix} = \frac{1}{r}\left(\frac{d^2}{d\varphi^2}\gamma(\varphi) + \gamma(\varphi)\right) e_\varphi \otimes e_\varphi = d^2\bar{\gamma}(r\mathbf{n}) \tag{2.6}$$

and $e_\varphi(\varphi(s)) = \mathbf{t}(\varphi(s))$. Convexity of $\bar{\gamma}$ is equivalent to

$$\frac{d^2}{d\varphi^2}\gamma(\varphi) + \gamma(\varphi) \geq 0. \tag{2.7}$$

In the simplest case we will rule out the possibility of equality above. Let us stress that, because of one-homogeneity of $\bar{\gamma}$, this function cannot be strictly convex in the usual sense.

We shall call a smooth curve *strictly convex* if its curvature never vanishes (in the literature such curves are often called convex). Let us notice that for strictly convex curves an *angle parameterization* $\varphi \mapsto \mathbf{x}(\varphi)$ is possible, because the function $s \mapsto \varphi(s)$ is strictly increasing.

Finally, we shall call a surface energy density function γ *strictly stable at* φ_0 if $\frac{d^2\gamma}{d\varphi^2}(\varphi_0) + \gamma(\varphi_0) > 0$, and γ will be called *strictly stable* if $\frac{d^2\gamma}{d\varphi^2}(\varphi) + \gamma(\varphi) > 0$ holds for all φ. The last condition is a substitute for strict convexity of $\bar{\gamma}$. We shall call a surface energy function γ *regular* if it is smooth and strictly stable.

We expect that (2.5) yields a well-posed differential equation having local solutions. The main difficulty will be associated with proving that solutions can be glued up to form a closed curve. Our tool in solving this problem will be the classical concept of the support function. Let us supposes that Θ is the angle set of a curve Γ for a point O. The *support function* $P_0 : \Theta \to \mathbb{R}$ of Γ is defined by the formula (see Figure 1).

The meaning of $P_0(\varphi)$ is the distance from the tangent to Γ at $\mathbf{x}(\varphi)$ to the origin. One can easily check using the Frenet formula, (2.4), and $\frac{d\mathbf{n}}{d\varphi} = \mathbf{t}$ that P_0 satisfies the following equations:

$$\mathbf{x}(\varphi) = P_0(\varphi)\mathbf{n}(\varphi) + \frac{dP_0}{d\varphi}(\varphi)\mathbf{t}(\varphi), \qquad \frac{d^2P_0}{d\varphi^2}(\varphi) + P_0(\varphi) = \frac{1}{\kappa} \tag{2.8}$$

(see also [Bl, section 94], [Gu, equation (1.10)]).

LEMMA 2.1 ([Gu, Lemma 1A(b)]). *If Γ is a convex curve with curvature $\kappa(\varphi)$, and if P_0 is any solution to $(2.8)_2$ on the angle set of Γ, then $(2.8)_1$ is the angle parameterization of a curve which differs from Γ by at most a translation.*

Proof. Identity $(2.8)_1$ defines a curve uniquely up to a translation. Hence assuming that the curves—mentioned in the lemma—are different, we conclude that they have to intersect at least in one point. Let us consider the difference u between P_0 and the support function of Γ. It satisfies $\frac{d^2u}{d\varphi^2} + u = 0$. The general solution to this equation is $u(\varphi) = a_1 \cos\varphi + a_2 \sin\varphi \equiv \mathbf{a} \cdot \mathbf{n}(\varphi)$. However, $(2.8)_1$ determined two parameters for the intersection, and hence $a_1 = a_2 = 0$. Our claim follows. ◻

We are now ready for our first local result.

LEMMA 2.2. *Let us suppose that γ is smooth and strictly stable. Then there exists a unique local solution to* (2.5) *augmented with the initial condition $\varphi(s_0) = \varphi_0$.*

Proof. The assumption of the strict stability of γ implies that there exist positive numbers a_1 and a_2 such that, for each point $p = (x(s), y(s))$ of curve Γ defined by (2.3), matrix K satisfies

$$a_1|\xi|^2 \le \xi \cdot K(p) \cdot \xi \le a_2|\xi|^2 \quad \text{for any } \xi \in T\Gamma_p, \tag{2.9}$$

where $T\Gamma_p$ denotes the tangent space to Γ at point p.

In this case the existence as well as the uniqueness of solutions are obvious since we reduce the problem to the following elementary ordinary differential equation:

$$\frac{d}{ds}\varphi(s) = (\mathbf{t} \cdot K(\mathbf{n}(\varphi(s))) \cdot \mathbf{t})^{-1}, \tag{2.10}$$

with initial data $\varphi(s_0) = \varphi_0$.

By the strict convexity of γ constants a_1 and a_2 in (2.9) are prescribed globally, and the solution to (2.10) exists for all $s \in \mathbb{R}$. ◻

Subsequently, we will use the above lemma a number of times. It is the necessary step to establishing our first goal, stated below.

PROPOSITION 2.1. *Let us suppose that γ is regular, i.e., smooth and globally strictly stable. Then solutions to* (2.5) *yield a unique closed C^2-curve, Γ being the boundary of a domain Ω.*

Proof. By Lemma 2.2 we have the local existence as well as the uniqueness of solutions to (2.5) with initial data $\varphi(s_0) = \varphi_0$. Now we are going to show that the solution to (2.10) yield a closed curve Γ. First, let us note that by (2.6) and (2.10) we conclude that

$$\frac{d^2}{d\varphi^2}\gamma(\varphi) + \gamma(\varphi) = \frac{1}{\kappa(\varphi)}. \tag{2.11}$$

Because of the global strict stability of γ, function φ is an angle parameterization of a strictly convex curve. Moreover, due to (2.9) there is an interval $[s_0, s_0 + d)$ such that φ is on $[\varphi_0, \varphi_0 + 2\pi)$; i.e., the angle set of Γ is $[0, 2\pi)$. Hence, by Lemma 2.1 we conclude that P_0 and γ differ by $a_1 \cos\varphi + a_2 \sin\varphi$. Since γ is periodic we deduce that P_0 is periodic as well. Thus, we infer that the curve, whose support function is periodic, must be closed; see $(2.8)_1$ and Lemma 2.1. Now we need to exclude self-intersections of the sought curve Γ. This follows immediately from the fact that $\mathbf{x}(\varphi)$ given by (2.8) is 2π-periodic. ◻

Remarks. Let us stress that Lemma 2.2 yields angle parameterizations which are solutions to a differential equation with arbitrary initial data. We showed that the corresponding geometric object, i.e., a closed curve whose angle parameterization we have constructed, exists, and it is unique. The uniqueness is a consequence of the fact that the support functions coincide (up to a periodic function) with γ. Moreover, if

γ is strictly stable, then the locus of the vector function given by (2.8), i.e.,

$$\mathbf{x}(\varphi) = \gamma(\varphi)\mathbf{n}(\varphi) + \gamma'(\varphi)\mathbf{t}(\varphi), \tag{2.12}$$

is the boundary of the Wulff shape W_γ (see [Gu, Theorem 7P]). We recall that, by definition,

$$W_\gamma = \{p \in \mathbb{R}^2 :\ p \cdot \mathbf{n}(\varphi) \le \gamma(\varphi)\}. \tag{2.13}$$

In other words, we have proved Theorem 1.1 for smooth, globally stable γ's.

Inequality (2.7) has a geometric interpretation. We recall that the *Frank diagram* F_γ of a one-homogeneous function $\bar{\gamma}$ is defined by

$$F_\gamma = \{p \in \mathbb{R}^2 :\ \bar{\gamma}(p) = 1\} = \{(r, \varphi) :\ r = \gamma^{-1}(\varphi)\}. \tag{2.14}$$

It is well-known fact that convexity of F_γ is equivalent to convexity of $\bar{\gamma}$ (e.g., see [Gu, Theorem 7B]). Moreover, one can see that the strict stability of γ is equivalent to smoothness and convexity of F_γ.

We are now ready to discuss the weighted mean curvature (WMC) equation for a general anisotropy.

3. The general case of the WMC equation for curves. The goal of this section is to restate problem (1.5) for general γ and define a generalization of the solution—see (3.13).

The approach presented in section 2 is not suitable for γ, which is not of class C^2. For this purpose we return to (2.1) and rewrite it again. Namely, we notice that for smooth γ (1.5) is equivalent to

$$\frac{d}{ds}[(\nabla_\xi \bar{\gamma}|_{\xi=\mathbf{n}}) \cdot \mathbf{t}] - (\nabla_\xi \bar{\gamma}|_{\xi=\mathbf{n}}) \frac{d}{ds}\mathbf{t} = 1, \tag{3.1}$$

but due to (2.2) we see that $\frac{d}{ds}\mathbf{t} = -(\frac{d}{ds}\varphi)\mathbf{n}$; hence, the second term of the LHS of (3.1) reads

$$-(\nabla_\xi \bar{\gamma}|_{\xi=\mathbf{n}}) \frac{d}{ds}\mathbf{t} = \left(\frac{d}{ds}\varphi\right)(\nabla_\xi \bar{\gamma}|_{\xi=\mathbf{n}}) \cdot \mathbf{n}; \tag{3.2}$$

however, by one-homogeneity of $\bar{\gamma}$—see (1.3)—we have $(\nabla_\xi \bar{\gamma}|_{\xi=\mathbf{n}}) \cdot \mathbf{n} = \bar{\gamma}(\mathbf{n})$. Hence we get

$$-(\nabla_\xi \bar{\gamma}|_{\xi=\mathbf{n}})\left(\frac{d}{ds}\mathbf{t}\right) = \left(\frac{d}{ds}\varphi\right)\bar{\gamma}(\mathbf{n}(\varphi)) = \frac{d}{ds}\int_\vartheta^{\varphi(s)} \bar{\gamma}(\mathbf{n}(t))dt, \tag{3.3}$$

where ϑ will be fixed later. Thus, (3.1) takes the following form:

$$\frac{d}{ds}\left[(\nabla_\xi \bar{\gamma}|_{\xi=\mathbf{n}}) \cdot \mathbf{t} + \int_\vartheta^{\varphi(s)} \bar{\gamma}(\mathbf{n}(t))dt\right] = 1. \tag{3.4}$$

We notice that the definition of the normal vector (2.2) implies that $(\nabla_\xi \bar{\gamma}|_{\xi=\mathbf{n}}) \cdot \mathbf{t} = \frac{d}{d\varphi}\bar{\gamma}(\mathbf{n}(\varphi))$. Keeping this in mind we rewrite (3.4) as follows:

$$\frac{d}{ds}\frac{d}{d\varphi}\left[\bar{\gamma}(\mathbf{n}(\varphi)) + \int_\vartheta^{\varphi} d\psi \int_\vartheta^{\psi} \bar{\gamma}(\mathbf{n}(t))dt\right] = 1 \tag{3.5}$$

or in a concise form

$$\frac{d}{ds}\frac{d}{d\varphi}I^{\gamma}_{\vartheta}(\varphi) = 1, \tag{3.6}$$

where

$$I^{\gamma}_{\vartheta}(\varphi) = \bar{\gamma}(\mathbf{n}(\varphi)) + \int_{\vartheta}^{\varphi} d\psi \int_{\vartheta}^{\psi} \bar{\gamma}(\mathbf{n}(t))dt. \tag{3.7}$$

We will drop the superscript for fixed γ.

Equation (3.6) does not make sense for $\bar{\gamma}$ which are only Lipschitz continuous. However, we know that $\bar{\gamma}$ is convex, and this property of $\bar{\gamma}$ is equivalent to the following distributional relation:

$$\frac{d^2}{d\varphi^2}\gamma(\varphi) + \gamma(\varphi) \geq 0 \quad \text{in } \mathcal{D}'(S^1). \tag{3.8}$$

Let us make the following observation.

PROPOSITION 3.1. *If one-homogeneous function $\bar{\gamma} : \mathbb{R}^2 \to \mathbb{R}$ is convex, then the function*

$$\mathbb{R} \ni \varphi \to I_{\vartheta}(\varphi) \in \mathbb{R}$$

is also a convex function for fixed ϑ.

Proof. If $\bar{\gamma}$ is as above, i.e., $\bar{\gamma}(x) = |x|\gamma(\varphi)$, where $x = |x|(\cos\varphi, \sin\varphi)$, then by (2.6) convexity of $\bar{\gamma}$ is equivalent to inequality $\gamma'' + \gamma \geq 0$ in the sense of distributions (see (3.8)). Once we notice that $\frac{\partial^2}{\partial\varphi^2}I_{\vartheta}(\varphi) = \gamma'' + \gamma$, our claim follows. ∎

Thus, the subdifferential $\partial_{\varphi}I_{\vartheta}$ of a convex function I_{ϑ} is always well-defined. Hence, our task consists of finding a suitable section of $\partial_{\varphi}I_{\vartheta}$, and we can formally write (3.5) as the following inclusion:

$$\frac{d}{ds}\partial_{\varphi}\left[\bar{\gamma}(\mathbf{n}(\varphi)) + \int_{\vartheta}^{\varphi} d\psi \int_{\vartheta}^{\psi} \bar{\gamma}(\mathbf{n}(t))dt\right] \ni 1. \tag{3.9}$$

We will make it precise below. Let us first make some observations about I_{ϑ}. If γ is smooth, then $\frac{d}{d\varphi}I_{\vartheta_1}$ and $\frac{d}{d\varphi}I_{\vartheta_2}$ differ by a constant, i.e.,

$$\frac{d}{d\varphi}I_{\vartheta_1}(\varphi) = \frac{d}{d\varphi}I_{\vartheta_2}(\varphi) + \int_{\vartheta_1}^{\vartheta_2} \gamma(t)\, dt. \tag{3.10}$$

Let us notice that this identity will remain valid if γ is Lipschitz continuous and we replace $\frac{d}{d\varphi}$ with ∂_{φ}. Mere continuity of γ yields

$$I_{\vartheta}(2\pi + \varphi) = I_{\vartheta}(\varphi) + I_{\vartheta}(2\pi + \vartheta) - \gamma(\vartheta) + (\varphi - \vartheta)\int_{\vartheta}^{2\pi+\vartheta} \gamma(t)\, dt. \tag{3.11}$$

Indeed, for $\varphi \geq \vartheta$ we have

$$I_{\vartheta}(2\pi+\varphi) - I_{\vartheta}(2\pi+\vartheta) = \gamma(\varphi) - \gamma(\vartheta) + \int_{2\pi+\vartheta}^{2\pi+\varphi}\int_{\vartheta}^{2\pi+\vartheta} \gamma(t)\, dt d\psi + \int_{2\pi+\vartheta}^{2\pi+\varphi}\int_{2\pi+\vartheta}^{\psi} \gamma(t)\, dt d\psi.$$

Hence, (3.11) follows.

Suppose again that γ is smooth, and then we can infer from (3.10) that by an appropriate choice of ϑ we may achieve $\varphi(\bar{s}) = \bar{\varphi}$ for arbitrary $\bar{\varphi}$, $\bar{s} \in \mathbb{R}$.

Furthermore, let us integrate (3.5) over $[\bar{s}, s]$, where $\bar{s}$ is an arbitrary parameter and $s - \bar{s}$ is less than the length of the curve Γ we are looking for.

After integration we obtain

$$\frac{d}{d\varphi} I_\vartheta(\varphi(s)) - \frac{d}{d\varphi} I_\vartheta(\bar{\varphi}) = s - \bar{s}, \tag{3.12}$$

where $\bar{\varphi} = \varphi(\bar{s})$ may be chosen at will.

If γ is smooth, then the LHS of (3.12) is well-defined for all $\varphi \in \mathbb{R}$. However, this is no longer the case if γ is merely Lipschitz continuous. For such a γ, the LHS of (3.12) is well-defined only a.e.

Now we are in a position to define a notion of a generalized solution to (1.5) with minimal assumptions on $\bar{\gamma}$.

DEFINITION 3.1. *If γ is one-homogeneous and convex, then by a* solution to (1.5) *we mean a closed curve Γ whose angle parameterization $\varphi(\cdot)$ is a monotone (increasing) multivalued function, which can be treated locally as an L_1 function, and the following differential inclusion is valid:*

$$\partial_\varphi I_\vartheta(\varphi) \ni s - \bar{s} + s^\star \qquad a.e., \tag{3.13}$$

with initial data $\varphi|_{s=\bar{s}} = \bar{\varphi}$ and $s^\star \in \partial_\varphi I_\vartheta(\bar{\varphi})$. The inclusion (3.13) *holds for such s that $\varphi(s)$ is single-valued.*

Let us notice that (3.10) (with $\frac{d}{d\varphi}$ replaced by ∂_φ) again yields the possibility of choosing ϑ so that $\bar{s} = s^\star \in \partial_\varphi I_\vartheta(\bar{\varphi})$ for any $\bar{s}$, $\bar{\varphi} \in \mathbb{R}$. Thus, (3.13) will read

$$\partial_\varphi I_\vartheta(\varphi) \ni s. \tag{3.14}$$

We note an expected result.

COROLLARY 3.1. *Let us suppose that γ is smooth and a function $\varphi : \mathbb{R} \to \mathbb{R}$ such that $\varphi(\bar{s}) = \bar{\varphi}$ is a solution to* (3.13)*, and then φ is a solution to* (3.12)*. Moreover, if in addition γ is globally stable, then the solution is unique.*

Proof. This fact is obvious, once we realize that, for smooth γ, the function $I_\vartheta(\cdot)$ is also smooth and

$$\partial_\varphi I_\vartheta(\varphi) = \left\{ \frac{d}{d\varphi} I_\vartheta(\varphi) \right\},$$

where $\frac{d}{d\varphi} I_\vartheta(\varphi)$ is the usual derivative of $I_\vartheta(\varphi)$. The uniqueness of the solutions follows from strict monotonicity of $\partial_\varphi I_\vartheta(\varphi)$, provided that γ is strictly stable. □

4. The general existence and stability. In this section we prove Theorems 1.1 and 1.2. We want to proceed in a manner similar to that of section 2. We first show the local existence of solutions to (3.13), which replaces (2.5) for nonsmooth γ. However, in order to complete the proof of Theorem 1.1, we have to show that the solutions we have constructed yield closed curves. Indeed, this goal is achieved through an approximation procedure. We show for this purpose Theorem 1.2. It states that solutions to (3.13) depend continuously on $\bar{\gamma}$. Thus, the approximation of any energy density function γ by a sequence of smooth, strictly stable energy density functions γ_ϵ yields the desired result.

We begin with an analog of Lemma 2.2 for a nonsmooth energy density function.

PROPOSITION 4.1. *Let us suppose that $\bar{\gamma}$ is convex and one-homogeneous and ϑ is fixed. Then*

(a) *the multivalued operator* $\partial I_\vartheta : \mathbb{R} \to \mathbb{R}$ *is maximal monotone;*

(b) *for any* $s \in \mathbb{R}$ *there exists* $\varphi \in \mathbb{R}$ *a solution to* $\partial I_\vartheta(\varphi) \ni s$.

Proof. Part (a) is a conclusion from the general theory; see, e.g., [Ba] or [Br]. (b) By (3.11) we notice

$$\partial I_\vartheta(2\pi + \varphi) = \partial I_\vartheta(\varphi) + L(\gamma),$$

where $L(\gamma) = \int_\vartheta^{2\pi+\vartheta} \gamma(t)\, dt > 0$. We recall that γ is 2π-periodic, and thus $L(\gamma)$ is independent from θ. Since $R(\partial I_\vartheta)$, the range of ∂I_ϑ is connected and contains a sequence which is neither bounded from below nor from above, and we conclude that $R(\partial I_\vartheta)$ is equal to $\mathbb{R}$. ☐

Once we decide upon $\bar{s}$ and $\bar{\varphi}$, we fix ϑ, and we will drop for notational convenience the subscript ϑ. The choice of value $\bar{s}$ which is the beginning of counting the length of the solution curve seems irrelevant; however, it is not quite so. We will explain it momentarily. Since the function $\varphi \mapsto I(\varphi)$ is convex, then it has one-sided derivatives everywhere. They are equal; i.e., I is differentiable at all points of $\mathbb{R}$ except at most countably many where $\frac{d^- I}{d\varphi}(\psi) < \frac{d^+ I}{d\varphi}(\psi)$. In other words, the derivative $\frac{dI}{d\varphi}$ is well-defined a.e., and its graph $\Gamma(\frac{dI}{d\varphi})$ is a subset of ∂I. Since ∂I is maximal monotone, so is its inverse graph $(\partial I)^{-1}$. If we now consider any function $h : \mathbb{R} \to \mathbb{R}$, with $h(x) \in (\partial I)^{-1}(x)$, i.e., $\Gamma(h) \subset (\partial I)^{-1}$, then it is monotone and hence continuous, except possibly a countable number of points where h suffers jumps. For this reason, any two such functions differ on a set of measure zero and define a unique element of $L_{p(loc)}(\mathbb{R})$. This observation permits us to identify (when necessary) ∂I or $(\partial I)^{-1}$ with monotone functions defining unique elements of $L_{p(loc)}(\mathbb{R})$.

We are now ready for a definition. If we fix $\vartheta \in \mathbb{R}$ and a one-homogeneous convex γ, then we shall call $s^\star \in \mathbb{R}$ *a regular point of* I_ϑ^γ, if it is a continuity point of any function h defined on $\mathbb{R}$, such that $\Gamma(h) \subset (\partial I_\vartheta^\gamma)^{-1}$. This scant continuity of $(\partial I_\vartheta^\gamma)^{-1}$ will play an important role.

We stress that regularity of $s^\star$ depends upon ϑ; i.e., $s^\star$ may cease to be regular if we change ϑ. In particular, we may assume that in (3.13) $s_0 = s^\star - \bar{s}$ is a regular point, possibly after adjusting ϑ and $\bar{\varphi}$. In this case we obtain a unique definition of the solution

$$\varphi(\cdot) = (\partial I)^{-1}(\cdot + s_0), \tag{4.1}$$

where φ is treated as a multivalued function, but we can write $\{\varphi_0\} = (\partial I)^{-1}(s_0)$, because $(\partial I)^{-1}(s_0)$ is a singleton.

Part (b) of Proposition 4.1 immediately yields the existence of a solution to (3.13), that is, a function $\varphi : \mathbb{R} \to \mathbb{R}$, $\varphi(0) = \varphi_0$, which is an angle parameterization of a curve Γ we seek. We recall that this curve is given (up to a translation) by (2.3). We also notice that this formula yields the same result for all L^1 equivalent functions $\varphi(\cdot)$. We should comment on the domain of definition of $\mathbf{x}$, i.e., the length of Γ. We will also explain the meaning of $L(\gamma)$ and the role of regular points. We have just constructed a candidate for a solution.

PROPOSITION 4.2. *Let us suppose that* s_0, φ_0 *are fixed, and then*

(a) $(\partial I_\vartheta^\gamma)^{-1}[s_0, s_0 + L(\gamma)]$ *contains an interval of length* 2π*;*

(b) *if* $0 < b < L(\gamma)$ *and* s_0 *is a regular point of* I_ϑ^γ, *then* $(\partial I_\vartheta^\gamma)^{-1}[s_0, s_0 + b]$ *does not contain any interval of length* 2π.

Proof. Part (a) follows immediately from (3.11).

(b) For a regular point s_0 (3.11) takes the form

$$\frac{dI}{d\varphi}(\varphi_0+2\pi)=\frac{dI}{d\varphi}(\varphi_0)+L(\gamma). \tag{4.2}$$

Hence, $\frac{dI}{d\varphi}(\varphi_0+2\pi)>\frac{dI}{d\varphi}(\varphi_0)+b$, and our claim follows. ∎

We can draw some corollaries.

COROLLARY 4.1. *Let us suppose that s_0 is a regular point of I^{γ}_{ϑ} and curve Γ defined by* (2.3), *where $\varphi(\cdot)$ is given by* (4.1).

(a) *If γ is smooth and strictly stable, then the length of Γ, $d(\Gamma)$, is equal to $L(\gamma)$.*

(b) *If $\{\gamma_\epsilon\}$ is a sequence of smooth, strictly stable functions converging to γ_0 in $W^1_p(S^1)$, then $d(\Gamma_0)$ is equal to $L(\gamma_0)$.*

Proof. (a) By (4.2) we notice that $\varphi(s_0)+2\pi=\varphi(s_0+L(\gamma))$. Thus, we see that $d(\Gamma)=L(\gamma)$.

(b) If $\gamma_\epsilon\to\gamma_0$ in $W^1_p(S^1)$ and $\vartheta_\epsilon\to\vartheta_0$, then the parameterizations given by (2.3) converge as well, i.e., $\mathbf{x}_\epsilon\to\mathbf{x}_0$ in W^1_p. Thus,

$$L(\gamma_0)=\lim_{\epsilon\to0}L(\gamma_\epsilon)=\lim_{\epsilon\to0}d(\Gamma_\epsilon)=d(\Gamma_0),$$

and our claim follows. ∎

The key point is to show that indeed Γ is a closed curve. The main step toward this goal is the following result on the continuous dependence of solutions to (3.13) on the anisotropy $\bar\gamma$. For this purpose we will investigate solutions to (2.1), i.e., curves parameterized by $\mathbf{x}(\cdot)$ (see (2.3)), upon γ. The simple lemma below is our basic tool.

LEMMA 4.1. *Let $1\le p<\infty$ and $\gamma_\epsilon(\cdot)\to\gamma_0(\cdot)$ in $W^1_p(S^1)$; then there exists a sequence $\{\vartheta_\epsilon\}_{\epsilon>0}$ converging to ϑ_0, and for $I_\epsilon:=I^{\epsilon}_{\vartheta_\epsilon}$ we have*

$$(\partial I_\epsilon)^{-1}(\cdot)\to(\partial I_0)^{-1}(\cdot)\quad in\quad L_{p(loc)}(\mathbb{R}). \tag{4.3}$$

Moreover, if s_0 is a regular point of $I^{\gamma_0}_{\vartheta_0}$ (i.e., $\{\varphi_0\}=(\partial I_0)^{-1}(s_0)$), then we can find a sequence $\{s_\epsilon\}$ such that s_ϵ is a regular point of $I^{\gamma_\epsilon}_{\vartheta_\epsilon}$ (i.e., $\{\varphi_\epsilon\}=(\partial I_\epsilon)^{-1}(s_\epsilon)$) and $s_\epsilon\to s_0$.

Proof. Since $I_\epsilon=I^{\gamma_\epsilon}_{\vartheta_\epsilon}$, then for a.e. convergence of $\frac{d}{d\varphi}I_\epsilon$ to $\frac{d}{d\varphi}I_0$ it is necessary that $\vartheta_\epsilon\to\vartheta_0$. To show (4.3) it is enough to recall the definition (3.7) and recall that all $(\partial I_\epsilon)^{-1}$ are monotone. The last fact implies that our functions are continuous except at most countable sets, so the rest of Lemma 4.1 is clear, too. Once we realize that, the remaining details become quite elementary, and they are left to the interested reader. ∎

THEOREM 4.1. *If $1\le p<\infty$ and $\gamma_\epsilon\to\gamma_0$ in $W^1_p(S^1)$, then $\Gamma_\epsilon\to\Gamma_0$ in W^1_p.*

Proof. It is sufficient to show that φ_ϵ (the angle parameterization of Γ_ϵ) converges to φ_0, as long as $\mathbf{v_0}^\epsilon$ converges to $\mathbf{v_0}$; see (2.3). These functions are defined on intervals of changing length $d(\Gamma_\epsilon)$. However, due to convergence of $d(\Gamma_\epsilon)$ to $d(\Gamma_0)$, we may restrict our attention to functions φ_ϵ and φ_0 only on interval $[0,d(\Gamma_0)]$, which is sufficient to prove convergence in the L_p-space. By Lemma 4.1 by a proper choice of ϑ_0 and ϑ_ϵ we can make sure s_ϵ, s_0 are regular points, in particular

$$\{\varphi_0(0)\}=(\partial I_0)^{-1}(s_0),\qquad\{\varphi_\epsilon(0)\}=(\partial I_\epsilon)^{-1}(s_\epsilon),$$

and moreover $s_\epsilon\to s_0$. Subsequently we will identify $(\partial I_\epsilon)^{-1}$ with its L_p-representative.

Let us note that

$$\begin{aligned}\varphi_\epsilon(s)-\varphi_0(s)&=\partial I_\epsilon^{-1}[s+s_\epsilon]-\partial I_0^{-1}[s+s_0]\\&=\partial I_\epsilon^{-1}[s+s_\epsilon]-\partial I_0^{-1}[s+s_\epsilon]+\partial I_0^{-1}[s+s_\epsilon]-\partial I_0^{-1}[s+s_0].\end{aligned} \tag{4.4}$$

Since $\partial I_\epsilon^{-1} \to \partial I_0^{-1}$ in L_p on compact sets, we have

(4.5)

$$||\partial I_\epsilon^{-1}[\,\cdot + s_\epsilon] - \partial I_0^{-1}[\,\cdot + s_\epsilon]||_{L_p(0,d(\Gamma_0))} \leq \left(\int_0^{d(\Gamma_0)} |\partial I_\epsilon^{-1}(s+s_\epsilon) - \partial I_0^{-1}(s+s_\epsilon)|^p\,ds\right)^{1/p}$$
$$= \left(\int_{s_\epsilon}^{d(\Gamma_0)+s_\epsilon} |\partial I_\epsilon^{-1}(s) - \partial I_0^{-1}(s)|^p\,ds\right)^{1/p} \leq ||\partial I_\epsilon^{-1} - \partial I_0^{-1}||_{L_p(s_0-1,d(\Gamma_0)+s_0+1)} \to 0$$

as $\epsilon \to 0$. Next, we note that

$$||\partial I_0^{-1}(\,\cdot + s_\epsilon) - \partial I_0^{-1}(\,\cdot + s_0)||_{L_p(0,d(\Gamma_0))} \to 0, \tag{4.6}$$

which follows from convergence of s_ϵ to s_0 regularity of s_0 and continuity of the shift operator in the L_p-spaces for $p < \infty$.

As a consequence we get

$$\varphi_\epsilon(\cdot) \to \varphi_0(\cdot) \quad \text{in} \quad L_p(0, d(\Gamma_0)). \tag{4.7}$$

Lemma 4.1 implies $\Gamma_\epsilon \to \Gamma_0$ in W_p^1. ☐

Once the above theorem is at hand, we conclude that the limiting curve Γ_0 is closed.

Lemma 4.2. *Let us suppose that $\bar\gamma_0$ is one-homogeneous and convex and there exists a sequence $\bar\gamma_\epsilon$ of one-homogeneous, convex, and strictly stable functions such that $\gamma_\epsilon \to \gamma_0$ in $W_p^1(S^1)$. Then any curve Γ whose angle parameterization is a solution to* (3.13) *is a closed curve, and moreover it is the Wulff shape of $\bar\gamma_0$, up to translation.*

Proof. We have to prove that the curve Γ is closed. For smooth strictly convex γ's the answer is given by Proposition 2.1 in section 2.

Let us consider an arbitrary $\gamma_0(\cdot)$ and a curve Γ_0 given by solution φ to (3.13). Let us assume that Γ_0 is not closed; i.e., there exists $\varphi_0 \in [0, 2\pi)$ such that

$$\mathbf{x}_0(s(\varphi_0)) \neq \mathbf{x}_0(s(\varphi_0 + 2\pi)),$$

in particular

$$|\mathbf{x}_0(s(\varphi_0) - \mathbf{x}_0(s(\varphi_0 + 2\pi))| = T > 0 \tag{4.8}$$

for a number $T > 0$. Since γ_ϵ is a smooth approximation of γ_0 after a proper selection of $\mathbf{v_0}^\epsilon$ in (2.3), we guarantee that

$$||\Gamma_\epsilon - \Gamma||_{C(0,2\pi)} < T/3. \tag{4.9}$$

This is always possible having Theorem 4.1 at hand. But by Proposition 2.1 we know that Γ_ϵ is closed, so by (4.9) we have

(4.10)

$$|\mathbf{x}_0(s(\varphi_0)) - \mathbf{x}_0(s(\varphi_0 + 2\pi))| \leq |\mathbf{x}_0(s(\varphi_0) - \mathbf{x}_\epsilon(s(\varphi_0))| + |\mathbf{x}_\epsilon(s(\varphi_0)) - \mathbf{x}_\epsilon(s(\varphi_0 + 2\pi))|$$
$$+|\mathbf{x}_\epsilon(s(\varphi_0 + 2\pi)) - \mathbf{x}_0(s(\varphi_0 + 2\pi))| < T/3 + 0 + T/3 = 2/3T,$$

which contradicts (4.8). Finally, we know that Γ_ϵ is the Wulff shape of γ_ϵ. By (2.13) and the uniform convergence of γ_ϵ to γ_0, we deduce that Γ_0 defined by (4.1) and (2.3) is the Wulff shape of γ_0. ☐

In order to finish the proofs of Theorems 1.1 and 1.2, it is enough to note that finding a sequence approximating general function γ is elementary. For example, for a proper choice of smooth functions $\rho_\epsilon : S^1 \to \mathbb{R}$ converging to δ_1 on the circle, we may take

$$\gamma_\epsilon(\varphi) = \epsilon + \gamma \star \rho_\epsilon(\varphi). \tag{4.11}$$

Here $\gamma\star\rho_\epsilon$ denotes the convolution on S^1, and as a result γ_ϵ is smooth and 2π-periodic.

Hence Lemma 4.2 implies the first of our main results, stated in Theorem 1.1, i.e., the existence of φ fulfilling the definition of the generalized solution (3.13). Theorem 1.2 follows from Theorem 4.1. ∎

5. Properties of the solutions. Since our case concerns only plane domains, Hessian $K(\mathbf{n}(\cdot))$—see (2.6)—can be zero on an interval, i.e., $\gamma'' + \gamma = 0$ on (a, b). This in turn is equivalent to F_γ containing a line segment between the angles a and b, (see [Gu, Lemma 7D]). The Euclidean curvature of the Frank diagram is zero there. We want to control precisely the behavior of the system at such points, because singularities may appear there.

For general γ, we concentrate our attention only on the support of K, i.e., on the following set:

$$\operatorname{supp} K := \overline{\{\varphi \in S^1 : \mathbf{t}(\varphi) \cdot K(\mathbf{n}(\varphi)) \cdot \mathbf{t}(\varphi) \neq 0\}}. \tag{5.1}$$

It is implicitly understood that the points of nondifferentiability of γ belong to $\operatorname{supp} K$. Roughly speaking, $\operatorname{supp} K$ consists of points of strict stability of γ. Certainly, due to our previous considerations $(\partial I)^{-1}$ is single-valued and continuous on this set. In order to understand the whole structure of solutions we need to control the behavior of intervals $(a, b) \subset S^1$ such that

$$\mathbf{t}(\varphi) \cdot K(\mathbf{n}(\varphi)) \cdot \mathbf{t}(\varphi) \equiv 0 \qquad \text{for } \varphi \in (a, b). \tag{5.2}$$

In order to make our considerations precise we introduce the following definition. We call an interval (a, b) on which (5.2) holds *maximal* iff for any sufficiently small $\epsilon > 0$ there exist a_* and b^* such that $0 < a - a_* < \epsilon$, $0 < b^* - b < \epsilon$ and for all $w_a \in \partial I(a_*)$ and $w_b \in \partial I(b^*)$ the following relations are satisfied:

$$w_a < \partial I(a + \epsilon) \quad \text{and} \quad \partial I(b - \epsilon) < w_b. \tag{5.3}$$

Obviously, all points of interval (a, b) are continuity points of ∂I. However, we have no control over the behavior of the ends of this interval, which explains the form of (5.3). The above remarks imply that, to analyze the behavior on the interval (a, b), it is enough to consider sets $(a - \epsilon, b + \epsilon)$.

The control over these singularities is guaranteed by the following lemma.

LEMMA 5.1. *Let us suppose that γ is not strictly stable. If* (5.2) *holds, then for any solution $\varphi(\cdot)$ of system* (2.5) *with the range in* $\operatorname{supp} K$ *and for each interval* (a, b) *satisfying* (5.2) *there exists a parameter s_0 such that*

$$\lim_{s\to s_0{}^-} \varphi(s) \le a \quad \textit{and} \quad \lim_{s\to s_0^+} \varphi(s) \ge b. \tag{5.4}$$

Moreover, if (a, b) is maximal (in the meaning of (5.3)*), then*

$$\lim_{s\to s_0{}^-} \varphi(s) = a \quad \textit{and} \quad \lim_{s\to s_0^+} = b. \tag{5.5}$$

Proof. By considerations in the last section we can introduce a strictly stable approximation of the studied γ. We put:

$$\mathbf{t}(\varphi_\epsilon)\cdot K_\epsilon(\varphi_\epsilon)\cdot\mathbf{t}(\varphi_\epsilon)=\epsilon \qquad \text{for} \quad \varphi_\epsilon\in(a,b) \tag{5.6}$$

for $\epsilon > 0$ and

$$\mathbf{t}(\varphi_\epsilon)\cdot K_\epsilon(\varphi_\epsilon)\cdot\mathbf{t}(\varphi_\epsilon)=\mathbf{t}(\varphi_\epsilon)\cdot K(\varphi_\epsilon)\cdot\mathbf{t}(\varphi_\epsilon)+\epsilon \qquad \text{for} \quad S^1\setminus(a,b). \tag{5.7}$$

By (2.6) we can put $\gamma_\epsilon=\gamma+\epsilon$, getting (5.6) and (5.7). The above definition guarantees that $\gamma_\epsilon\to\gamma$ in W^1_p at least on a neighborhood of interval (a,b), and hence the application of Theorem 1.2 is possible.

Let us suppose that s_a is defined by the following relation $\varphi_\epsilon(s_a)=a$ if (a,b) is maximal. Otherwise, we require $\varphi_\epsilon(s_a)\le a$ in general. Then by (2.10) we see that for the approximation satisfying (5.6) we have

$$\frac{d}{ds}\varphi_\epsilon=\frac{1}{\epsilon} \quad \text{as} \quad \varphi\in(a,b), \tag{5.8}$$

and hence, we are able to find s_b such that $\varphi_\epsilon(s_b)=b$ in the “maximal” case or $\varphi_\epsilon(s_b)\ge b$ in general. Formula (5.8) yields $\varphi_\epsilon(s)=\frac{1}{\epsilon}(s-s_a)+a$, so we conclude that $s_b-s_a=\epsilon(b-a)$. Passing with ϵ to 0 we get at the limit $s_a=s_b=:s_0$. Remembering that φ_ϵ and φ are monotone we easily deduce (5.4). From the definition (5.3) we obtain (5.5). We can prove the same result by applying the approximation by regular γ_ϵ's defined by (4.11). However, this approach yields additional technical difficulties. Thus, one can conclude that any regular approximation of this case leads to (5.4) or (5.5), respectively. ∎

Remark. We are tempted to write formally at point s_0, where function φ has a jump, that

$$\frac{d}{ds}\varphi\Big|_{s=s_0}=(b-a)\delta(s-s_0).$$

From the geometrical point of view the sought curve Γ will have an interior angle of measure $(b-a)$ at point s_0.

Examples. Here we present examples of solutions which are constructed by our method. We present only the extreme cases: the isotropic and a crystalline $\bar\gamma$.

(1) The isotropic energy density is given by $\bar\gamma(x)=|x|$, i.e., $\gamma(\varphi)=1$. Let us take $\vartheta=0$, $s_0=0$; hence, $I(\varphi)=\bar\gamma(\mathbf{n}(\varphi))+\int_0^\varphi d\psi\int_0^\psi\bar\gamma(\mathbf{n}(t))dt=1+\frac{1}{2}\varphi^2$ and thus $\partial_\varphi I(\varphi)=\{\frac{d}{d\varphi}I(\varphi)\}=\{\varphi\}$ and $L(\gamma)=2\pi$. Equation (3.13) takes the form $\varphi(s)=s$. Then we conclude that the solution is a circle, whose length is given by $L(\gamma)=2\pi$.

(2) This case is sometimes called *crystalline*, because the Wulff shape is a polygon. We may consider the following anisotropy: $\bar\gamma(x)=\max\{|x_1|,|x_2|\}$, and hence

$$\gamma(\varphi)=\max\{|\cos\varphi|,|\sin\varphi|\}.$$

Elementary calculations, which are left to the interested reader (we may take $s_0=0,\vartheta=0$ in I), and Lemma 5.1 lead us to a conclusion that $d(\Gamma)=4\sqrt{2}$ and Γ is a square (up to translation) with vertexes at $(\pm1,0),(0,\pm1)$.

Acknowledgment. The second author thanks professor Robert Gulliver for pointing to his attention reference [Bl] and an enlightening discussion on the topic of Lemma 2.1.

REFERENCES

[AW] F. Almgren and L. Wang, *Mathematical existence of crystal growth with Gibbs–Thompson curvature effects*, J. Geom. Anal. 10 (2000), pp. 1–100.

[A] R. Almgren, *Crystalline Saffman-Taylor fingers*, SIAM J. Appl. Math., 55 (1995), pp. 1511–1535.

[Ba] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, 1976.

[BNP] G. Bellettini, M. Novaga, and M. Paolini, *On a Crystalline Variational Problem, Part* I: *First Variation and Global Regularity*, Arch. Ration. Mech. Anal., 157 (2001), pp. 165–191.

[Bl] W. Blaschke, *Differential Geometrie*, Dover, New York, 1945.

[Br] H. Brezis, *Operateurs Maximaux Monotones et Semi-group de Contraction dans les Espace de Hilbert*, North-Holland, Amsterdam, 1973.

[CR] X. Chen and F. Reitich, *Local existence and uniqueness of solutions to the Stefan problem with surface tension and kinetic undercooling*, J. Math. Anal. Appl., 164 (1992), pp. 350–362.

[DE] K. Deckelnick and Ch. M. Elliott, *Local and global existence results for anisotropic Hele-Shaw flows*, Proc. Roy. Soc. Edinburgh Sect. A, 129 (1999), pp. 265–294.

[FM] I. Fonseca and S. Müller, *A uniqueness proof for the Wulff theorem*, Proc. Roy. Soc. Edinburgh Sect. A, 119 (1991), pp. 125–136.

[FR] A. Friedman and F. Reitich, *On the existence of spatially patterned dormant malignancies in a model for the growth of non-necrotic vascular tumors*, Math. Models Methods Appl. Sci., 11 (2001), pp. 601–625.

[GR1] Y. Giga and P. Rybka, *Quasi-static evolution of* 3*-D crystals grown from supersaturated vapor*, Differential Integral Equations, 15 (2002), pp. 1–15.

[GR2] Y. Giga and P. Rybka, *Stability of facets of self-similar motion of a crystal*, Adv. Differential Equations, 10 (2005), pp. 601–634.

[Gu] M. Gurtin, *Thermomechanics of Evolving Phase Boundaries in the Plane*, Clarendon Press, Oxford, 1993.

[Lu] S. Luckhaus, *Solutions for the two-phase Stefan problem with the Gibbs–Thomson law for the melting temperature*, European J. Appl. Math., 1 (1990), pp. 101–111.

[Mo] F. Morgan, *Planar Wulff shape is unique equilibrium*, Proc. Amer. Math. Soc., 133 (2005), pp. 809–813.

[Pa] B. Palmer, *Stability of the Wulff shape*, Proc. Amer. Math. Soc., 126 (1998), pp. 3661–3667.

[Ra] E. V. Radkevich, *The Gibbs–Thomson correction and condition for the existence of classical solution of the modified Stefan problem*, Soviet Math. Dokl., 43 (1991), pp. 274–278.

[T] J. E. Taylor, *Crystalline variational problem*, Bull. Amer. Math. Soc., 84 (1978), pp. 568–588.

SIAM J. MATH. ANAL.
Vol. 39, No. 4, pp. 1135–1149

WEAKLY COUPLED SCHRÖDINGER OPERATORS ON REGULAR METRIC TREES*

HYNEK KOVAŘÍK†

Abstract. Spectral properties of the Schrödinger operator $A_\lambda = -\Delta + \lambda V$ on regular metric trees are studied. It is shown that as λ goes to zero the asymptotical behavior of the negative eigenvalues of A_λ depends on the global structure of the tree.

Key words. Schrödinger operator, Sturm–Liouville problems, metric trees

AMS subject classifications. 34L40, 34B24, 34B45

DOI. 10.1137/070682733

1. Introduction.

The spectrum of a Schrödinger operator

$$-\Delta + \lambda V, \quad \lambda > 0, \quad \text{in} \quad L^2(\mathbb{R}^n)$$

is given by the disjoint union of the essential spectrum $\sigma_{ess}(-\Delta+\lambda V)$ and the discrete spectrum $\sigma_d(-\Delta+\lambda V)$. Under certain decay conditions on V at infinity the essential spectrum covers the half-line $[0,\infty)$ so that the discrete spectrum consists of negative eigenvalues of finite multiplicity. It is a well-known fact that the behavior of these eigenvalues for small values of λ depends on the spacial dimension n [11]. Namely, for $n < 3$ the negative eigenvalues of $-\Delta+\lambda V$ appear for any $\lambda > 0$, provided $\int_{\mathbb{R}^n} V < 0$, while for $n \geq 3$ the negative spectrum of $-\Delta+\lambda V$ remains empty for λ small enough. Moreover for the lowest eigenvalue $-\varepsilon(\lambda)$ of $-\Delta+\lambda V$ the following asymptotics hold true (see [11]):

$$n = 1 : \varepsilon(\lambda) \sim \lambda^2, \qquad \lambda \to 0, \tag{1}$$

$$n = 2 : \varepsilon(\lambda) \sim e^{-\lambda^{-1}}, \quad \lambda \to 0. \tag{2}$$

In this paper we want to find the asymptotic behavior of $\varepsilon(\lambda)$ for a Schrödinger operator

$$A_\lambda = -\Delta + \lambda V, \quad \lambda > 0, \quad \text{in} \quad L^2(\Gamma)$$

defined on a regular metric tree Γ. Such a metric tree consists of the set of vertices and the set of edges (branches), i.e., one-dimensional intervals connecting the vertices; see section 2 for details. Metric trees form a special subclass of so-called quantum graphs. The latter serve as mathematical models for nanotechnological devices that consist of connected very thin strips. The motion of an electron in such a "web" is then governed by the Schrödinger equation. Therefore the study of Laplace and Schrödinger operators on these structures has recently attracted considerable attention; see, e.g., [4, 5, 6, 7, 8, 9, 12, 13], and references therein.

*Received by the editors February 15, 2007; accepted for publication (in revised form) July 30, 2007; published electronically November 7, 2007. The work has been partially supported by the Czech Academy of Sciences and by DAAD within the project D-CZ 5/05-06.
http://www.siam.org/journals/sima/39-4/68273.html

†Institute of Analysis, Dynamics and Modeling, Universität Stuttgart, PF 80 11 40, D-70569 Stuttgart, Germany (kovarik@mathematik.uni-stuttgart.de).

We are interested in the spectral behavior of A_λ when $\lambda \to 0$. The intuitive expectation is that this should depend on the rate of the growth, or branching, of the tree Γ. In order to quantify this branching, we assign to Γ a so-called *global dimension* d; see Definition 2 below. Roughly speaking it tells us how fast the number of branches of Γ increases as a function of the distance from its root. If the latter grows with the power $d-1$ at infinity, then we say, in analogy with the Euclidean spaces, that d is the global dimension of Γ. We use the notation global in order to distinguish d from the local dimension, which is of course one. Since d can be in general any real number larger or equal to one, it is natural to ask how the weak coupling behavior looks for noninteger values of d and what the condition on V is under which the eigenvalues appear.

Our main result says (see Theorem 3 and section 5.2), that if $d \in [1,2]$ and $\int_\Gamma V < 0$, then A_λ possesses at least one negative eigenvalue for any $\lambda > 0$, and for λ small enough this eigenvalue is unique and satisfies

$$\begin{aligned} &\varepsilon(\lambda) \sim \lambda^{\frac{2}{2-d}}, \qquad 1 \le d < 2\,, \\ &\varepsilon(\lambda) \sim e^{-\lambda^{-1}}, \qquad d = 2. \end{aligned} \tag{3}$$

As expected, the faster the branching of Γ, i.e., larger d, the faster the eigenvalue tends to zero. The borderline is reached at $d = 2$, in which case $\varepsilon(\lambda)$ converges to zero faster than any power of λ similarly as in (2). Finally, if the tree grows too fast, i.e., $d > 2$, then the discrete spectrum of A_λ generically remains empty for λ small enough; see section 5.3. Note also that the condition $\int_\Gamma V < 0$ is almost optimal in the sense that if $\int_\Gamma V > 0$, then A_λ has no negative eigenvalues for small λ as shown in Theorem 3b.

To study the operator A_λ we make use of the decomposition (5), Theorem 1, which was proved in [8, 9]; see also [4]. In section 3.1 we introduce certain auxiliary operators, whose eigenvalues will give us the estimate on $\varepsilon(\lambda)$ from above and from below. In order to establish (3) we find the asymptotics of the lowest eigenvalues of the auxiliary operators, which are of the same order. This is done in section 5.1. In addition, in section 4 we give some estimates on the number of eigenvalues of the individual operators in the decomposition (5), which might be of an independent interest as well.

Throughout the text we will employ the notation $\alpha := d-1$ and $\nu := \frac{2-d}{2}$. For a real-valued function f and a real noninteger number μ, we will use the shorthand

$$f^\mu := \operatorname{sign} f\, |f|^\mu = \frac{f\,|f|^\mu}{|f|}.$$

Finally, given a self-adjoint operator T on a Hilbert space $\mathcal{H}$ we denote by $N_-(T;s)$ the number of eigenvalues, taking into account their multiplicities, of T on the left of the point s. For $s = 0$ we will write $N_-(T)$ instead of $N_-(T;0)$.

2. Preliminaries. We define a metric tree Γ with the root o following the construction given in [8]. Let $\mathcal{V}(\Gamma)$ be the set of vertices and $\mathcal{E}(\Gamma)$ be the set of edges of Γ. The distance $\rho(y,z)$ between any two points $y, z \in \Gamma$ is defined in a natural way as the length of the unique path connecting y and z. Consequently, $|y|$ is equal to $\rho(y,o)$. We write $y \preceq z$ if y lies on the unique simple path connecting o with z. For $y \preceq z$ we define

$$\langle y, z\rangle := \{x \in \Gamma \,:\, y \preceq x \preceq z\}.$$

If $e = \langle y, z\rangle$ is an edge, then y and z are its end points. For any vertex z its generation $Gen(z)$ is defined by

$$Gen(z) = \#\{x \in \mathcal{V} : o \prec x \preceq z\}.$$

For an edge $e \in \mathcal{E}(\Gamma)$ we define its generation as the generation of the vertex, from which e is emanating. The branching number $b(z)$ of the vertex z is equal to the number of edges emanating from z. We assume that $b(z) > 1$ for any $z \neq o$ and $b(o) = 1$.

Definition 1. *A tree* Γ *is called* regular *if all of the vertices of the same generation have equal branching numbers and all of the edges of the same generation have equal length.*

We denote by $t_k > 0$ the distance between the root and the vertices of the kth generation and by $b_k \in \mathbb{N}$ their corresponding branching number. For each $k \in \mathbb{N}$ we define the so-called branching function $g_k : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$g_k(t) := \begin{cases} 0 & \text{if} \quad t < t_k, \\ 1 & \text{if} \quad t_k \le t < t_{k+1}, \\ b_{k+1}b_{k+2}\dots b_n & \text{if} \quad t_n \le t < t_{n+1},\ k < n, \end{cases}$$

and

$$g_0(t) := b_0\, b_1 \dots b_n, \quad t_n \le t < t_{n+1}.$$

It follows directly from the definition that

$$g_0(t) = \#\{x \in \Gamma : |x| = t\}.$$

Obviously $g_0(\cdot)$ is a nondecreasing function, and the rate of growth of g_0 determines the rate of growth of the tree Γ. In particular, if one denotes by $\Gamma(t) := \{x \in \Gamma : |x| \le t\}$ the "ball" of radius t, then g_0 tells us how fast the surface of $\Gamma(t)$ grows with t. This motivates the following.

Definition 2. *If there exist positive constants* a^-, a^+, *and* T_0, *such that for all* $t \ge T_0$ *the inequalities*

$$a^- \le \frac{g_0(t)}{t^{d-1}} \le a^+$$

hold true, then we say that d *is the* global dimension *of the tree* Γ.

We note that in the case of the so-called homogeneous metric trees treated in [12] the function $g_0(t)$ grows faster than any power of t. Formally, this corresponds to $d = \infty$ in the above definition. From now on we will work under the assumption that $d < \infty$.

3. Schrödinger operators on Γ. We will consider potential functions V which satisfy the following assumption.

Assumption A. $V : \mathbb{R}_+ \to \mathbb{R}$ *is measurable and bounded and* $\lim_{t\to\infty} V(t) = 0$.

For a given function V which satisfies Assumption A we define the Schrödinger operator A_λ as the self-adjoint operator in $L^2(\Gamma)$ associated with the closed quadratic form

$$Q_\lambda[u] := \int_\Gamma \left(|u'|^2 + \lambda V(|x|)\, |u|^2\right) dx,$$

with the form domain $D(Q) = H^1(\Gamma)$ consisting of all continuous functions u such that $u \in H^1(e)$ on each edge $e \in \mathcal{E}(\Gamma)$ and

$$\int_\Gamma \left(|u'|^2 + |u|^2\right)\, dx < \infty.$$

The domain of A_λ consists of all continuous functions u such that $u'(o) = 0$, $u \in H^2(e)$ for each $e \in \mathcal{E}(\Gamma)$ and such that at each vertex $z \in \mathcal{V}(\Gamma) \setminus \{o\}$ the matching conditions

$$u_-(z) = u_1(z) = \cdots = u_{b(z)}(z), \quad u_1'(z) + \cdots + u_{b(z)}'(z) = u_-'(z) \tag{4}$$

are satisfied, where u_- denotes the restriction of u on the edge terminating in z and u_j, $j = 1, \ldots, b(z)$, denote, respectively, the restrictions of u on the edges emanating from z; see [8] for details. Notice that A_λ satisfies the Neumann boundary condition at the root o.

The following result by Naimark and Solomyak (see [8, 9]), also established by Carlson in [4], makes it possible to reduce the spectral analysis of A_λ to the analysis of one-dimensional Schrödinger operators in weighted $L^2(\mathbb{R}_+)$ spaces.

THEOREM 1. *Let V be measurable and bounded, and suppose that Γ is regular. Then A_λ is unitarily equivalent to the following orthogonal sum of operators:*

$$A_\lambda \sim A_{\lambda,0} \oplus \sum_{k=1}^{\infty} \oplus\, A_{\lambda,k}^{[b_1 \ldots b_{k-1}(b_k - 1)]}. \tag{5}$$

Here the symbol $A_{\lambda,k}^{[b_1 \ldots b_{k-1}(b_k - 1)]}$ means that the operator $A_{\lambda,k}$ enters the orthogonal sum $[b_1 \ldots b_{k-1}(b_k - 1)]$ times. For each $k \in \mathbb{N}$ the corresponding self-adjoint operator $A_{\lambda,k}$ acts in $L^2((t_k, \infty),\, g_k)$ and is associated with the closed quadratic form

$$Q_k[f] = \int_{t_k}^\infty \left(|f'|^2 + \lambda V(t)\, |f|^2\right)\, g_k(t)\, dt,$$

whose form domain is given by the the weighted Sobolev space $D(Q_k) = H_0^1((t_k, \infty),\, g_k)$, which consists of all functions f such that

$$\int_{t_k}^\infty \left(|f'|^2 + |f|^2\right)\, g_k(t)\, dt < \infty, \quad f(t_k) = 0.$$

The operator $A_{\lambda,0}$ acts in the weighted space $L^2(\mathbb{R}_+, g_0)$ and is associated with the closed form

$$Q_0[f] = \int_0^\infty \left(|f'|^2 + \lambda V(t)\, |f|^2\right)\, g_0(t)\, dt,$$

with the form domain $D(Q_0) = H^1(\mathbb{R}_+,\, g_0)$, which consists of all functions f such that

$$\int_0^\infty \left(|f'|^2 + |f|^2\right)\, g_0(t)\, dt < \infty;$$

see also [13].

3.1. Auxiliary operators. Let d be the global dimension of Γ. Definition 2 implies that there exist positive constants b^- and b^+ such that

$$b^-\,(1+t)^\alpha =: g_k^-(t) \le g_k(t) \le g_k^+(t) := b^+(1+t)^\alpha\,, \quad t\in[t_k,\infty). \tag{6}$$

Now assume that the Rayleigh quotient

$$\frac{\int_{t_k}^\infty \left(|f'|^2+\lambda V(t)\,|f|^2\right)\, g_k(t)\,dt}{\int_{t_k}^\infty |f|^2\, g_k(t)\,dt}$$

of the operator $A_{\lambda,k}$, $k\ge 0$ is negative for some $f\in D(Q_k)$. From (6) follows that

$$\begin{aligned}\frac{\int_{t_k}^\infty \left(|f'|^2+\lambda V_k^-(t)\,|f|^2\right)\,(1+t)^\alpha\,dt}{\int_{t_k}^\infty |f|^2\,(1+t)^\alpha\,dt} &\le \frac{\int_{t_k}^\infty \left(|f'|^2+\lambda V(t)\,|f|^2\right)\, g_k(t)\,dt}{\int_{t_k}^\infty |f|^2\, g_k(t)\,dt}\\ &\le \frac{\int_{t_k}^\infty \left(|f'|^2+\lambda V_k^+(t)\,|f|^2\right)\,(1+t)^\alpha\,dt}{\int_{t_k}^\infty |f|^2\,(1+t)^\alpha\,dt},\end{aligned} \tag{7}$$

where

$$V_k^-(t) := \frac{g_k(t)}{g_k^-(t)}\,V(t), \quad V_k^+(t) := \frac{g_k(t)}{g_k^+(t)}\,V(t).$$

It is thus natural to introduce the auxiliary operators $A^\pm_{\lambda,k}$ acting in the Hilbert space $L^2((t_k,\infty),\,(1+t)^\alpha)$ and associated with the quadratic forms

$$Q_k^\pm[f] = \int_{t_k}^\infty \left(|f'|^2 + \lambda V_k^\pm(t)\,|f|^2\right)\,(1+t)^\alpha\,dt, \quad f\in D(Q_k), k\in\mathbb{N}_0. \tag{8}$$

The variational principle (see, e.g., [3]) and (7) thus imply that

$$N_-(A^+_{\lambda,k};s) \le N_-(A_{\lambda,k};s) \le N_-(A^-_{\lambda,k};s), \quad s\le 0,\ k\in\mathbb{N}_0. \tag{9}$$

Let $E_{n,k}(\lambda)$ be the nondecreasing sequence of negative eigenvalues of the operators $A_{\lambda,k}$, and let $E^\pm_{n,k}(\lambda)$ be the analogous sequences corresponding to the operators $A^\pm_{\lambda,k}$, respectively. In all of these sequences each eigenvalue occurs according to its multiplicity. Relation (9) and the variational principle then yield

$$E^-_{n,k}(\lambda) \le E_{n,k}(\lambda) \le E^+_{n,k}(\lambda), \quad k\in\mathbb{N}_0, n\in\mathbb{N}, \tag{10}$$

and

$$\inf\sigma_{ess}(A^-_{\lambda,k}) \le \inf\sigma_{ess}(A_{\lambda,k}) \le \inf\sigma_{ess}(A^+_{\lambda,k})\,, \quad k\in\mathbb{N}_0. \tag{11}$$

Next we introduce the transformation U by

$$(U\,f)(t) = (1+t)^{\alpha/2}\,f(t) =: \varphi(t),$$

which maps $L^2((t_k,\infty),(1+t)^\alpha)$ unitarily onto $L^2((t_k,\infty))$. We thus get the following.

Lemma 1. *Let V satisfy the assumptions of Theorem 1. Then*

(i) *for each $k \in \mathbb{N}$ the operators $A^{\pm}_{\lambda,k}$ are unitarily equivalent to the self-adjoint operators $B^{\pm}_{\lambda,k}$ in $L^2((t_k,\infty))$, which act as*

$$\left(B^{\pm}_{\lambda,k}\,\varphi\right)(t) = -\varphi''(t) + \frac{(d-1)(d-3)}{4(1+t)^2}\,\varphi(t) + \lambda V^{\pm}_k(t)\,\varphi(t), \tag{12}$$

and whose domains consist of all functions $\varphi \in H^2((t_k,\infty))$ such that

$$\varphi(t_k) = 0.$$

(ii) *$A^{\pm}_{\lambda,0}$ are unitarily equivalent to the self-adjoint operators $B^{\pm}_{\lambda,0}$ in $L^2(\mathbb{R}_+)$, acting as*

$$\left(B^{\pm}_{\lambda,0}\,\varphi\right)(t) = -\varphi''(t) + \frac{(d-1)(d-3)}{4(1+t)^2}\,\varphi(t) + \lambda\, V^{\pm}_0(t)\,\varphi(t), \tag{13}$$

with the domain that consists of all $\varphi \in H^2(\mathbb{R}_+)$ such that

$$\varphi'(0) = \frac{d-1}{2}\,\varphi(0). \tag{14}$$

Proof. For each $k \in \mathbb{N}_0$ we have

$$B^{\pm}_{\lambda,k} = U\,A^{\pm}_{\lambda,k}\,U^{-1}, \quad \|f\|_{L^2((t_k,\infty),\,(1+t)^\alpha)} = \|Uf\|_{L^2((t_k,\infty))}.$$

The statement of the lemma then follows by a direct calculation keeping in mind that the functions f from the domain of the operators $A^{\pm}_{\lambda,0}$ satisfy $f'(0) = 0$. ∎

Remark 1. If V satisfies Assumption A, then the inequalities (11) and standard arguments from the spectral theory of Schrödinger operators (see, e.g., [10, Chap. 13.4]) imply that

$$\inf\,\sigma_{ess}(A^-_{\lambda,k}) = \inf\,\sigma_{ess}(A_{\lambda,k}) = \inf\,\sigma_{ess}(A^+_{\lambda,k}) = 0 \quad \forall\, k \in \mathbb{N}_0.$$

Moreover, constructing suitable Weyl sequences for the operators $A_{\lambda,k}$ in the similar way as it was done in [13] for the Laplace operator, one can easily show that

$$\sigma_{ess}(A_{\lambda,k}) = [0,\infty) \quad \forall\, k \in \mathbb{N}_0. \tag{15}$$

4. Number of bound states. From Theorem 1 and (15) we can see that if V satisfies Assumption A, then

$$\sigma_{ess}\,(A_\lambda) = [0,\infty). \tag{16}$$

In order to analyze the discrete spectrum of A_λ we first study the number of bound states of the individual operators in the decomposition (5).

We start by proving an auxiliary proposition. Given a real-valued measurable bounded function $\tilde V$ we consider the self-adjoint operator $\tilde B_\lambda$ acting in $L^2(\mathbb{R}_+)$ as

$$\left(\tilde B_\lambda\,\varphi\right)(t) = -\varphi''(t) + \frac{(d-1)(d-3)}{4\,t^2}\,\varphi(t) + \lambda \tilde V(t)\,\varphi(t) \tag{17}$$

with the Dirichlet boundary condition at zero. This operator is associated with the closure of the quadratic form

$$\int_{\mathbb{R}_+} \left(|\varphi'(t)|^2 + \frac{(d-1)(d-3)}{4\,t^2}\;|\varphi(t)|^2 + \lambda \tilde V(t)\,|\varphi(t)|^2\right) dt$$

defined on $C_0^\infty(\mathbb{R}_+)$.

PROPOSITION 1. *Let $d \in [1,2)$. Assume that $\tilde V$ satisfies Assumption A and that $\int_0^\infty t\,|\tilde V(t)|\,dt < \infty$. Then*

$$N_-(\tilde B_\lambda) \le \lambda\,\tilde K(d) \int_0^\infty t\,|\tilde V(t)|\,dt, \tag{18}$$

where

$$\tilde K(d) = \frac{\pi}{2\sin(\nu\pi)\Gamma(1-\nu)\Gamma(1+\nu)}.$$

Proof. We write

$$\tilde B_\lambda = \tilde B_0 + \lambda\,\tilde V, \quad \tilde B_0 := -\frac{d^2}{d\,t^2} + \frac{(d-1)(d-3)}{4t^2}.$$

Moreover, without loss of generality we may assume that $\tilde V$ is negative and continuous. The statement for the general case then follows by a standard approximation argument.

By the Birman–Schwinger principle (see, e.g., [3]), the number of eigenvalues of $\tilde B_\lambda$ to the left of the point $-\kappa^2$ then does not exceed the trace of the operator

$$\lambda|\tilde V|^{1/2}\,(\tilde B_0+\kappa^2)^{-1}\,|\tilde V|^{1/2}.$$

The integral kernel $\tilde G(t,t',\kappa)$ of the operator $(\tilde B_0+\kappa^2)^{-1}$ can be calculated by using the Sturm–Liouville theory. We get

$$\tilde G(t,t',\kappa) = \begin{cases} \frac{\pi\,i}{4}\,v_1(t,\kappa)\,v_2(t',\kappa), & t \ge t', \\ \frac{\pi\,i}{4}\,v_1(t',\kappa)\,v_2(t,\kappa), & t < t', \end{cases} \tag{19}$$

with

$$\begin{aligned} v_1(t,\kappa) &= \sqrt{t}\;H_\nu^{(1)}(i\kappa t), \\ v_2(t,\kappa) &= \sqrt{t}\,H_\nu^{(1)}(i\kappa t) + \sqrt{t}\,H_\nu^{(2)}(i\kappa t), \end{aligned}$$

where $H_\nu^{(1)}$ (respectively, $H_\nu^{(2)}$) denote Hankel's functions of the first (respectively, second) kind; see, e.g., [14]. Since $\int_0^\infty t\,|\tilde V(t)|\,dt < \infty$, we can pass to the limit $\kappa \to 0$ in the corresponding integral, using the Lebesgue dominated convergence theorem, and calculate the trace to get

$$N_-(\tilde B_\lambda) \le \lambda \int_0^\infty |\tilde V(t)|\,|\tilde G(t,t,0)|\,dt = \lambda\,\tilde K(d) \int_0^\infty t\,|\tilde V(t)|\,dt. \tag{20}$$

Here we have used the fact that $\tilde G(t,t,\kappa) \to t\,\tilde K(d)$ pointwise as $\kappa \to 0$, which follows from the asymptotic behavior of the Hankel functions at zero; see, e.g., [1]. ∎

Remark 2. For $d = 1$ we have $\tilde K(1) = 1$, and (18) gives the well-known Bargmann inequality [2]. On the other hand, $\tilde K(d)$ diverges as $d \to 2$. This is expected because the operator $-\frac{d^2}{d\,t^2} - \frac{1}{4t^2} + \lambda V$ with Dirichlet boundary conditions. at zero does have at least one negative eigenvalue for any $\lambda > 0$ if the integral of V is negative.

Armed with Proposition 1 we can prove the following.

COROLLARY 1. *Let $1 \le d < 2$. Assume that V satisfies Assumption A and that $\int_0^\infty t\,|V(t)|\,dt < \infty$. Then*

$$N_-(A_{\lambda,0}) \le 1 + \lambda\,K(d) \int_0^\infty |V(t)|\,g_0(t)\,t^{2-d}dt. \tag{21}$$

Proof. We introduce the operator $A^D_{\lambda,0}$, which is associated with the quadratic form

$$Q^D_0[f] := \int_0^\infty \left(|f'|^2 + \lambda V(t)\,|f|^2\right)\, g_0(t)\, dt, \quad D(Q^D_0) = H^1_0(\mathbb{R}_+,\, g_0),$$

where $H^1_0(\mathbb{R}_+,\, g_0) := \{f \in H^1(\mathbb{R}_+,\, g_0),\ f(0) = 0\}$. First we observe that

$$a\, t^{d-1} \leq g_0(t), \quad t \in \mathbb{R}_+,$$

for a suitable $a > 0$. We can thus mimic the analysis of section 3.1 and define the operator $\tilde{A}_\lambda$ acting in $L^2(\mathbb{R}_+,\, t^{d-1})$ associated with the quadratic form

$$\tilde{Q}[f] = \int_0^\infty \left(|f'|^2 + \lambda \tilde{V}(t)\,|f|^2\right) t^{d-1}\, dt, \quad f \in D(Q), \tag{22}$$

where $D(Q) = H^1_0((\mathbb{R}_+),\, t^{d-1})$ and $\tilde{V}(t) := \frac{g_0(t)}{a\, t^{d-1}}\, V(t)$. Repeating the arguments of section 3.1 we claim that

$$N_-(A^D_{\lambda,0}) \leq N_-(\tilde{A}_\lambda)$$

and that $\tilde{A}_\lambda$ is unitarily equivalent to $\tilde{B}_\lambda$ by means of the transformation $\tilde{U}\, f(t) = t^{(d-1)/2}\, f(t)$, which maps $L^2(\mathbb{R}_+,\, t^{d-1})$ unitarily onto $L^2(\mathbb{R}_+)$. Since the codimension of $H^1_0(\mathbb{R}_+,\, g_0)$ in $H^1(\mathbb{R}_+,\, g_0)$ is equal to one, the variational principle gives

$$N_-(A_{\lambda,0}) \leq 1 + N_-(A^D_{\lambda,0}) \leq 1 + N_-(\tilde{A}_\lambda) = 1 + N_-(\tilde{B}_\lambda).$$

Application of Proposition 1 with $\tilde{K}(d) = a\, K(d)$ concludes the proof. □

Corollary 2. *Let $1 \leq d < 2$. Let V satisfy Assumption A, and assume that $\int_0^\infty t\,|V(t)|\, dt < \infty$. Then there exists $\lambda_c > 0$, so that for $\lambda \in [0, \lambda_c]$ the discrete spectra of the operators $A_{\lambda,k}$, $k \geq 1$, are empty. In particular we have*

$$\sigma_d\ (A_\lambda) = \sigma_d\ (A_{\lambda,0})\,, \quad 0 \leq \lambda \leq \lambda_c, \tag{23}$$

where the multiplicities of the eigenvalues are taken into account.

Proof. In view of Lemma 1 it suffices to show that the discrete spectra of the operators $B^-_{\lambda,k}$ for $k \geq 1$ are empty, provided λ is small enough. Since $(d-1)(d-3) \leq 0$, the following inequality holds true in the sense of quadratic forms:

$$B^-_{\lambda,k} \geq \mathcal{B}_{\lambda,k} := -\frac{d^2}{dt^2} + \frac{(d-1)(d-3)}{4(t-t_k)^2} + \lambda V^-_k(t), \tag{24}$$

where $\mathcal{B}_{\lambda,k}$ acts in $L^2((t_k, \infty))$ with Dirichlet boundary conditions at t_k. A simple translation $s = t - t_k$ then shows that $\mathcal{B}_{\lambda,k}$ is unitarily equivalent to the operator

$$-\frac{d^2}{ds^2} + \frac{(d-1)(d-3)}{4s^2} + \lambda V^-_k(s + t_k) \quad \text{in} \quad L^2(\mathbb{R}_+),$$

which is defined in the similar way as the operator $\tilde{B}_\lambda$ in (17). Since $\int_0^\infty s\,|V^-_k(s + t_k)|\, ds$ is uniformly bounded with respect to k, it follows from Proposition 1 that for λ small enough we have $N_-(\mathcal{B}_{\lambda,k}) = 0$ for all $k \geq 1$. In view of (24) this concludes the proof. □

5. Weak coupling.

5.1. The case $1 \le d < 2$. In this section we will show that if $d \in [1,2)$ and V is attractive in a certain sense, then the operator A_λ possesses at least one negative eigenvalue for any $\lambda > 0$. Since for small values of λ the discrete spectra of A_λ and $A_{\lambda,0}$ coincide (see Corollary 2), we will focus on the operator $A_{\lambda,0}$ only. More exactly, in view of (10), we will study the operators $B^\pm_{\lambda,0}$. Clearly we have

$$B^\pm_{\lambda,0} = B_0 + \lambda\, V_0^\pm, \quad B_0 := -\frac{d^2}{d\,t^2} + \frac{(d-1)(d-3)}{4(1+t)^2},$$

with the boundary condition $v'(0) = \frac{d-1}{2}\, v(0)$. Note that, by Lemma 1, the operator B_0 is nonnegative. We shall first calculate the Green function of B_0 at a point $-\kappa^2$, $\kappa > 0$, using the Sturm–Liouville theory again. In the same manner as in the previous section we obtain

$$G(t,t',\kappa) := \begin{cases} \frac{\pi}{4i\beta(\kappa)}\, v_1(t,\kappa)\, v_2(t',\kappa), & t \ge t', \\ \frac{\pi}{4i\beta(\kappa)}\, v_1(t',\kappa)\, v_2(t,\kappa), & t < t', \end{cases} \tag{25}$$

where

$$\begin{aligned} v_1(t,\kappa) &= \sqrt{1+t}\ H^{(1)}_\nu(i\kappa(1+t)), \\ v_2(t,\kappa) &= \sqrt{1+t}\ \Big(H^{(1)}_\nu(i\kappa(1+t)) - \beta(\kappa)\, H^{(2)}_\nu(i\kappa(1+t))\Big), \\ \beta(\kappa) &= \frac{H^{(1)}_{\nu-1}(i\kappa)}{H^{(2)}_{\nu-1}(i\kappa)}. \end{aligned}$$

Consider a function W which satisfies Assumption A. According to the Birman–Schwinger principle the operator $B_0 + \lambda W$ has an eigenvalue $-\kappa^2$ if and only if the operator

$$K(\kappa) := |W|^{1/2}\, (B_0 + \kappa^2)^{-1}\, W^{1/2}$$

has eigenvalue $-\lambda^{-1}$. The integral kernel of $K(\kappa)$ is equal to

$$K(t,t',\kappa) = |W(t)|^{1/2}\, G(t,t',\kappa)\, (W(t'))^{1/2}.$$

We will use the decomposition

$$K(t,t',\kappa) = L(t,t',\kappa) + M(t,t',\kappa),$$

with

$$L(t,t',\kappa) := \frac{\pi\, 2^{2\nu-1}\, \kappa^{-2\nu}}{(\Gamma(1-\nu))^2\, \sin(\nu\pi)}\ |W(t)|^{1/2}\, [(1+t)(1+t')]^{-\nu+\frac{1}{2}}\, W(t')^{1/2},$$

and denote by $L(\kappa)$ and $M(\kappa)$ the integral operators with the kernels $L(t,t',\kappa)$ and $M(t,t',\kappa)$ respectively. Furthermore, we denote by $M(0)$ the integral operator with the kernel

$$M(t,t',0) := C_M(\nu)\, \big(|W(t)| W(t')\, (1+t)(1+t')\big)^{\frac{1}{2}}\, \left(\frac{1+t}{1+t'}\right)^{\nu\, \mathrm{sign}\,(t-t')},$$

where

$$C_M(\nu) := -\frac{\pi}{2\sin(\nu\pi)\Gamma(1-\nu)\Gamma(1+\nu)}.$$

Lemma 2 in the appendix says that $M(\kappa)$ converges in the Hilbert–Schmidt norm to the operator $M(0)$ as $\kappa \to 0$, provided W decays fast enough at infinity. This allows us to prove the following.

THEOREM 2. *Let W satisfy Assumption A and let $\int_0^\infty (1+t)^{3-d}\,|W(t)|\,dt < \infty$, where $1 \le d \le 2$. Then the following statements hold true:*

(a) *If*

$$\int_0^\infty W(t)\,(1+t)^{d-1}\,dt < 0,$$

then the operator $B_0 + \lambda W$ has at least one negative eigenvalue for all $\lambda > 0$. For λ small enough this eigenvalue, denoted by $E(\lambda)$, is unique and satisfies

$$(E(\lambda))^{\frac{2-d}{2}} = C(\nu)\left(\lambda \int_0^\infty W(t)\,(1+t)^{d-1}\,dt + \mathcal{O}(\lambda^2)\right), \tag{26}$$

where

$$C(\nu) = \frac{\pi\, 2^{2\nu-1}}{(\Gamma(1-\nu))^2\,\sin(\nu\pi)}.$$

(b) *If*

$$\int_0^\infty W(t)\,(1+t)^{d-1}\,dt > 0,$$

then the operator $B_0 + \lambda W$ has no negative eigenvalues for λ positive and small enough.

Proof. Part (a). The operator $B_0 + \lambda W$ has eigenvalue $E = -\kappa^2$ if and only if the operator

$$\lambda K(\kappa) = \lambda M(\kappa) + \lambda L(\kappa)$$

has an eigenvalue -1 for certain $\kappa(\lambda)$. On the other hand, Lemma 1 and (9) imply that

$$N_-\left(B_0 + \lambda\,\frac{g_0}{g_0^+}\,V\right) \le N_-(A_{\lambda,0}).$$

The uniqueness of E, and so of $\kappa(\lambda)$, for λ small enough thus follows from (21) by taking $V = \frac{g_0^+}{g_0}\,W$. Next we note that by Lemma 2 for λ small we have $\lambda\,\|M(\kappa)\| < 1$ and

$$(I + \lambda K(\kappa))^{-1} = \left[I + \lambda(I + \lambda M(\kappa))^{-1}\,L(\kappa)\right]^{-1}\,(I + \lambda\,M(\kappa))^{-1}.$$

Hence $\lambda K(\kappa)$ has an eigenvalue -1 if and only if $\lambda(I + \lambda M(\kappa))^{-1}\,L(\kappa)$ has an eigenvalue -1. Since $\lambda(I + \lambda M(\kappa))^{-1}\,L(\kappa)$ is of rank one we get the equation for $\kappa(\lambda)$ in the form

$$\operatorname{tr}\left(\lambda(I + \lambda M(\kappa(\lambda))^{-1}\,L(\kappa(\lambda)))\right) = -1. \tag{27}$$

Using the decomposition

$$(I+\lambda M(\kappa))^{-1} = I - \lambda M(0) - \lambda(M(\kappa)-M(0)) + \lambda^2 M^2(\kappa)(I+\lambda M(\kappa))^{-1},$$

we obtain

$$\begin{aligned}
&\operatorname{tr}\left(\lambda(I+\lambda M(\kappa))^{-1}\, L(\kappa)\right)\\
&= \lambda\, C(\nu)\kappa^{-2\nu}\, \left(|W(t)|^{1/2}\,(1+t)^{-\nu+\frac12}, (I+\lambda M(\kappa))^{-1} W(t)^{1/2}\,(1+t)^{-\nu+\frac12}\right)\\
&= C(\nu)\,\kappa^{-2\nu}\left(\lambda\int_0^\infty W(t)\,(1+t)^{d-1}\,dt + \mathcal{O}(\lambda^2)\right).
\end{aligned}$$

It thus follows from (27) that

$$E^\nu(\lambda) = -\kappa^{2\nu}(\lambda) = C(\nu)\left(\lambda\int_0^\infty W(t)\,(1+t)^{d-1}\,dt + \mathcal{O}(\lambda^2)\right). \tag{28}$$

To finish the proof of part (a) of the theorem, we mimic the argument used in [11] and notice that if $(\varphi,(B_0+\lambda W)\,\varphi)<0$, then $(\varphi, W\varphi)<0$, since B_0 is nonnegative, and therefore $(\varphi,(B_0+\tilde\lambda W)\,\varphi)<0$ if $\lambda<\tilde\lambda$. So if $B_0+\lambda W$ has a negative eigenvalue for λ small enough, then, by the variational principle, it has at least one negative eigenvalue for all λ positive.

Part (b). From the proof of part (a) it can be easily seen that if

$$\int_0^\infty W(t)\,(1+t)^{d-1}\,dt > 0,$$

then $\operatorname{tr}\left(\lambda(I+\lambda M(\kappa))^{-1}\, L(\kappa)\right)$ is positive for λ small, and therefore $K(\kappa)$ cannot have an eigenvalue -1. □

Remark 3. Note that if

$$W_0 := \int_{\mathbb{R}^2_+} W(t)\,W(t')(1+t)^{1-\nu}(1+t')^{1-\nu}\left(\frac{1+t}{1+t'}\right)^{\nu\,\operatorname{sign}(t-t')} dt\,dt' < 0,$$

then the operator $B_0+\lambda W$ has a negative eigenvalue for λ small, positive or negative, also in the critical case when

$$\int_0^\infty W(t)\,(1+t)^{d-1}\,dt = 0.$$

Moreover, it follows from the proof of Theorem 2 that this eigenvalue then satisfies

$$E^\nu(\lambda) = C(\nu)\left(-\lambda^2\, C_M(\nu)\, W_0 + o\,(\lambda^2)\right), \quad \lambda\to 0. \tag{29}$$

As an immediate consequence of Theorem 2 and inequalities (10) we get the following.

Theorem 3. *Let V satisfy Assumption A, and let $\int_0^\infty (1+t)^{3-d}\,|V(t)|\,dt<\infty$, where $1\le d<2$. Then the following statements hold true:*

(a) *If*

$$\int_0^\infty V(t)\,g_0(t)\,dt = \int_\Gamma V(|x|)\,dx < 0,$$

then the operator A_λ has at least one negative eigenvalue $E_{1,0}(\lambda)$ for all $\lambda > 0$. For λ small enough this eigenvalue is unique and satisfies

$$C_1 \left| \lambda \int_\Gamma V(|x|)\, dx \right|^{\frac{2}{2-d}} \leq |E_{1,0}(\lambda)| \leq C_2 \left| \lambda \int_\Gamma V(|x|)\, dx \right|^{\frac{2}{2-d}} \tag{30}$$

for suitable positive constants C_1 and C_2.

(b) *If*

$$\int_0^\infty V(t)\, g_0(t)\, dt = \int_\Gamma V(|x|)\, dx > 0,$$

then the discrete spectrum of A_λ is empty for λ positive and small enough.

Proof. Part (a). From (10) we get

$$E^-_{1,0}(\lambda) \leq E_{1,0}(\lambda) \leq E^+_{1,0}(\lambda).$$

Moreover, by Lemma 1, $E^\pm_{1,0}(\lambda)$ are the lowest eigenvalues of operators $B^\pm_{\lambda,0}$. The existence and uniqueness of $E_{1,0}$ thus follows from part (a) of Theorem 2 applied with $W(t) = V_0^+(t)$ and $W(t) = V_0^-(t)$, respectively. At the same time, (26) implies (30).

Similarly, part (b) of the statement follows immediately from Lemma 1 and part (b) of Theorem 2 applied with $W(t) = V_0^-(t)$. □

Remark 4. We note that the strong coupling behavior of A_λ is, contrary to (30), typically one-dimensional, i.e., determined by the local dimension of Γ. Namely, if V is continuous and compactly supported, then the standard Dirichlet–Neumann bracketing technique shows that the Weyl asymptotic formula

$$\lim_{\lambda\to\infty} \lambda^{-\gamma-\frac{1}{2}} \sum_j |E_j|^\gamma = L^{cl}_{\gamma,1} \int_\Gamma |V|^{\gamma+\frac{1}{2}}\, dx, \quad \gamma \geq 0,$$

holds true, where E_j are the negative eigenvalues of A_λ and $L^{cl}_{\gamma,1} = \frac{\Gamma(\gamma+1)}{2\sqrt{\pi}\,\Gamma(\gamma+3/2)}$.

Remark 5. Notice that our result qualitatively agrees with the precise asymptotic formula for $\varepsilon(\lambda)$ on branching graphs with one vertex and finitely many edges which was found in [5]. Such graphs correspond to $d = 1$ in our setting.

5.2. The case $d = 2$. For $d = 2$ one can mimic the above procedure by replacing the Hankel functions $H^{(1,2)}_\nu$ by $H^{(1,2)}_0$. The latter have a logarithmic singularity at zero, and therefore it turns out that the lowest eigenvalue of A_λ then converges to zero exponentially fast. Indeed, here instead of (26) one obtains

$$E(\lambda) \sim -e^{-\lambda^{-1}}, \quad \lambda \to 0,$$

as for the two-dimensional Schrödinger operator; see [11]. Since the analysis of this case is completely analogous to the previous one, we skip it.

5.3. The case $d > 2$. Here we will show, under some assumptions on V, that for $d > 2$ and λ small enough the discrete spectrum of A_λ remains empty no matter what the sign of $\int_\Gamma V$ is.

Proposition 2. *Let $d > 2$, and let V satisfy Assumption A. If $V \in L^\infty(\mathbb{R}_+) \cap L^{p/2}(\mathbb{R}_+, g_0)$, with $p < d$, then there exists $\lambda_0 > 0$ such that the discrete spectrum of A_λ is empty for all $\lambda \in [0, \lambda_0]$.*

Proof. From the definition of the function g_k it follows that

$$\frac{\int_{t_k}^{\infty}\left(|f'|^2+\lambda V|f|^2\right)g_k(t)\,dt}{\int_{t_k}^{\infty}|f|^2 g_k(t)\,dt}=\frac{\int_{t_k}^{\infty}\left(|f'|^2+\lambda V|f|^2\right)g_0(t)\,dt}{\int_{t_k}^{\infty}|f|^2 g_0(t)\,dt}.$$

Since every function $f\in H_0^1((t_k,\infty),g_k)$ can be extended by zero to a function in $H^1(\mathbb{R}_+,g_0)$, the variational principle shows that

$$\sigma_d(A_{\lambda,0})=\emptyset \implies \sigma_d(A_{\lambda,k})=\emptyset \quad \forall\, k\geq 1.$$

Hence it suffices to prove the statement for the operator $A_{\lambda,0}$, i.e., to show that $A_{\lambda,0}$ is nonnegative. Consider a function $f\in D(Q_0)$. Since $f\in H^1(\mathbb{R}_+)$, which is continuously embedded in $L^\infty(\mathbb{R}_+)$, it follows that $f\to 0$ at infinity, and we can write

$$f(t)=-\int_t^\infty f'(s)\,ds.$$

In view of (6) we have $g_0^{-1}\in L^1(\mathbb{R}_+)$. Using the Cauchy–Schwarz inequality we thus find out that for any $q>q_0$, where $\frac{1}{q_0}+\frac{1}{d}=\frac{1}{2}$, the following estimate holds true:

$$\begin{aligned}\left(\int_0^\infty |f(t)|^q\,g_0(t)\,dt\right)^{\frac{1}{q}} &\leq \left(\int_0^\infty\left(\int_t^\infty |f'(s)|ds\right)^q g_0(t)\,dt\right)^{\frac{1}{q}}\\ &\leq\left(\int_0^\infty\left(\int_t^\infty |f'(s)|^2 g_0(s)\,ds\right)^{\frac{q}{2}}\left(\int_t^\infty\frac{ds}{g_0(s)}\right)^{\frac{q}{2}} g_0(t)\,dt\right)^{\frac{1}{q}}\\ &\leq C(q)\left(\int_0^\infty |f'(s)|^2\,g_0(s)\,ds\right)^{\frac{1}{2}}, \qquad (31)\end{aligned}$$

with a constant $C(q)$ independent of f. Take q such that $\frac{1}{q}+\frac{1}{p}=\frac{1}{2}$. The Hölder inequality and (31) then give

$$\begin{aligned}\int_0^\infty |V|\,|f|^2 g_0(t)\,dt &\leq\left(\int_0^\infty |V|^{p/2}\,g_0(t)\,dt\right)^{\frac{2}{p}}\left(\int_0^\infty |f|^q\,g_0(t)\,dt\right)^{\frac{2}{q}}\\ &\leq C^2(q)\int_0^\infty |f'|^2\,g_0(t)\,dt\left(\int_0^\infty |V|^{p/2}\,g_0(t)\,dt\right)^{\frac{2}{p}},\end{aligned}$$

which implies

$$Q_0[f]\geq\int_0^\infty |f'|^2\,g_0(t)\,dt\left[1-\lambda\,C^2(q)\left(\int_0^\infty |V|^{p/2}\,g_0(t)\,dt\right)^{\frac{2}{p}}\right].$$

To show that the negative spectrum of $A_{\lambda,0}$ is empty, it suffices to take λ small enough so that $Q_0[f]\geq 0$. $\square$

Appendix.

LEMMA 2. *Let W be bounded, and assume that $\int_0^\infty(1+t)^{1+2\nu}\,|W(t)|\,dt<\infty$. Then $M(\kappa)$ converges in the Hilbert–Schmidt norm to the operator $M(0)$ as $\kappa\to 0$.*

Proof. We first notice that $M(0)$ is Hilbert–Schmidt, since

$$\int_0^\infty\int_0^\infty |M(t,t',0)|^2\,dt\,dt'<\infty$$

by assumption. We will also need the asymptotic behavior of the Bessel functions with a purely imaginary argument near zero:

$$J_\nu(i\kappa(1+t)) = e^{i\pi\nu/2}\, I_\nu(\kappa(1+t)) \sim e^{i\pi\nu/2}\, \frac{\kappa^\nu(1+t)^\nu}{2^\nu\,\Gamma(\nu+1)}, \quad \kappa(1+t)\to 0; \tag{32}$$

see [1, 14]. From the definition of Hankel's functions we thus get

$$\beta(\kappa) = \frac{J_{1-\nu}(i\kappa) - e^{i(1-\nu)\pi}\, J_{\nu-1}(i\kappa)}{e^{i(\nu-1)\pi}\, J_{\nu-1}(i\kappa) - J_{1-\nu}(i\kappa)} \to -e^{-2i\nu\pi}, \quad \kappa\to 0.$$

This together with the asymptotics (32) implies

$$\lim_{\kappa\to 0} M(t,t',\kappa) = M(t,t',0)\,. \tag{33}$$

Now using the asymptotic behavior of Hankel's functions at infinity [1], we find out that

$$G(t,t',\kappa) \sim ((1+t)(1+t'))^{1/2}\ \frac{e^{-\kappa(t+t')} - \beta(\kappa)\, e^{-\kappa|t-t'|}}{\kappa\,(1+t)(1+t')^{1/2}}, \quad \kappa^2(1+t)(1+t')\to\infty.$$

Since $|\beta(\kappa)|$ is bounded, we obtain the following estimates:

(i). For $\kappa^2(1+t)(1+t')\geq 1$:

$$|K(t,t',\kappa)|, |L(t,t',\kappa)| \leq C\ |W(t')\,W(t)(1+t)(1+t')|^{1/2}\,.$$

(ii). For $\kappa^2(1+t)(1+t')<1$:

$$|M(t,t',\kappa)| \leq C'\,|W(t')\,W(t)\,|\left[1+((1+t)(1+t'))^{\nu+\frac{1}{2}}\right],$$

where we have used (33). Note that the constants C and C' may be chosen independent of κ, which enables us to employ the Lebesgue dominated convergence theorem to conclude that

$$\lim_{\kappa\to 0}\int_{\mathbb{R}_+^2} |M(t,t',\kappa) - M(t,t',0)|^2\, dt\, dt' = 0. \qquad \square$$

REFERENCES

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, 1964.

[2] V. Bargmann, *On the number of bound states in a central field of force*, Proc. Natl. Acad. Sci. USA, 38 (1952). pp. 961–966.

[3] M.S. Birman and M.Z. Solomyak, *Schrödinger operator. Estimates for number of bound states as function-theoretical problem*, Amer. Math. Soc. Transl. (2), 150 (1992), pp. 1–54.

[4] R. Carlson, *Nonclassical Sturm-Liouville problems and Schrödinger operators on radial trees*, Electron. J. Differential Equations, 71 (2000), pp. 1–24.

[5] P. Exner, *Weakly coupled states on branching graphs*, Lett. Math. Phys., 38 (1996), pp. 313–320.

[6] P. Kuchment, *Quantum graphs:* I. *Some basic structures*, Waves Random Media, 14 (2004), pp. 107–128.

[7] V. Kostrykin and R. Schrader, *Kirchhoff's rules for quantum wires*, J. Phys. A, 32 (1999), pp. 595–630.

[8] K. Naimark and M. Solomyak, *Geometry of the Sobolev spaces on the regular trees and Hardy's inequalities*, Russ. J. Math. Phys., 8 (2001), pp. 322–335.

[9] K. Naimark and M. Solomyak, *Eigenvalue estimates for the weighted Laplacian on metric trees*, Proc. London Math. Soc., 80 (2000), pp. 690–724.

[10] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* IV, Academic Press, New York, 1978.

[11] B. Simon, *The bound state of weakly coupled Schrödinger operators in one and two dimensions*, Ann. Physics, 97 (1976), pp. 279–288.

[12] A. Sobolev and M. Solomyak, *Schrödinger operators on homogeneous metric trees: Spectrum in gaps*, Rev. Math. Phys., 14 (2002), pp. 421–467.

[13] M. Solomyak, *On the spectrum of the Laplacian on metric trees*, Waves Random Media, 14 (2004), pp. S155–S171.

[14] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, Cambridge, 1958.

SIAM J. MATH. ANAL.
Vol. 39, No. 4, pp. 1150–1190

TOTAL VARIATION REGULARIZATION FOR IMAGE DENOISING, I. GEOMETRIC THEORY*

WILLIAM K. ALLARD†

Abstract. Let Ω be an open subset of $\mathbf{R}^n$, where $2 \leq n \leq 7$; we assume $n \geq 2$ because the case $n = 1$ has been treated elsewhere (see [S. S. Alliney, *IEEE Trans. Signal Process.*, 40 (1992), pp. 1548–1562] and is quite different from the case $n > 1$; we assume $n \leq 7$ because we will make use of the regularity theory for area minimizing hypersurfaces. Let $\mathcal{F}(\Omega) = \{f \in \mathbf{L}_1(\Omega) \cap \mathbf{L}_\infty(\Omega) : f \geq 0\}$. Suppose $s \in \mathcal{F}(\Omega)$ and $\gamma : \mathbb{R} \to [0, \infty)$ is locally Lipschitzian, positive on $\mathbb{R} \sim \{0\}$, and zero at zero. Let $F(f) = \int_\Omega \gamma(f(x) - s(x))\, d\mathcal{L}^n x$ for $f \in \mathcal{F}(\Omega)$; here $\mathcal{L}^n$ is Lebesgue measure on $\mathbb{R}^n$. Note that $F(f) = 0$ if and only if $f(x) = s(x)$ for $\mathcal{L}^n$ almost all $x \in \mathbb{R}^n$. In the denoising literature F would be called a *fidelity* in that it measures deviation from s, which could be a noisy grayscale image. Let $\epsilon > 0$ and let $F_\epsilon(f) = \epsilon \mathbf{TV}(f) + F(f)$ for $f \in \mathcal{F}(\Omega)$; here $\mathbf{TV}(f)$ is the total variation of f. A minimizer of F_ϵ is called a *total variation regularization of* s. Rudin, Osher, and Fatemi and Chan and Esedoḡlu have studied total variation regularizations where $\gamma(y) = y^2$ and $\gamma(y) = |y|$, $y \in \mathbb{R}$, respectively. As these and other examples show, the geometry of a total variation regularization is quite sensitive to changes in γ. Let f be a total variation regularization of s. The first main result of this paper is that the reduced boundaries of the sets $\{f > y\}$, $0 < y < \infty$, are embedded $C^{1,\mu}$ hypersurfaces for any $\mu \in (0, 1)$ where $n > 2$ and any $\mu \in (0, 1]$ where $n = 2$; moreover, the generalized mean curvature of the sets $\{f \geq y\}$ will be bounded in terms of y, ϵ and the magnitude of $|s|$ near the point in question. In fact, this result holds for a more general class of fidelities than those described above. A second result gives precise curvature information about the reduced boundary of $\{f > y\}$ in regions where s is smooth, provided F is convex. This curvature information will allow us to construct a number of interesting examples of total variation regularizations in this and in a subsequent paper. In addition, a number of other theorems about regularizations are proved.

Key words. total variation, regularization, image denoising

AMS subject classifications. 49Q20, 58E30

DOI. 10.1137/060662617

1. Introduction and statement of main results.

Throughout this paper, n is an integer, $2 \leq n \leq 7$, Ω is an open subset of $\mathbb{R}^n$, and $\mathcal{L}^n$ is Lebesgue measure on $\mathbb{R}^n$.

We require $n \geq 2$ because the problems we consider are very different in case $n = 1$; see [Alli]. We require $n \leq 7$ because we will be using the regularity theory of mass minimizing integral currents in $\mathbb{R}^n$ of codimension one; as is well known, these currents are free of interior singularities when $n \leq 7$ but may possess singularities if $n > 7$; see [FE, sect. 5.4.15]. This work is motivated by image denoising applications in which it is often the case that $1 \leq n \leq 4$.

After a fairly lengthy discussion of results which occur in a setting more general than that of denoising, we treat denoising in section 1.8. See also sections 1.9, 8, and 10 as well as the examples in section 11 for more on denoising.

1.1. Some basic notation and conventions.

Whenever $E \subset \Omega$ we frequently identify "E" with "1_E," the indicator function of E.

The first appearance of any term which is about to be defined will always appear in italics or be displayed.

*Received by the editors June 9, 2006; accepted for publication (in revised form) April 20, 2007; published electronically November 16, 2007. This work was supported in part by Los Alamos National Laboratory.
http://www.siam.org/journals/sima/39-4/66261.html
†Department of Mathematics, Duke University, Durham, NC 27708-0320 (wka@math.duke.edu).

We let

$$\mathcal{F}(\Omega) = \{f \in \mathbf{L}_1(\Omega) \cap \mathbf{L}_\infty(\Omega) : f \geq 0\}$$

and

$$\mathcal{M}(\Omega) = \{D : D \subset \Omega \text{ and } 1_D \in \mathcal{F}(\Omega)\}\,;$$

thus a subset D of Ω belongs to $\mathcal{M}(\Omega)$ if and only if D is Lebesgue measurable and $\mathcal{L}^n(D) < \infty$. We endow $\mathbf{L}_1^{loc}(\Omega)$ with the topology induced by the seminorms $\mathbf{L}_1^{loc}(\Omega) \ni f \mapsto \int_K |f|\, d\mathcal{L}^n$ corresponding to compact subsets K of Ω. Whenever $f \in \mathbf{L}_1^{loc}(\Omega)$ and K is a compact subset of Ω we let

$$\mathbf{k}(f, K) = \{g \in \mathbf{L}_1^{loc}(\Omega) : g(x) = f(x) \text{ for } \mathcal{L}^n \text{ almost all } x \in \Omega \sim K\};$$

in other words, $g \in \mathbf{k}(f, K)$ if the support of the generalized function corresponding to $g - f$ is a subset of K. We let

$$\mathbf{k}(f) = \cup\{\mathbf{k}(f, K) : K \text{ is a compact subset of } \Omega\}.$$

Whenever D is a Lebesgue measurable subset of Ω and K is a compact subset of Ω we let

$$\mathbf{k}(D, K) = \{E : E \subset \Omega \text{ and } 1_E \in \mathbf{k}(1_D, K)\}$$

and

$$\mathbf{k}(D) = \cup\{\mathbf{k}(D, K) : K \text{ is a compact subset of } \Omega\}.$$

Whenever A, D, E are Lebesgue measurable subsets of Ω we let

$$\Sigma_A(D, E) = \mathcal{L}^n(A \cap ((D \sim E) \cup (E \sim D))) = \int_A |1_D - 1_E|\; d\mathcal{L}^n;$$

note that $\mathcal{M}(\Omega) \times \mathcal{M}(\Omega) \ni (D, E) \mapsto \Sigma_A(D, E)$ is a pseudometric on $\mathcal{M}(\Omega)$.

Whenever $a \in \mathbb{R}^n$ and $0 < r < \infty$ we let

$$\mathbf{U}^n(a, r) = \{x \in \mathbb{R}^n : |x - a| < r\} \quad \text{and} \quad \mathbf{B}^n(a, r) = \{x \in \mathbb{R}^n : |x - a| \leq r\}.$$

We let

$$\mathbf{int}, \quad \mathbf{cl}, \quad \text{and} \quad \mathbf{bdry}$$

stand for "interior," "closure," and "boundary," respectively, with respect to Ω.

Whenever $A \subset \mathbb{R}^n$ and a is an accumulation point of A we let

$$\mathbf{Tan}(A, a) = \bigcap_{0<r<\infty} \mathbf{cl}\,\{t(x - a) : 0 < t < \infty \text{ and } x \in A \cap (\mathbf{B}^n(a, r) \sim \{a\})\}$$

and

$$\mathbf{Nor}(A, a) = \bigcap_{w \in \mathbf{Tan}(A,a)} \{v \in \mathbb{R}^n : v \bullet w \leq 0\}.$$

Whenever $0 \le m < \infty$ we let

$$\mathcal{H}^m$$

be an m-dimensional Hausdorff measure on $\mathbb{R}^n$.

We let

$$\mathcal{X}(\Omega)$$

be the space of smooth compactly supported vector fields on Ω.

Whenever $y, z \in \mathbb{R}$ we let

$$y \vee z = \max\{y, z\}, \qquad y \wedge z = \min\{y, z\},$$

and we note that $y + z = y \vee z + y \wedge z$.

1.2. Total variation. This work is based on the notion of the total variation of a locally summable function, which we now define.

DEFINITION 1.1. *Suppose $f \in \mathbf{L}_1^{loc}(\Omega)$. Then $\mathbf{TV}(f, \cdot)$, the* total variation *of f, is the largest Borel regular measure on Ω such that, for any open subset U of Ω,*

$$\mathbf{TV}(f, U) = \sup\left\{ \int_U f \,\mathbf{div}\, X \, d\mathcal{L}^n : X \in \mathcal{X}(U) \text{ and } |X| \le 1 \right\}.$$

In particular, if f is C^1 on Ω and B is a Borel subset of Ω, then

$$\mathbf{TV}(f, B) = \int_B |\nabla f| \, d\mathcal{L}^n. \tag{1.1}$$

Moreover, if E a Lebesgue measurable subset of Ω with Lipschitz boundary, then $\mathbf{TV}(E, B)$ equals the $(n-1)$-dimensional Hausdorff measure of the intersection of the boundary of E with B.

Suppose $f \in \mathbf{L}_1^{loc}(\Omega)$. We say f is of *bounded variation on* Ω if $\mathbf{TV}(f, \Omega)$ is finite. If $\mathbf{TV}(f, \cdot)$ is a Radon measure on Ω, which will be the case if and only if $\mathbf{TV}(f, K) < \infty$ whenever K is a compact subset of Ω, we say f is of *locally bounded variation on* Ω. We let

$$\mathbf{BV}(\Omega) \quad \text{and} \quad \mathbf{BV}^{loc}(\Omega)$$

be the vector spaces of those $f \in \mathbf{L}_1(\Omega)$ which are of bounded variation on Ω and those $f \in \mathbf{L}_1^{loc}(\Omega)$ which are of locally bounded variation on Ω, respectively.

If E is a Lebesgue measurable subset of Ω, the *perimeter of* E is, by definition, $\mathbf{TV}(E, \Omega)$; we say E is of *locally finite perimeter* if $E \in \mathbf{BV}^{loc}(\Omega)$. As is well known, if $f \in \mathbf{BV}^{loc}(\Omega)$, then $\{f > y\}$ is of locally finite perimeter for $\mathcal{L}^1$ almost all y. As is well known, sets of locally finite perimeter have nice rectifiability properties; see section 2.8 below.

1.3. Total variation regularization.

DEFINITION 1.2. *Suppose $F : \mathcal{F}(\Omega) \to \mathbb{R}$ and $0 < \epsilon < \infty$. We let*

$$F_\epsilon : \mathcal{F}(\Omega) \to \mathbb{R} \cup \{\infty\},$$

the total variation regularization of F (with respect to ϵ)*, be defined by setting*

$$F_\epsilon(g) = \epsilon \mathbf{TV}(g) + F(g) \quad \text{for } g \in \mathcal{F}(\Omega).$$

We let

$$\mathbf{m}_\epsilon^{loc}(F) = \{f \in \mathcal{F}(\Omega) \cap \mathbf{BV}(\Omega) : F_\epsilon(f) \leq F_\epsilon(g) \text{ whenever } g \in \mathcal{F}(\Omega) \cap \mathbf{k}(f)\}.$$

All of the statements and proofs of this paper, after straightforward modification, go through with the condition "$f \geq 0$," omitted in the definition of $\mathcal{F}(\Omega)$; however, the modified statements and proofs often break into two cases because if $f \in \mathbf{L}_1(\Omega)$ and $y \in \mathbb{R}$, one can only be sure that $\mathcal{L}^n(\{f > y\}) < \infty$ if $y > 0$ and that $\mathcal{L}^n(\{f < y\}) < \infty$ if $y < 0$.

It will be useful to extend the foregoing notions to functionals defined on sets, as follows.

DEFINITION 1.3. *Suppose $M : \mathcal{M}(\Omega) \to \mathbb{R}$ and $0 < \epsilon < \infty$. We let*

$$M_\epsilon : \mathcal{M}(\Omega) \to \mathbb{R} \cup \{\infty\},$$

the total variation regularization of M (with respect to ϵ)*, be defined by setting*

$$M_\epsilon(E) = \epsilon\mathbf{TV}(E) + M(E) \quad \text{for } E \in \mathcal{M}(\Omega).$$

We let

$$\mathbf{n}_\epsilon^{loc}(M) = \{D \in \mathcal{M}(\Omega) \cap \mathbf{BV}(\Omega) : M_\epsilon(D) \leq M_\epsilon(E) \text{ whenever } E \in \mathcal{M}(\Omega) \cap \mathbf{k}(D)\}.$$

The main purpose of this paper is to study the geometry and regularity of the sets $\{f > y\}$ and $\{f \geq y\}$, $y \in (0, \infty)$, when $f \in \mathbf{m}_\epsilon^{loc}(F)$, provided F satisfies certain conditions which we now describe. We will relate these results to certain methods for image denoising.

1.4. Admissibility. Suppose

$$F : \mathcal{F}(\Omega) \to \mathbb{R}.$$

All our results will require F to be *admissible*, a notion we now define.

DEFINITION 1.4. *We say F is* admissible *if the restriction of F to any subset of $\mathcal{F}(\Omega)$ which is bounded with respect to $||\cdot||_{\mathbf{L}_\infty(\Omega)}$ is Lipschitz with respect to $||\cdot||_{\mathbf{L}_1(\Omega)}$.*

In other words, F is admissible if whenever $0 < Y < \infty$ we have

$$\mathbf{l}(F, Y) < \infty,$$

where $\mathbf{l}(F, Y)$ is the infimum of the set of $L \in (0, \infty)$ such that

$$|F(f) - F(g)| \leq L \int_\Omega |f - g| \, d\mathcal{L}^n$$

whenever $f, g \in \mathcal{F}(\Omega)$ and $\max\{||f||_{\mathbf{L}_\infty(\Omega)}, ||g||_{\mathbf{L}_\infty(\Omega)}\} \leq Y$.

The notion of admissibility extends naturally to functionals on sets, as follows.

DEFINITION 1.5. *Suppose $M : \mathcal{M}(\Omega) \to \mathbb{R}$. We let*

$$\mathbf{l}(M)$$

be the infimum of the set of $L \in (0, \infty)$ such that

$$|M(D) - M(E)| \leq L\, \Sigma_\Omega(D, E) \quad \text{whenever } D, E \in \mathcal{M}(\Omega).$$

We say M is admissible *if $\mathbf{l}(M) < \infty$.*

1.5. $\mathcal{B}_\lambda(\Omega)$ and $\mathcal{C}_\lambda(\Omega)$. These spaces will be indispensable in this work.

DEFINITION 1.6. *Suppose* $0 \le \lambda < \infty$. *We let*

$$\mathcal{B}_\lambda(\Omega)$$

be the set of those $f \in \mathbf{BV}^{loc}(\Omega)$ *such that for each compact subset* K *of* Ω *we have*

$$\mathbf{TV}(f,K) \le \mathbf{TV}(g,K) + \lambda \int_\Omega |f-g|\, d\mathcal{L}^n \quad \textit{whenever } g \in \mathbf{k}(f,K).$$

We let

$$\mathcal{C}_\lambda(\Omega)$$

be the set of those Lebesgue measurable subsets D *of* Ω *with locally finite perimeter such that for each compact subset* K *we have*

$$\mathbf{TV}(D,K) \le \mathbf{TV}(E,K) + \lambda \Sigma_\Omega(D,E) \quad \textit{whenever } E \in \mathbf{k}(D,K).$$

The following result is based on ideas found in [BDG].

THEOREM 1.1. *Suppose* $0 \le \lambda < \infty$ *and* $f \in \mathbf{BV}^{loc}(\Omega)$. *Then*

$$f \in \mathcal{B}_\lambda(\Omega) \;\Rightarrow\; \{f > y\} \in \mathcal{C}_\lambda(\Omega) \quad \textit{for } y \in \mathbb{R}.$$

Conversely, if Y *is a dense subset of* $\mathbb{R}$, *then*

$$\{f > y\} \in \mathcal{C}_\lambda(\Omega) \textit{ for } y \in Y \;\Rightarrow\; f \in \mathcal{B}_\lambda(\Omega).$$

An immediate corollary is that $D \in \mathcal{C}_\lambda(\Omega)$ if and only if $1_D \in \mathcal{B}_\lambda(\Omega)$.

More results on $\mathcal{C}_\lambda(\Omega)$ and $\mathcal{B}_\lambda(\Omega)$ may be found in section 5.1.

1.5.1. The regularity theorem for $\mathcal{C}_\lambda(\Omega)$. The proof of the following theorem is an exercise, carried out in section 5.4, in the use of techniques from geometric measure theory which have been in the literature for over thirty years.

Note that in the following theorem, θ *does not depend on* D.

THEOREM 1.2 (regularity). *Suppose* $0 < \mu < \infty$ *and* $0 < \beta < 1$. *There exists* θ *such that* $0 < \theta < 1$ *and with the following property:*

Suppose

(i) $a \in \mathbb{R}^n$ *and* $0 < R < \infty$;

(ii) $0 \le \lambda < \infty$, $\lambda R \le \theta$, *and* $D \in \mathcal{C}_\lambda(\mathbf{U}^n(a,R))$;

(iii) Γ *is the interior of the support of the generalized function corresponding to the indicator function of* D *and* M *is the boundary of* Γ *relative to* $\mathbf{U}^n(a,R)$.

Then $\Sigma_{\mathbf{U}^n(a,R)}(D,\Gamma) = 0$ *and* M *is an embedded hypersurface in* $\mathbf{U}^n(a,R)$ *of class* $C^{1,\mu}$; *moreover, if* N *is a continuous field of unit normals to* M *and* $r = \theta R$, *then*

$$|N(x) - N(w)| \le \beta\, (|x-w|/r)^\mu \quad \textit{whenever } x, w \in M \cap \mathbf{U}^n(a,r);$$

finally, if L *is a line perpendicular to* $\mathbf{Tan}(M,a)$, *then* L *intersects* $M \cap \mathbf{U}^n(a,r)$ *in at most one point.*

In case $n = 2$ *we may take* $\mu = 1$.

1.5.2. The relationship between admissibility and $\mathcal{C}_\lambda(\Omega)$. The following simple proposition relates the notion of admissibility to the spaces $\mathcal{B}_\lambda(\Omega)$.

PROPOSITION 1.1. *Suppose $F : \mathcal{F}(\Omega) \to \mathbb{R}$, F is admissible, $0 < \epsilon < \infty$, $f \in \mathbf{m}_\epsilon^{loc}(F)$, $Y = ||f||_{\mathbf{L}_\infty(\Omega)}$, and $\lambda = \mathbf{l}(F, Y)/\epsilon$.*

Then $f \in \mathcal{B}_\lambda(\Omega)$.

Proof. Suppose $g \in \mathbf{k}(f, K)$. Let $h = (g \wedge Y) \vee 0$. Then $h \in \mathbf{k}(f, K)$, so

$$\epsilon\,(\mathbf{TV}(f, K) - \mathbf{TV}(h, K)) \le F(h) - F(f) \le \mathbf{l}(F, Y) \int_\Omega |f - h|.$$

As is well known and shown in Proposition 2.3 below, $\mathbf{TV}(h, K) \le \mathbf{TV}(g, K)$, and it is evident that $\int_\Omega |f - h|\, d\mathcal{L}^n \le \int_\Omega |f - g|\, d\mathcal{L}^n$, so the proposition is proved. ∎

We leave the even simpler proof of the following proposition to the reader.

PROPOSITION 1.2. *Suppose $M : \mathcal{M}(\Omega) \to \mathbb{R}$, M is admissible, $0 < \epsilon < \infty$, $D \in \mathbf{n}_\epsilon^{loc}(M)$, and $\lambda = \mathbf{l}(M)/\epsilon$.*

Then $D \in \mathcal{C}_\lambda(\Omega)$.

Remark 1.1. Thus if $f \in \mathbf{m}_\epsilon^{loc}(F)$, where F is admissible, the regularity theorem, Theorem 1.2, for $\mathcal{C}_\lambda(\Omega)$ applies to the sets $\{f > y\}$, $0 < y < \infty$. In particular, if $n > 2$ and $0 < \mu < 1$ or if $n = 2$ and $0 < \mu \le 1$, the boundaries of the support of $[\{f > y\}]$, $0 < y < \infty$, are always embedded $C^{1,\mu}$ hypersurfaces.

In order to obtain yet more information about $\{f > y\}$ we need to assume more about F, as follows.

1.6. Locality.

DEFINITION 1.7. *Suppose $F : \mathcal{F}(\Omega) \to \mathbb{R}$. We say F is* local *if F is admissible and*

$$\hat{F}(f + g) = \hat{F}(f) + \hat{F}(g) \quad \textit{whenever } f, g \in \mathcal{F}(\Omega) \textit{ and } fg = 0,$$

where we have set

$$\hat{F}(f) = F(f) - F(0) \quad \textit{for } f \in \mathcal{F}(\Omega).$$

The notion of locality extends naturally to functionals on sets, as follows.

DEFINITION 1.8. *Suppose $M : \mathcal{M}(\Omega) \to \mathbb{R}$. We say M is* local *if M is admissible and*

$$\hat{M}(D \cup E) = \hat{M}(D) + \hat{M}(E) \quad \textit{whenever } D, E \in \mathcal{M}(\Omega) \textit{ and } D \cap E = \emptyset,$$

where we have set

$$\hat{M}(E) = M(E) - M(\emptyset) \quad \textit{for } E \in \mathcal{M}(\Omega).$$

The proofs of the following four propositions, which we carry out in section 6, are exercises in real variable theory.

PROPOSITION 1.3. *Suppose $M : \mathcal{M}(\Omega) \to \mathbb{R}$, M is admissible, and*

$$\mu(x) = \limsup_{r \downarrow 0} \frac{\hat{M}(\mathbf{B}^n(x, r))}{\mathcal{L}^n(\mathbf{B}^n(x, r))} \quad \textit{for } x \in \Omega.$$

Then M is local if and only if

$$M(E) = M(\emptyset) + \int_E m\, d\mathcal{L}^n \quad \textit{for } E \in \mathcal{M}(\Omega) \tag{1.2}$$

for some bounded Borel function $m : \Omega \to \mathbb{R}$, in which case $m(x) = \mu(x)$ for $\mathcal{L}^n$ almost all $x \in \Omega$.

PROPOSITION 1.4. *Suppose $F : \mathcal{F}(\Omega) \to \mathbb{R}$, F is admissible, and*

$$\kappa(x,y) = \limsup_{r\downarrow 0} \frac{\hat{F}(y 1_{\mathbf{B}^n(x,r)})}{\mathcal{L}^n(\mathbf{B}^n(x,r))} \quad \textit{for } (x,y) \in \Omega \times [0,\infty).$$

Then F is local if and only if

$$F(f) = F(0) + \int_\Omega k(x, f(x))\, d\mathcal{L}^n x \quad \textit{whenever } f \in \mathcal{F}(\Omega) \tag{1.3}$$

for some Borel function $k : \Omega \times [0,\infty) \to \mathbb{R}$ such that

(i) *$k(x,0) = 0$ for $\mathcal{L}^n$ almost all $x \in \Omega$;*

(ii) *whenever $0 < Y < \infty$ there is $L \in [0,\infty)$ such that if $0 \le y < z < Y$, then*

$$|k(x,y) - k(x,z)| \le L|y-z| \quad \textit{for } \mathcal{L}^n \textit{ almost all } x \in \Omega,$$

in which case, for each $y \in [0,\infty)$, $k(x,y) = \kappa(x,y)$ for $\mathcal{L}^n$ almost $x \in \Omega$.

Remark 1.2. Suppose $\beta : \mathbb{R} \to \mathbb{R}$ is locally Lipschitzian and

$$F(f) = \beta\left(\int_\Omega f\, d\mathcal{L}^n\right) \quad \text{for } f \in \mathcal{F}(\Omega).$$

It is evident that F is admissible but not local unless β is an affine function.

Perhaps a more interesting example is as follows. Suppose $K, s \in \mathbf{L}_1(\mathbb{R}^n)$ and

$$F(f) = \int_{\mathbb{R}^n} |K * f - s|\, d\mathcal{L}^n \quad \text{for } f \in \mathcal{F}(\mathbb{R}^n).$$

Evidently, $\mathbf{l}(F,Y) = ||K||_{\mathbf{L}_1(\mathbb{R}^n)} < \infty$ whenever $0 < Y < \infty$, so F is admissible. If κ is as in Proposition 1.4, we find that $\kappa(x,y) = |K(x)y - s(x)|$ for $\mathcal{L}^n$ almost all x and all $y \in (0,\infty)$. It is easy to see that F is not local if both $\{K > 0\}$ and $\{K < 0\}$ have positive Lebesgue measure.

PROPOSITION 1.5. *Suppose $F : \mathcal{F}(\Omega) \to \mathbb{R}$, F is admissible, κ is as in Proposition 1.4,*

$$u(x,y) = \limsup_{z\to y} \frac{\kappa(x,z) - \kappa(x,y)}{z-y} \quad \textit{for } (x,y) \in \Omega \times (0,\infty),$$

and, for each $y \in (0,\infty)$,

$$U_y(E) = \limsup_{z\to y} \frac{F(z1_E) - F(y1_E)}{z-y} \quad \textit{for } E \in \mathcal{M}(\Omega).$$

Then

(i) *u is a Borel function and $|u(x,y)| \le l(F,Y)$ whenever $x \in \Omega$ and $0 < y < Y < \infty$;*

(ii) *$|U_y(E)| \le l(F,Y)\mathcal{L}^n(E)$ whenever $0 < y < Y$ and $E \in \mathcal{M}(\Omega)$;*

(iii) *for any $f \in \mathcal{F}(\Omega)$, $(0,\infty) \ni y \mapsto U_y(\{f > y\})$ is a Borel function.*

Moreover, if F is local, then

(iv) *for* $\mathcal{L}^1$ *almost all* $y \in (0,\infty)$, U_y *is local,* $\mathbf{l}(U_y) \leq \mathbf{l}(F,Y)$ *whenever* $y < Y < \infty$, *and*

$$U_y(E) = \int_E u(x,y)\, d\mathcal{L}^n x \quad \textit{whenever } E \in \mathcal{M}(\Omega);$$

(v) *for any* $f \in \mathcal{F}(\Omega)$,

$$F(f) = F(0) + \int_0^\infty U_y(\{f > y\})\, d\mathcal{L}^1 y.$$

PROPOSITION 1.6. *Suppose* $F : \mathcal{F}(\Omega) \to \mathbb{R}$, F *is local,* κ *is as in Proposition* 1.4, *and* U_y, $0 < y < \infty$, *and* u *are as in Proposition* 1.5.

The following are equivalent:

(i) F *is convex.*

(ii) $\mathbb{R} \ni y \mapsto F(y 1_E)$ *is convex for any* $E \in \mathcal{M}(\Omega)$.

(iii) *For any* $x \in \Omega$,

$$\mathbb{R} \ni y \mapsto \kappa(x,y) \textit{ is convex.}$$

(iv) $\mathbb{R} \ni y \mapsto U_y(E)$ *is nondecreasing for any* $E \in \mathcal{M}(\Omega)$.

(v) *For any* $x \in \Omega$,

$$\mathbb{R} \ni y \mapsto u(x,y) \textit{ is nondecreasing.}$$

Moreover, if F *is convex and* $0 < y < \infty$, *then* U_y *is local and*

$$\lim_{z \downarrow y} U_z(E) = U_y(E) \quad \textit{whenever } E \in \mathcal{F}(\Omega). \tag{1.4}$$

Remark 1.3. Suppose $F : \mathcal{F}(\Omega) \to \mathbb{R}$, F is admissible, κ is as in Proposition 1.4,

$$l(x,y) = \liminf_{z \to y} \frac{\kappa(x,z) - \kappa(x,y)}{z - y} \quad \text{for } (x,y) \in \Omega \times (0,\infty),$$

and, for each $y \in (0,\infty)$,

$$L_y(E) = \liminf_{z \to y} \frac{F(z 1_E) - F(y 1_E)}{z - y} \quad \text{for } E \in \mathcal{M}(\Omega).$$

Modifying the proof of Proposition 1.5 in a straightforward way one finds that this proposition holds with u and U_y, $0 < y < \infty$, replaced by l and L_y, $0 < y < \infty$, respectively. Evidently, for any $E \in \mathcal{M}(\Omega)$ we have

$$L_y(E) \leq U_y(E) \quad \text{whenever } 0 < y < \infty$$

with equality for $\mathcal{L}^1$ almost all $y \in (0,\infty)$.

Moreover, if F is local, one finds that by modifying the proof of Proposition 1.6 in a straightforward way that this proposition holds with u and U_y, $0 < y < \infty$, replaced by l and L_y, $0 < y < \infty$, respectively, except that (1.4) must be replaced by

$$\lim_{z \uparrow y} L_z(E) = L_y(E) \quad \text{whenever } E \in \mathcal{F}(\Omega). \tag{1.5}$$

We will show at the end of section 6 that for all but countably many $y \in (0,\infty)$

$$L_y(E) = U_y(E) \quad \text{whenever } E \in \mathcal{M}(\Omega). \tag{1.6}$$

See section 1.9 for a natural example where $L_y \neq U_y$ for some $y \in (0,\infty)$.

1.6.1. When F is local and convex. Things get a lot more interesting when F is local and convex. An important result, which will be proved in section 6, is the following.

THEOREM 1.3. *Suppose $F : \mathcal{F}(\Omega) \to \mathbb{R}$, F is local and convex, and $f \in \mathbf{m}_\epsilon^{loc}(F)$. Then*

$$\{f \geq y\} \in \mathbf{n}_\epsilon^{loc}(L_y) \quad \textit{and} \quad \{f > y\} \in \mathbf{n}_\epsilon^{loc}(U_y) \quad \textit{whenever } 0 < y < \infty.$$

When I wrote the initial version of this paper I thought this theorem was completely new. I was wrong. As a referee has pointed out, essentially the following result appears as Proposition 2.2 of [CA2].

Suppose $s \in \mathcal{F}(\Omega)$,

$$F(g) = \frac{1}{2}\int_\Omega |g - s|^2 \, d\mathcal{L}^2 \quad \textit{for } g \in \mathbf{L}_2(\Omega),$$

$0 < \epsilon < \infty$, and $f \in \mathbf{L}_2(\Omega)$ is such that

$$\epsilon \mathbf{TV}(f, \Omega) + F(f) \leq \epsilon \mathbf{TV}(g, \Omega) + F(g) \quad \textit{for } g \in \mathbf{L}_2(\Omega).$$

Then $f(x) \geq 0$ for $\mathcal{L}^n$ almost all $x \in \Omega$ and, whenever $0 < y < \infty$,

$$\epsilon \mathbf{TV}(\{f > y\}) + U_y(\{f > y\}) \leq \epsilon \mathbf{TV}(E) + U_y(E) \quad \textit{for } E \in \mathcal{M}(\Omega),$$

where $U_y(E) = \int_E y - s \, d\mathcal{L}^n$ for $E \in \mathcal{M}(\Omega)$.

Note that f above is a *global* minimizer of $\mathbf{L}_2(\Omega) \ni g \mapsto \epsilon \mathbf{TV}(g, \Omega) + F(g)$. In fact, the method used to prove this result in [CA2] can be extended to a very general class of local and convex F's but still for global minimizers. For example, I do not see how to apply this method to the case when Ω has Lipschitz boundary and one minimizes in the class of f's with a given trace on the boundary of Ω, a situation in which Theorem 1.3 clearly applies.

The following theorem, which will be proved in section 6, is more than a converse of the preceding theorem. This result is of particular interest when $\gamma(y) = |y|$ for $y \in \mathbb{R}$ in section 1.8; it is the essential ingredient in the proof of Theorem 1.7.

THEOREM 1.4. *Suppose $F : \mathcal{F}(\Omega) \to \mathbb{R}$, F is local and convex, G is a $\mathcal{L}^n \times \mathcal{L}^1$ measurable subset of $\Omega \times (0, \infty)$ such that*

(i) $(\mathcal{L}^n \times \mathcal{L}^1)(G) < \infty$*;*
(ii) $\{y \in (0, \infty) : \{(x, y) \in G\} \neq \emptyset\}$ *is bounded;*
(iii) $\{x : (x, y) \in G\} \in \mathbf{n}_\epsilon^{loc}(U_y)$ *for $\mathcal{L}^1$ almost all $y \in (0, \infty)$,*

and $f : \Omega \to [0, \infty)$ is such that

$$f(x) = \mathcal{L}^1\{y : (x, y) \in G\} \quad \textit{for } x \in \Omega.$$

Then $f \in \mathbf{m}_\epsilon^{loc}(F)$.

It turns out that a set G as in the previous theorem is essentially unique, provided F is strictly convex. Simple examples in [CE] show this is not the case if F is merely convex. Bear in mind that f below is essentially unique because F is strictly convex.

THEOREM 1.5. *Suppose F, G, and f are as in the preceding theorem and F is strictly convex. Then*

$$(\mathcal{L}^n \times \mathcal{L}^1)((G \sim \{(x, y) : f(x) > y > 0\}) \cup (\{(x, y) : f(x) > y > 0\} \sim G)) = 0.$$

See section 9.3 for the proof.

1.7. Results on curvature. A good deal of the following theorem, which will be proved in section 7, is well known. If one assumes that M below is of class C^2, the formula for H in (1.7) may be derived by a straightforward variational argument which appears in [M]; in our case, in the light of the regularity theorem for $\mathcal{C}_\lambda(\Omega)$ we know only that M is of class $C^{1,\mu}$, $0 < \mu < 1$, so one must proceed a bit more carefully. We represent M locally as a graph of a function which satisfies an elliptic equation and appeal to higher regularity results for such equations as appear, for example, in [GT]. One may then obtain the second variation formula (1.8) which, obviously, is a *global* constraint on a member of $\mathbf{n}_\epsilon^{loc}(Z)$ to which it applies. I believe (1.8) is new; it will be used in section 11 and [AW2] when we construct minimizers.

See section 4 for the definitions of mean curvature and second fundamental form which we use.

THEOREM 1.6. *Suppose*

(i) $\zeta \in \mathbf{L}_\infty(\Omega)$ *and* $Z(E) = \int_E \zeta \, d\mathcal{L}^n$ *whenever* $E \in \mathcal{M}(\Omega)$*;*

(ii) U *is an open subset of* Ω*,* j *is a nonnegative integer,* $0 < \mu < 1$*, and* $\zeta | U$ *is of class* $C^{j,\mu}$*;*

(iii) $0 < \epsilon < \infty$ *and* $D \in \mathbf{n}_\epsilon^{loc}(Z)$*;*

(iv) Γ *is the intersection of* U *with the interior of the support of the generalized function corresponding to the indicator function of* D *and* M *is the boundary of* Γ *relative to* U*.*

Then $\Sigma_U(D, \Gamma) = 0$*,* M *is an embedded hypersurface of class* $C^{j+2,\mu}$*, and*

$$H(x) = -\frac{1}{\epsilon}\zeta(x)\nu(x) \quad \text{for } x \in M, \tag{1.7}$$

where H *is the mean curvature vector of* M *and* ν *is the unit normal along* M *which points out of* Γ*.*

Moreover, if ζ *is of class* C^1 *on* U *and* Q *is the square of the length of the second fundamental form of* M*, then*

$$\int_M \epsilon \left(|\nabla_M \phi(x)|^2 - \phi(x)^2 Q(x) \right) - \phi(x)^2 \nabla\zeta(x) \bullet \nu(x) \, d\mathcal{H}^{n-1}x \geq 0 \tag{1.8}$$

for any smooth compactly supported function ϕ *on* U*; here, for each* $x \in M$*,* $\nabla_M \phi(x)$ *is the orthogonal projection of* $\nabla\phi(x)$ *on* $\mathbf{Tan}(M, x)$*.*

In case $n = 2$ *we may take* $\mu = 1$*.*

This theorem will apply in the context of denoising if s as in section 1.8 is sufficiently regular in U.

1.8. Denoising. Suppose

(i) $s \in \mathcal{F}(\Omega)$;

(ii) $\gamma : \mathbb{R} \to [0, \infty)$, γ is locally Lipschitzian, γ is decreasing on $(-\infty, 0)$, $\gamma(0) = 0$, and γ is increasing on $(0, \infty)$;

(iii) $F(f) = \int_\Omega \gamma(f(x) - s(x)) \, d\mathcal{L}^n x$ for $f \in \mathcal{F}(\Omega)$.

Here s could be a grayscale representation of a degraded image which we wish to denoise. In the context of denoising F would be called a *fidelity*; it is a measure of how much f differs from s. If $0 < \epsilon < \infty$, the members of $\mathbf{m}_\epsilon^{loc}(F)$ would be called *total variation regularizations of* s *(with respect to the fidelity* F *and smoothing parameter* ϵ*).*

In the literature F_ϵ is often replaced by F_ϵ/ϵ and $\lambda = 1/\epsilon$ is thought of as a Lagrange multiplier.

For a very informative discussion of the use of total variation regularizations in the field of image processing, see the introduction of [CE]. We will not discuss image processing any further except to note that the notion of total variation regularization in image processing is useful for other purposes besides denoising.

Evidently, F is admissible, so Proposition 1.1 holds and the results of section 1.5 apply. It is also evident that F is local.

Let us now assume γ is convex. It follows that F is convex. Set

$$\alpha(y) = \liminf_{z\to y} \frac{\gamma(z)-\gamma(y)}{z-y} \quad \text{and} \quad \beta(y) = \limsup_{z\to y} \frac{\gamma(z)-\gamma(y)}{z-y} \qquad \text{for } y \in \mathbb{R}$$

and let L_y and U_y, $0 < y < \infty$, be as in Remark 1.3 and Proposition 1.5, respectively. It is a simple matter to verify that if $0 < y < \infty$ and $E \in \mathcal{M}(\Omega)$, then

$$L_y(E) = \int_E \alpha(y - s(x))\, d\mathcal{L}^n x \quad \text{and} \quad U_y(E) = \int_E \beta(y - s(x))\, d\mathcal{L}^n x. \tag{1.9}$$

In view of Theorem 1.3 the results of section 1.7 apply when α or β and s are sufficiently regular.

Of particular interest is when $1 \le p < \infty$ and

$$\gamma(y) = \frac{1}{p}|y|^p \quad \text{whenever } y \in \mathbb{R}. \tag{1.10}$$

Rudin, Osher, and Fatemi [ROF] studied the case $p = 2$ and Chan and Esedoḡlu [CE] studied the case $p = 1$. The results of section 1.7 will allow us to construct a number of interesting examples of minimizers in [AW2], a sequel to this paper; we believe these examples provide insights into the nature of total variation regularization. At the end of this paper we will determine $\mathbf{m}_\epsilon^{loc}(F)$ when $\Omega = \mathbb{R}^2$, s is the indicator function of a square, and γ is as in (1.10). Note that F is strictly convex if $p > 1$ and merely convex if $p = 1$.

Suppose $K \in \mathbf{L}_1(\mathbb{R}^n)$. Let

$$F(f) = \int_{\mathbb{R}^n} \gamma(K * f(x) - s(x))\, d\mathcal{L}^n x \quad \text{for } f \in \mathcal{F}(\mathbb{R}^n).$$

It is easy to see that F is admissible but not local except in degenerate situations. Nonetheless, the results of section 1.5 apply.

1.9. Some results on the Chan–Esedoḡlu functional. Suppose s, γ, F are as in section 1.8 with $\gamma(y) = |y|$ for $y \in \mathbb{R}$. Whenever $0 < y < \infty$ and $E \in \mathcal{M}(\Omega)$ we use (1.9) to obtain

$$L_y(E) = \mathcal{L}^n(E \cap \{y > s\}) - \mathcal{L}^n(E \cap \{y \le s\}) = \widehat{N_{\{y\le s\}}}(E),$$
$$U_y(E) = \mathcal{L}^n(E \cap \{y \ge s\}) - \mathcal{L}^n(E \cap \{y < s\}) = \widehat{N_{\{y< s\}}}(E),$$

where for each $S \in \mathcal{M}(\Omega)$ we have set

$$N_S(E) = \Sigma_\Omega(E, S) \quad \text{for } E \in \mathcal{M}(\Omega).$$

We use Theorem 1.4 to obtain interesting results about N_S, $S \in \mathcal{M}(\Omega)$, one of which is as follows; it was suggested by a similar result in a different context in [CA1] and will be used in [AW2] in determining $\mathbf{n}_\epsilon^{loc}(N_S)$, $0 < \epsilon < \infty$, for certain S, which in view of the above formulae for L_y and U_y, $0 < y < \infty$, and Theorem 1.3 will allow us to determine $\mathbf{m}_\epsilon^{loc}(F)$, $0 < \epsilon < \infty$, where $F(f) = \int_{\mathbb{R}^2} |f - 1_S| \, d\mathcal{L}^2$ for $f \in \mathcal{F}(\mathbb{R}^2)$.

THEOREM 1.7. *Suppose $S \in \mathcal{M}(\Omega)$, $0 < \epsilon < \infty$, and $\mathcal{A}$ is a nonempty subfamily of $\mathbf{n}_\epsilon^{loc}(N_S)$. Then*

$$\cap\mathcal{A} \in \mathbf{n}_\epsilon^{loc}(N_S) \quad \textit{and, provided } \mathcal{L}^n(\cup\mathcal{A}) < \infty, \quad \cup\,\mathcal{A} \in \mathbf{n}_\epsilon^{loc}(N_S).$$

2. Geometric measure theoretic background. We find the mathematical infrastructure of normal and integral currents to be convenient in carrying out this work. For that reason we will adopt, for the most part, the notation and terminology of [FE]; note the extensive glossary, list of notation, and index starting on page 669 of that book. We avoided using that notation and terminology in the introduction in order to make it more accessible to readers not familiar with [FE].

2.1. More notations and conventions. Suppose μ *measures* Ω, which is to say μ maps the power set of Ω countably subadditively into $[0, \infty]$; whenever $A \subset \Omega$ we let

$$\mu \llcorner A(B) = \mu(A \cap B) \quad \text{whenever } B \subset \Omega$$

and note that $\mu \llcorner A$ measures Ω.

Whenever f is a function mapping a subset of a normed vector space into another normed vector space, a is an interior point of the domain of f, and f is Fréchet differentiable at a, we let

$$\partial f(a)$$

be the Fréchet differential of f at a.

If V is a vector space, $v \in V$, and ψ belongs to the dual space of V, we frequently write

$$\langle v, \psi \rangle \quad \text{instead of} \quad \psi(v).$$

We use **spt** as an abbreviation for "support."

2.2. Spaces of smooth functions; currents. Whenever Y is a Banach space we let

$$\mathcal{E}(\Omega, Y) \quad \text{and} \quad \mathcal{D}(\Omega, Y)$$

be the space of smooth Y valued functions on Ω and the space of compactly supported members of $\mathcal{E}(\Omega, Y)$, respectively, with the strong topologies as described in [FE, sect. 4.1.1]. Thus $\mathcal{X}(\Omega) = \mathcal{D}(\Omega, \mathbb{R}^n)$.

We let

$$\mathcal{E}(\Omega) \quad \text{and} \quad \mathcal{D}(\Omega)$$

equal $\mathcal{E}(\Omega, \mathbf{R})$ and $\mathcal{D}(\Omega, \mathbf{R})$, respectively. For each nonnegative integer m we let

$$\mathcal{E}^m(\Omega) \quad \text{and} \quad \mathcal{D}^m(\Omega)$$

equal $\mathcal{E}(\Omega, Y)$ and $\mathcal{D}(\Omega, Y)$, respectively, with $Y = \bigwedge^m \mathbb{R}^n$. Thus $\mathcal{E}^m(\Omega)$ is the space of smooth differential m-forms on Ω, and $\mathcal{D}^m(\Omega)$ is the space of those members of $\mathcal{E}^m(\Omega)$ with compact support. We let

$$\mathcal{E}_m(\Omega) \quad \text{and} \quad \mathcal{D}_m(\Omega)$$

be the duals of $\mathcal{E}^m(\Omega)$ and $\mathcal{D}^m(\Omega)$, respectively. Thus $\mathcal{D}_m(\Omega)$ is the space of m dimensional *currents* on Ω, and $\mathcal{E}_m(\Omega)$ is the space of m-dimensional currents with compact support on Ω. We define the *boundary operator*

$$\partial : \mathcal{D}_m(\Omega) \to \mathcal{D}_{m-1}(\Omega)$$

by setting $\partial T(\omega) = T(d\omega)$ whenever $T \in \mathcal{D}_m(\Omega)$ and $\omega \in \mathcal{D}_{m-1}(\Omega)$; here d is exterior differentiation.

Suppose $T \in \mathcal{D}_m(\Omega)$. As in [FE, sect. 4.1.5] we let

$$||T||,$$

the *total variation measure of* T, be the largest Borel regular measure on Ω such that

$$||T||(G) = \sup\{|T(\omega)| : \omega \in \mathcal{D}^m(\Omega),\ ||\omega|| \le 1 \text{ and } \mathbf{spt}\,\omega \subset G\}$$

for each open subset G of Ω; here $||\cdot||$ is the *comass* which in case $m \in \{0, 1, n-1, n\}$ is the Euclidean norm; these are the only cases we will encounter in this paper. It follows immediately from this definition that

$$||T||(G) \le \liminf_{\nu\to\infty} ||S_\nu||(G) \quad \text{for any open subset } G \text{ of } \Omega \tag{2.1}$$

whenever S is a sequence in $\mathcal{D}_m(\Omega)$ such that $S_\nu(\omega) \to T(\omega)$ as $\nu \to \infty$ whenever $\omega \in \mathcal{D}^m(\Omega)$. We let

$$\mathbf{M}(T) = ||T||(\Omega)$$

and call this nonnegative extended real number the *mass* of T. We say T is *representable by integration* if $||T||$ is a Radon measure which is equivalent to the statement that $||T||(K) < \infty$ whenever K is a compact subset of Ω. If this is the case and $\overrightarrow{T}$ is the $||T||$ measurable function with values in $\{\xi \in \bigwedge_m \mathbb{R}^n : ||\xi|| = 1\}$ defined in [FE, sect. 4.1.7], there is a unique extension of T to the $||T||$ summable functions on Ω with values in $\bigwedge^m \mathbb{R}^n$, which we continue to denote by T, such that

$$T(\omega) = \int \langle \overrightarrow{T}(x), \omega(x) \rangle \, d||T||x$$

whenever ω is a $||T||$ summable function on Ω with values in $\bigwedge^m \mathbb{R}^n$. If $T \in \mathcal{D}_m(\Omega)$ is representable by integration, l is a nonnegative integer not exceeding m, and η is a bounded Borel function on Ω with values in $\bigwedge^l \mathbb{R}^n$, then we let

$$T \llcorner \eta \in \mathcal{D}_{m-l}(\Omega)$$

be such that

$$T \llcorner \eta(\omega) = \int \langle \overrightarrow{T}(x), (\eta \wedge \omega)(x) \rangle \, d||T||x \quad \text{for } \omega \in \mathcal{D}^{m-l}(\Omega).$$

2.3. The current corresponding to a locally summable function. We let

$$\mathbf{e}_1, \ldots, \mathbf{e}_n \quad \text{and} \quad \mathbf{e}^1, \ldots, \mathbf{e}^n$$

be the standard basis vectors and covectors for $\mathbb{R}^n$ and its dual space, respectively. We let

$$\mathbf{E}^n = \mathbf{e}^1 \wedge \cdots \wedge \mathbf{e}^n \in \bigwedge^n \mathbb{R}^n$$

be the standard orientation on $\mathbb{R}^n$.

We let

$$\mathbf{V}^n \in \mathcal{D}^n(\Omega) \tag{2.2}$$

be such that $\mathbf{V}^n(x) = \mathbf{E}^n$ for $x \in \Omega$.

DEFINITION 2.1. *Whenever* $f \in \mathbf{L}_1^{loc}(\Omega)$ *we define*

$$[f] \in \mathcal{D}_n(\Omega)$$

by setting

$$[f](\phi \mathbf{V}^n) = \int_\Omega \phi f \, d\mathcal{L}^n \quad \textit{whenever } \phi \in \mathcal{D}(\Omega).$$

Suppose $f \in \mathbf{L}_1^{loc}(\Omega)$. For any $X \in \mathcal{X}(\Omega)$ we have $d(X \lrcorner \mathbf{V}^n) = (-1)^{n-1}(\operatorname{div} X)\mathbf{V}^n$ so that

$$\partial[f](X \lrcorner \mathbf{V}^n) = (-1)^{n-1} \int f \operatorname{div} X \, d\mathcal{L}^n; \tag{2.3}$$

here $\lrcorner$ is as in [FE, sect. 1.5.1]. It follows that

$$\mathbf{TV}(f, B) = ||\partial[f]||(B) \quad \text{whenever } B \text{ is a Borel subset of } \Omega. \tag{2.4}$$

2.4. Mapping currents. Whenever $T \in \mathcal{D}_m(\Omega)$ and F is a smooth map from Ω to the open subset Γ of some Euclidean space whose restriction to the support of T is proper, we let

$$F_\# T \in \mathcal{D}_m(\Gamma)$$

be such that $F_\# T(\omega) = T(F^\# \omega)$ for any $\omega \in \mathcal{D}^m(\Gamma)$; here the *pullback* $F^\#$ is as in [FE, sect. 4.1.6]. If F carries Ω diffeomorphically onto Γ, T is representable by integration, and $\overrightarrow{T}(x)$ is decomposable for $||T||$ almost all $x \in \Omega$, then we have

$$\int b(y) \, d||F_\# T||y = \int b(F(x)) \left| \bigwedge\nolimits_m \partial F(x)(\overrightarrow{T}(x)) \right| d||T||x \tag{2.5}$$

for nonnegative Borel function b on Γ. By a simple approximation argument one need only assume that F is of class C^1 if T is representable by integration.

2.5. A mapping formula. Suppose Γ is an open subset of $\mathbb{R}^n$; $f \in \mathbf{L}_1^{loc}(\Omega)$; $F : \Omega \to \Gamma$ is locally Lipschitzian; the restriction of F to the support of $[f]$ is proper; A is the set of $y \in \Gamma$ such that $F^{-1}[\{y\}]$ is finite and such that if $F(x) = y$, then F is differentiable at x; and $g : \Gamma \to \mathbb{R}$ is such that

$$g(y) = \begin{cases} \sum_{x \in F^{-1}[\{y\}]} f(x) \operatorname{sgn} \det \partial F(x) & \text{if } y \in A, \\ 0 & \text{else.} \end{cases}$$

Then $g \in \mathbf{L}_1^{loc}(\Gamma)$ and

$$F_\#[f] = [g]. \tag{2.6}$$

In particular, if F is univalent and $\det \partial F(x) > 0$ for $\mathcal{L}^n$ almost all $x \in \Omega$, then

$$F_\#[f] = [f \circ F^{-1}].$$

See [FE, sect. 4.1.25] for the proof.

2.6. Slicing. Suppose m, l are positive integers, $m \geq l$, $T \in \mathcal{D}^m(\Omega)$, T is locally flat as defined in [FE, sect. 4.1.12], and $f : \Omega \to \mathbb{R}^l$ is locally Lipschitzian. Note that if both T and ∂T are representable by integration, then T is locally flat; this will always be the case when we apply slicing in this paper. For $y \in \mathbb{R}^l$ we follow [FE, sect. 4.3.1] and define

$$\langle T, f, y \rangle$$

the *slice of* T *in* $f^{-1}[\{y\}]$ to be that member of $\mathcal{D}_{m-l}(\Omega)$ which, if it exists, satisfies

$$\langle T, f, y\rangle(\psi) = \lim_{r \downarrow 0} \frac{T \llcorner [f^\#(\mathbf{B}^l(y,r) \wedge \mathbf{V}^l)](\psi)}{\mathcal{L}^l(\mathbf{B}^l(y,r))} \quad \text{whenever } \psi \in \mathcal{D}^{m-l}(\Omega),$$

where $T \llcorner [f^\#(\mathbf{B}^l(y,r) \wedge \mathbf{V}^l)]$ is defined as in [FE, sect. 4.3.1]. Then, by [FE, sect. 4.3.1], the slice $\langle T, f, y\rangle$ exists for $\mathcal{L}^l$ almost all y and satisfies

$$\mathbf{spt}\, \langle T, f, y\rangle \subset f^{-1}[\{y\}] \quad \text{and} \quad \partial\langle T, f, y\rangle = (-1)^l \langle \partial T, f, y\rangle. \tag{2.7}$$

Moreover, we have from [FE, sect. 4.3.2] that

$$\int \Phi(y)\langle T, f, y\rangle(\psi)\, d\mathcal{L}^l y = [T \llcorner f^\#(\Phi \wedge \mathbf{V}^l)](\psi) \tag{2.8}$$

whenever Φ is a bounded Borel function on $\mathbb{R}^l$ and $\psi \in \mathcal{D}^{m-l}(\Omega)$ and that

$$\int \left(\int b ||\langle T, f, y\rangle|| \right) d\mathcal{L}^l y = \int b\, d||T \llcorner f^\# \mathbf{V}^l]|| \tag{2.9}$$

whenever b is a nonnegative Borel function on Ω.

PROPOSITION 2.1. *Suppose K is a compact subset of Ω, $u(x) = \mathbf{dist}\,(x, K)$ for $x \in \Omega$, R is the supremum of the set of $r \in (0, \infty)$ such that $\{u \leq r\} \subset \Omega$, $f, g \in \mathbf{BV}^{loc}(\Omega)$, and*

$$h_r = g 1_{\{u \leq r\}} + f 1_{\{u > r\}} \quad \textit{for each } r \in (0, R).$$

Then $h_r \in \mathbf{BV}^{loc}(\Omega)$ for $\mathcal{L}^1$ almost all $r \in (0, R)$, and whenever $0 < r < s < R$ we have

$$\int_r^s ||\partial[h_\rho]||(\{u \leq \rho\})\, d\mathcal{L}^1\rho \leq \int_{\{r < u < s\}} |f - g|\, d\mathcal{L}^n + \int_r^s ||\partial[g]||(\{u \leq \rho\})\, d\mathcal{L}^1 \rho. \tag{2.10}$$

Proof. From [FE, sect. 4.2.1] and [FE, sect. 4.3.4] we find that

$$\partial[h_\rho] = \langle [g] - [f], u, \rho\rangle + (\partial[g]) \llcorner \{u \leq \rho\} + (\partial[f]) \llcorner \{u > \rho\}$$

for $\mathcal{L}^1$ almost all $\rho \in (0, R)$. Now multiply by $1_{\{u \leq \rho\}}$, integrate from r to s, and invoke (2.9). ◻

2.7. Densities and density ratios. Suppose μ measures Ω, m is a nonnegative integer, and $\alpha(m) = \mathcal{L}^m(\mathbf{U}^m(0,1))$. For each $a \in \Omega$ we set

$$\Theta^m(\mu, a, r) = \frac{\mu(\mathbf{B}(a,r))}{\alpha(m) r^m} \quad \text{whenever } 0 < r < \mathbf{dist}\,(a, \mathbb{R}^n \sim \Omega)$$

and

$$\Theta^m(\mu, a) = \lim_{r \to 0} \Theta^m(\mu, a, r),$$

provided this limit exists.

2.8. Sets of finite perimeter. Suppose E is a Lebesgue measurable subset of Ω. Proceeding as in [FE, sect. 4.5.5], we say $u \in \mathbb{R}^n$ is an *exterior normal to E at $b \in \Omega$* if $|u| = 1$ and

$$\Theta^n(\mathcal{L}^n \llcorner \{x \in E : (x-b) \bullet u > 0\} \cup \{x \in \Omega \sim E : (x-b) \bullet u < 0\}, b) = 0.$$

We let

$$\mathbf{n}_E$$

be the set of $(b, u) \in \Omega \times \mathbb{R}^n$ such that *either* u is an exterior normal to E at b *or* $u = 0$ and there is no exterior normal to E at b; note that $\mathbf{n}_E$ is a function with domain Ω. We let

$$\mathbf{b}(E),$$

the *reduced boundary of E*, be equal to the set of points $b \in \Omega$ such that there is an exterior normal to E at b.

THEOREM 2.1 (see [FE, sect. 4.5.6]). *Suppose E is a subset of Ω with locally finite perimeter. The following statements hold:*

(i) $\mathbf{b}(E)$ *is a Borel set which is countably $(\mathcal{H}^{n-1}, n-1)$ rectifiable.*

(ii) $||\partial[E]|| = \mathcal{H}^{n-1} \llcorner \mathbf{b}(E)$.

(iii) *For $\mathcal{H}^{n-1}$ almost all $b \in \mathbf{b}(E)$ we have*

$$*\mathbf{n}_E(b) = \overrightarrow{\partial[E]}(b) \quad \textit{and} \quad \Theta^{n-1}(||\partial[E]||, b) = 1;$$

here $*$ is the Hodge star operator *as defined in* [FE, sect. 1.7.8].

(iv) *For $\mathcal{H}^{n-1}$ almost all $b \in \Omega \sim \mathbf{b}(E)$, $\Theta^{n-1}(||\partial[E]||, b) = 0$ and*

$$\textit{either } \Theta^n(\mathcal{L}^n \llcorner E, b) = 0 \quad \textit{or} \quad \Theta^n(\mathcal{L}^n \llcorner (\Omega \sim E), b) = 0.$$

It follows that if E is a subset of Ω with locally finite perimeter, then

$$\partial[E](X \lrcorner \mathbf{V}^n) = (-1)^{n-1} \int X \bullet \mathbf{n}_E \, d||\partial[E]|| \quad \text{whenever } X \in \mathcal{X}(\Omega). \tag{2.11}$$

PROPOSITION 2.2. *Suppose E is a subset of Ω with finite perimeter and C is a closed convex subset of $\mathbb{R}^n$. Then*

$$\mathbf{M}(\partial[C \cap E]) \le \mathbf{M}(\partial[E]). \tag{2.12}$$

Proof. Let $\rho : \mathbb{R}^n \to C$ be such that $|x - \rho(x)| = \mathbf{dist}\,(x, C)$ for $x \in \mathbb{R}^n$. In case $\mathbf{spt}\,[E]$ is compact we infer from (2.6) that $[C \cap E] = \rho_\#[E]$ so that, as $\mathbf{Lip}\,\rho \le 1$, (2.12) holds. In case $\mathbf{spt}\,[E]$ is not compact we let $E_r = E \cap \mathbf{U}^n(0, r)$, $0 < r < \infty$, and apply the result just obtained together with (2.10) and (2.1). ∎

2.9. Basic facts about functions of bounded variation. Proofs of the following formulae, which are absolutely fundamental for this work, may be found in [FE, sect. 4.5.9, eq. (13)]. Suppose $f \in \mathbf{BV}^{loc}(\Omega)$; then $\mathbb{R} \ni y \mapsto \partial[\{f \ge y\}](\omega)$ is $\mathcal{L}^1$ summable and

$$\partial[f](\omega) = \int \partial[\{f > y\}](\omega) \, d\mathcal{L}^1 y \quad \text{whenever } \omega \in \mathcal{D}^{n-1}(\Omega); \tag{2.13}$$

moreover, if B is a Borel subset of Ω, then $\mathbb{R} \ni y \mapsto ||\partial[\{f > y\}]||(B)$ is $\mathcal{L}^1$ measurable and

$$||\partial[f]||(B) = \int ||\partial[\{f > y\}]||(B)\, d\mathcal{L}^1 y. \tag{2.14}$$

The following well-known theorem follows from (2.1) and the discussion in [FE, sect. 4.5.7] concerning locally flat currents of dimension n in Ω.

THEOREM 2.2 (compactness theorem). *Suppose C is a sequence of nonnegative real numbers and K is a sequence of compact subsets of Ω such that $\cup_{\nu=1}^{\infty} K_\nu = \Omega$. Then*

$$\bigcap_{\nu=1}^{\infty} \left\{ f \in \mathbf{BV}^{loc}(\Omega) : \int_{K_\nu} |f|\, d\mathcal{L}^n + ||\partial[f]||(K_\nu) \leq C_\nu \right\}$$

is a compact subset of $\mathbf{L}_1^{loc}(\Omega)$.

PROPOSITION 2.3. *Suppose $f \in \mathbf{BV}^{loc}(\Omega)$ and $y \in \mathbf{R}$. Then $f \wedge y, f \vee y \in \mathbf{BV}^{loc}(\Omega)$ and*

$$||\partial[f \wedge y]|| + ||\partial[f \vee y]|| = ||\partial[f]||. \tag{2.15}$$

Proof. Since $f + y = f \wedge y + f \vee y$ it is trivial that the right-hand side of (2.15) does not exceed the left-hand side of (2.15). Using (2.13) one readily shows that

$$[f \wedge y](\omega) = \int_{-\infty}^{y} [\{f \geq z\}](\omega)\, d\mathcal{L}^1 z \quad \text{and} \quad [f \vee y](\omega) = \int_{y}^{\infty} [\{f > y\}](\omega)\, d\mathcal{L}^1 y$$

whenever $\omega \in \mathcal{D}^n(\Omega)$. Applying ∂ one infers

$$||\partial[f \wedge y]|| \leq \int_{-\infty}^{y} ||\partial[\{f > y\}]||\, d\mathcal{L}^1 y \quad \text{and} \quad ||\partial[f \vee y]|| \leq \int_{y}^{\infty} ||\partial[\{f > y\}]||\, d\mathcal{L}^1 y.$$

By (2.14) the sum of the right-hand sides of these inequalities is $||\partial[f]||$. Thus the left-hand side of (2.15) does not exceed the right-hand side. $\square$

2.10. The "layer cake" formula. Chan and Esedoḡlu in [CE] call the following elementary formula the "layer cake" formula; it is indispensable in this work.

PROPOSITION 2.4. *Suppose f, g are real valued Lebesgue measurable functions on Ω. Then*

$$\int_\Omega |f - g|\, d\mathcal{L}^n = \int_{-\infty}^{\infty} \Sigma_\Omega(\{f > y\}, \{g > y\})\, d\mathcal{L}^1 y. \tag{2.16}$$

Proof. Apply Tonelli's theorem to calculate the $\mathcal{L}^n \times \mathcal{L}^1$ measure of $\{(x, y) \in \Omega \times \mathbb{R} : g(x) < y \leq f(x)\}$ and $\{(x, y) \in \Omega \times \mathbb{R} : f(x) < y \leq g(x)\}$ and add the results. $\square$

3. Deformations and variations. We suppose throughout this section that

(i) $X : \Omega \to \mathbb{R}^n$ is continuously differentiable and $K = \mathbf{spt}\, X$ is compact;

(ii) I is an open interval containing 0 such that if $t \in I$ and

$$h_t(x) = x + tX(x) \quad \text{for } x \in \Omega,$$

then h_t carries Ω diffeomorphically (in the C^1 sense) onto itself;

(iii) D is a Lebesgue measurable subset of Ω with locally finite perimeter and

$$E_t = \{h_t(x) : x \in D\} \quad \text{for } t \in I;$$

(iv) for $x \in \mathbf{b}(D)$

$$P(x) \text{ is orthogonal projection of } \mathbb{R}^n \text{ onto } \{v \in \mathbb{R}^n : v \bullet \mathbf{n}_D(x) = 0\},$$

$$l_1(x) = P(x) \circ \partial X(x) \circ P(x) \quad \text{and} \quad l_2(x) = P(x)^\perp \circ \partial X(x) \circ P(x).$$

Note that given X as in (i) there is always I as in (ii).

3.1. Some useful variational formulae.

PROPOSITION 3.1. *Suppose*

$$A(t) = ||\partial[E_t]||(K) \quad \textit{for each } t \in I.$$

Then A is smooth, and

$$\dot{A}(0) = \int a_1 \, d||\partial[D]|| \quad \textit{and} \quad \ddot{A}(0) = \int a_2 \, d||\partial[D]||,$$

where for $x \in \mathbf{b}(D)$ we have set

$$a_1(x) = \operatorname{trace} l_1(x) \quad \textit{and} \quad a_2(x) = a_1(x)^2 + \operatorname{trace}(l_2(x)^* \circ l_2(x) - l_1(x) \circ l_1(x)).$$

Proof. It follows from (2.6) that $[E_t] = h_{t\#}[D]$ and therefore $\partial[E_t] = h_{t\#}\partial[D]$ for any $t \in I$. Now recall from Theorem 2.1(iii) that $*\mathbf{n}_D(x) = \overrightarrow{\partial[D]}(x)$ for $||\partial[D]||$ almost all x, differentiate under the integral sign in (2.5), and use the formulae

$$\left(\frac{d}{dt}\right)^j \bigwedge\nolimits_{n-1} \partial h_t(x)(\overrightarrow{\partial[D]}(x))\Big|_{t=0} = a_j(x), \quad j = 1, 2, \quad x \in \mathbf{b}(D),$$

proofs of which may be found in [FE, sect. 5.1.8]. $\square$

Since $[E_t] - [D]$ is compactly supported, $([E_t] - [D])(\phi \mathbf{V}^n)$ is well defined in the following proposition.

PROPOSITION 3.2. *For any $\phi \in \mathcal{E}(\Omega)$ we have*

$$([E_t] - [D])(\phi \mathbf{V}^n) = \int_0^t \left(\int \phi(h_\tau(x)) \, W_\tau(x) \, d||\partial[D]||x \right) d\mathcal{L}^1\tau,$$

where, for each $t \in I$, we have set

$$W_t(x) = \left\langle X(x) \wedge \bigwedge\nolimits_{n-1} \partial h_t(x)(*\mathbf{n}_D(x)), \mathbf{E}^n \right\rangle \quad \textit{for } x \in \mathbf{b}(D).$$

Proof. For each $t \in I$ let $J_t = [0, t] \in \mathcal{D}_1(\mathbb{R})$ as in [FE, sect. 4.1.8]. From [FE, sect. 4.1.8] we have $||J_t \times \partial[D]|| = ||J_t|| \times ||\partial[D]||$ for each $t \in I$. From [FE, sect. 4.1.8] and Theorem 2.1(iii) we have

$$\overrightarrow{J_t \times \partial[D]}(\tau, x) = (1, 0) \wedge \overrightarrow{\partial[D]}(x) = (1, 0) \wedge *\mathbf{n}_D(x) \quad \text{for } (\tau, x) \in (0, t) \times \mathbf{b}(D).$$

Suppose $t \in I$. We obtain

$$[E_t] - [D] = h_{t\#}[D] - [D] = h_\#(J_t \times \partial[D])$$

from the homotopy formula of [FE, sect. 4.1.9]; the formula to be proved now follows from (2.5). ∎

PROPOSITION 3.3. *Suppose $\mathcal{L}^n(D) < \infty$, $\zeta \in \mathbf{L}_\infty(\Omega)$, and*

$$B(t) = \int_{E_t} \zeta \, d\mathcal{L}^n \quad \textit{for } t \in I.$$

If ζ is continuous, then B is continuously differentiable and

$$\dot{B}(0) = \int \zeta(X \bullet \mathbf{n}_D)\, d\|\partial[D]\|. \tag{3.1}$$

If ζ is continuously differentiable, then B is twice continuously differentiable and

$$\ddot{B}(0) = \int \big(\zeta Y + (\nabla\zeta \bullet X)X\big) \bullet \mathbf{n}_D \, d\|\partial[D]\|, \tag{3.2}$$

where for $x \in \mathbf{b}(D)$ we have set

$$Y(x) = (\operatorname{trace} l_1(x))X(x) - l_2(x)(X(x)).$$

Proof. Using straightforward approximations if necessary, we may assume that ζ is smooth. For each $t \in I$ and $x \in \mathbf{b}(D)$ let

$$\xi_t(x) = \bigwedge\nolimits_{n-1} \partial h_t(x)(*\mathbf{n}_D(x)) \quad \text{and} \quad W_t(x) = \langle \dot{h}_t(x) \wedge \xi_t(x), \mathbf{E}^n\rangle.$$

Suppose $x \in \mathbf{b}(D)$. Let $u_1, \ldots, u_n$ be an orthonormal sequence of vectors in $\mathbb{R}^n$ such that $u_1 = \mathbf{n}_D(x)$ and $*u_1 = u_2 \wedge \cdots \wedge u_n$; since $\langle u_1 \wedge *u_1, \mathbf{E}^n\rangle = 1$ we have

$$\langle w \wedge u_2 \wedge \cdots \wedge u_n, \mathbf{E}^n\rangle = w \bullet u_1 \langle u_1 \wedge *u_1, \mathbf{E}^n\rangle = w \bullet u_1 \quad \text{for any } w \in \mathbb{R}^n;$$

see [FE, sect. 1.7.8] for the properties of $*$.

It should now be clear from Proposition 3.2 that (3.1) holds.

Let $u^1, \ldots, u^n$ be the sequence of covectors dual to $u_1, \ldots, u_n$ and let $\omega_1, \ldots, \omega_n$ be those covectors such that $\partial X(x) = \sum_{j=1}^n \omega_j u_j$. We have

$$\begin{aligned}
\frac{d}{dt} W_t(x)\Big|_{t=0} &= X(x) \wedge \frac{d}{dt}\xi_t(x)\Big|_{t=0} \\
&= X(x) \wedge \sum_{i=2}^n \partial X(x)(u_i) \wedge \big(u^i \lrcorner\, \xi_0(x)\big) \\
&= X(x) \wedge \sum_{i=2}^n \sum_{j=1}^n \langle u_i, \omega_j\rangle u_j \wedge \big(u^i \lrcorner\, \xi_0(x)\big) \\
&= X(x) \wedge \sum_{i=1}^n \langle u_i, \omega_i\rangle u_i \wedge \big(u^i \lrcorner\, \xi_0(x)\big) \\
&\qquad + X(x) \wedge \sum_{i=2}^n \langle u_i, \omega_1\rangle u_1 \wedge \big(u^i \lrcorner\, \xi_0(x)\big) \\
&= ((\operatorname{trace} l_1(x) X(x) - l_2(x)(X(x))) \bullet \mathbf{n}_D(x))\, u_1 \wedge *u_1,
\end{aligned}$$

so (3.2) holds. ∎

4. Second fundamental forms and mean curvature. Suppose M is an embedded hypersurface of class C^2 in Ω.

The *second fundamental form of* M is the function Π on M whose value at $a \in M$ is a linear map from $\mathbf{Nor}(M,a)$ into the symmetric linear maps from $\mathbf{Tan}(M,a)$ to itself characterized by the requirement that if U is an open subset of $\mathbb{R}^n$, $a \in U \cap M$; $N : U \to \mathbb{R}^n$; N is of class C^1; and $N(x) \in \mathbf{Nor}(M,x)$ whenever $x \in U \cap M$, then

$$\Pi(a)(N(a))(v) \bullet w = \partial N(a)(v) \bullet w \quad \text{for } v, w \in \mathbf{Tan}(M,a).$$

The *mean curvature vector of* M is, by definition, the function H on M whose value at a point a of M is that member $H(a)$ of $\mathbf{Nor}(M,a)$ whose inner product with $u \in \mathbf{Nor}(M,a)$ is the trace of $\Pi(a)(u)$. In the classical literature the mean curvature vector is $1/(n-1)$ times H as defined here, hence the word "mean." It turns out the factor $1/(n-1)$ is inconvenient when one is working, as we will be, with the first variation of area; for this reason we omit it. The direction of the mean curvature vector, and not just its magnitude, will be important in this work.

If $a \in M$, the *length of* $\Pi(a)$ is, by definition, the square root of the sum of the squares of the eigenvalues of $\Pi(a)(u)$ whenever $u \in \mathbf{Nor}(M,a)$ and $|u| = 1$.

Suppose $f : \Omega \to \mathbb{R}$ is C^2; $\nabla f(x) \neq 0$ whenever $x \in \Omega$; y is in the range of f; and $M = \{f = y\}$, so M is an embedded hypersurface of class C^2 in Ω. It follows that if $a \in M$ then

$$\Pi(a)(\nabla f(a))(u) \bullet v = \partial(\nabla f)(a)(u) \bullet v \quad \text{whenever } u, v \in \mathbf{Tan}(M,a).$$

Suppose $\Omega = \mathbb{R}^n \sim \{0\}$, $f(x) = |x|^2/2$ for $x \in \Omega$, $0 < R < \infty$, and $M = \{x \in \mathbb{R}^n : |x| = R\}$. Then $\nabla f(x) = x$ for $x \in \Omega$. It follows that if $a \in M$ then

$$\Pi(a)(a)(v) \bullet w = \frac{v \bullet w}{|a|} \quad \text{whenever } v, w \in \mathbf{Tan}(M,a), \qquad H(a) = \frac{n-1}{R^2} a$$

and the length $\Pi(a)$ equals the square root of $(n-1)/R^2$.

5. The spaces $\boldsymbol{\mathcal{B}_\lambda(\Omega)}$ and $\boldsymbol{\mathcal{C}_\lambda(\Omega)}$, $\boldsymbol{0 \le \lambda < \infty}$. We suppose throughout this section that $0 \le \lambda < \infty$, and we study the spaces $\mathcal{B}_\lambda(\Omega)$ and $\mathcal{C}_\lambda(\Omega)$.

5.1. Basic results on $\boldsymbol{\mathcal{B}_\lambda(\Omega)}$ and $\boldsymbol{\mathcal{C}_\lambda(\Omega)}$. In what follows we will frequently make use of the following simple observation. *Suppose $f, g \in \mathbf{L}_1^{loc}(\Omega)$, K is a compact subset of Ω, $g \in \mathbf{k}(f,K)$, and $y \in \mathbb{R}$. Then*

$$\{g > y\} \in \mathbf{k}(\{f > y\}). \tag{5.1}$$

Moreover, $\geq$ may be replaced by any of $\leq$, $>$, and $<$.

Remark 5.1. It is an elementary corollary of Theorem 5.1 below that if D is an open subset of Ω with smooth boundary M and $D \in \mathcal{C}_\lambda(\Omega)$, then the length of the mean curvature vector of M does not exceed λ. The converse of this statement is false as one sees in case $\lambda = 0$ by considering a set whose boundary is an unstable minimal surface.

However, if $f : \Omega \to \mathbb{R}$ is smooth with nowhere vanishing gradient and, for each y in the range of f, the length of the mean curvature vector of $\{f = y\}$ never exceeds λ, then a simple calibration argument shows that $f \in \mathcal{B}_\lambda(\Omega)$.

LEMMA 5.1. *Suppose $f \in \mathcal{B}_\lambda(\Omega)$, $g \in \mathbf{BV}^{loc}(\Omega)$, K is a compact subset of Ω, $u(x) = \mathbf{dist}\,(x,K)$ for $x \in \Omega$, $0 < h < \infty$, and $\{u \le h\}$ is a compact subset of Ω.*

Then

$$||\partial[f]||(K) \leq ||\partial[g]||(\{u \leq h\}) + \left(\lambda + \frac{1}{h}\right) \int_{\{u \leq h\}} |f - g| \, d\mathcal{L}^n.$$

In particular,

$$||\partial[f]||(K) \leq \left(\lambda + \frac{1}{h}\right) \int_{\{u \leq h\}} |f - y| \, d\mathcal{L}^n \quad \textit{for } y \in \mathbb{R}.$$

Proof. For each $r \in (0, h)$ let $h_r = g1_{\{u \leq r\}} + f1_{\{u > r\}}$. Then $h_r \in \mathbf{k}(f, \{u \leq r\})$ and $f - h_r = (f - g)1_{\{u \leq r\}}$, so

$$||\partial[f]||(\{u \leq r\}) \leq ||\partial[h_r]||(\{u \leq r\}) + \lambda \int_{\{u \leq r\}} |f - g|.$$

Now integrate this inequality from 0 to h and make use of (2.10) to prove the first inequality; to obtain the second, set $g(x) = y$ for $x \in \Omega$. ∎

THEOREM 5.1. *Suppose* $\lambda \in [0, \infty)$, $f \in \mathcal{B}_\lambda(\Omega)$, *and* $y \in \mathbf{R}$. *Then*

$$\{f + y, yf, f \wedge y, f \vee y\} \subset \mathcal{B}_\lambda(\Omega).$$

Proof. Suppose K is a compact subset of Ω. Obviously, $0f = 0 \in \mathcal{B}_\lambda(\Omega)$. Suppose $y \in \mathbb{R} \sim \{0\}$ and $g \in \mathbf{k}(yf, K)$. Then $g/y \in \mathbf{k}(f, K)$, so

$$\begin{aligned} ||\partial[yf]||(K) &= |y| ||\partial[f]||(K) \\ &\leq |y| \left(||\partial[g/y]||(K) + \lambda \int_\Omega |f - g/y| \, d\mathcal{L}^n \right) \\ &= ||\partial[g]||(K) + \lambda \int_\Omega |yf - g| \, d\mathcal{L}^n. \end{aligned}$$

Thus $yf \in \mathcal{B}_\lambda(\Omega)$.

Suppose $g \in \mathbf{k}(f + y, K)$. Then $g - y \in \mathbf{k}(f, K)$, so

$$\begin{aligned} ||\partial[f + y]||(K) &= ||\partial[f]||(K) \\ &\leq ||\partial[g - y]||(K) + \lambda \int_\Omega |f - (g - y)| \, d\mathcal{L}^n \\ &= ||\partial[g]||(K) + \lambda \int_\Omega |(f + y) - g| \, d\mathcal{L}^n, \end{aligned}$$

and thus $f + y \in \mathcal{B}_\lambda(\Omega)$.

Suppose $g \in \mathbf{k}(f \wedge y, K)$. Let $h = g + (f \vee y) - y$. Then $f - h = f + y - (f \vee y) - g = f \wedge y - g$, so $h \in \mathbf{k}(f, K)$. Using Proposition 2.3 we estimate

$$\begin{aligned} ||\partial[f \wedge y]||(K) &+ ||\partial[f \vee y]||(K) \\ &= ||\partial[f]||(K) \\ &\leq ||\partial[h]||(K) + \lambda \int_K |f - h| \, d\mathcal{L}^n \\ &\leq ||\partial[g]||(K) + ||\partial[f \vee y]||(K) + \lambda \int_K |f \wedge y - g| \, d\mathcal{L}^n \end{aligned}$$

and conclude that $f \wedge y \in \mathcal{B}_\lambda(\Omega)$.

Finally, $f \vee y = -((-f) \wedge (-y)) \in \mathcal{B}_\lambda(\Omega)$. $\square$

THEOREM 5.2. *Suppose $\lambda \in [0, \infty)$, f is a sequence in $\mathcal{B}_\lambda(\Omega)$, $F \in \mathbf{L}_1^{loc}(\Omega)$, and $f_\nu \to F$ in $\mathbf{L}_1^{loc}(\Omega)$. Then $F \in \mathcal{B}_\lambda(\Omega)$ and*

$$||\partial[f_\nu]|| \to ||\partial[F]|| \quad \textit{weakly as } \nu \to \infty.$$

Proof. Let K be a compact subset of Ω, let $u(x) = \mathbf{dist}\,(x, K)$ for $x \in \Omega$, and let $R = \sup\{r \in (0, \infty) : \{u \leq r\} \subset \Omega\}$.

Suppose $h \in (0, R)$ and for each positive integer ν let y_ν be the average of f_ν on $\{u \leq h\}$. Let Y be the average value of F on $\{u \leq h\}$. From Lemma 5.1 we obtain

$$||\partial[f_\nu]||(K) \leq \left(\lambda + \frac{1}{h}\right) \int_{\{u \leq h\}} |f_\nu - y_\nu|\, d\mathcal{L}^n \to \left(\lambda + \frac{1}{h}\right) \int_{\{u \leq h\}} |F - Y|\, d\mathcal{L}^n$$

as $\nu \to \infty$. Since K is arbitrary we infer from (2.1) that $F \in \mathbf{BV}^{loc}(\Omega)$.

For any $r \in (0, R)$ we infer from Lemma 5.1 that

$$||\partial[f_\nu]||(K) \leq ||\partial[F]||(\{u \leq r\}) + \left(\lambda + \frac{1}{h}\right) \int_{\{u \leq h\}} |f_\nu - F|\, d\mathcal{L}^n$$

for any positive integer ν. Keeping in mind (2.1) we conclude that $||\partial[f_\nu]||$ converges weakly to $||\partial[F]||$ as $\nu \to \infty$.

We now show that $F \in \mathcal{B}_\lambda(\Omega)$. To this end, let $G \in \mathbf{BV}^{loc}(\Omega) \cap \mathbf{k}(F, K)$. For each positive integer ν and each $\rho \in (0, R)$ we let

$$g_{\nu,\rho} = G \llcorner \{u \leq \rho\} + f_\nu \llcorner \{u > \rho\},$$

we note that $g_{\nu,\rho} \in \mathbf{k}(f_\nu, \{u \leq \rho\})$ and $f_\nu - g_{\nu,\rho} = (f_\nu - G)1_{\{u \leq \rho\}}$, and we conclude that

$$||\partial[f_\nu]||(\{u \leq \rho\}) \leq ||\partial[g_{\nu,\rho}||(\{u \leq \rho\}) + \lambda \int_{\{u \leq \rho\}} |G - f_\nu|\, d\mathcal{L}^n.$$

Suppose $0 < r < R$ and ν is a positive integer. Keeping in mind that $G - f_\nu = F - f_\nu$ at $\mathcal{L}^n$ almost all points of $\Omega \sim K$, we integrate this inequality from 0 to r and use (2.10) to obtain

$$\begin{aligned} r||\partial[f_\nu]||(K) &\leq \int_0^r ||\partial[f_\nu]||(\{u \leq \rho\})\, d\mathcal{L}^1 \rho \\ &\leq \int_{\{0 < u < r\}} |F - f_\nu|\, d\mathcal{L}^n + r||\partial[G]||(\{u \leq r\}) \\ &\qquad + \lambda r \int_{\{u \leq r\}} |G - f_\nu|\, d\mathcal{L}^n. \end{aligned}$$

Letting $\nu \to \infty$ we find that

$$\limsup_{\nu \to \infty} ||\partial[f_\nu]||(K) \leq ||\partial[G]||(\{u \leq r\}) + \lambda \int_{\{u \leq r\}} |G - F|\, d\mathcal{L}^n.$$

Letting $r \downarrow 0$ we infer that

$$||\partial[F]||(K) \leq ||\partial[G]||(K) + \lambda \int_K |G - F|\, d\mathcal{L}^n,$$

as desired. $\square$

THEOREM 5.3. *The following statements hold:*
(i) *If $f \in \mathcal{B}_\lambda(\Omega)$ and $y \in \mathbb{R}$, then $\{f > y\} \in \mathcal{C}_\lambda(\Omega)$.*
(ii) *If $f \in \mathbf{BV}^{loc}(\Omega)$, $D = \{y \in \mathbb{R} : \{f > y\} \in \mathcal{C}_\lambda(\Omega)\}$, and D is dense in $\mathbb{R}$, then $f \in \mathcal{B}_\lambda(\Omega)$.*
(iii) *If $\mathcal{E}$ is a nonempty nested subfamily of $\mathcal{C}_\lambda(\Omega)$, then $\cup\mathcal{E}$ and $\cap\mathcal{E}$ belong to $\mathcal{C}_\lambda(\Omega)$.*
(iv) *$E \in \mathcal{C}_\lambda(\Omega)$ if and only if $1_E \in \mathcal{B}_\lambda(\Omega)$ whenever $E \subset \Omega$.*

Proof. We begin with a lemma.

LEMMA 5.2. *Suppose $f \in \mathbf{BV}^{loc}(\Omega)$, $D = \{y \in \mathbb{R} : \{f > y\} \in \mathcal{C}_\lambda(\Omega)\}$, and $\mathcal{L}^1(\mathbb{R} \sim D) = 0$. Then $f \in \mathcal{B}_\lambda(\Omega)$.*

Proof. Suppose K is a compact subset of Ω and $g \in \mathbf{BV}^{loc}(\Omega) \cap \mathbf{k}(f, K)$. Keeping in mind (5.1) we infer from (2.14) and (2.16) that

$$\begin{aligned} ||\partial[f]||(K) &= \int_{-\infty}^{\infty} ||\partial[1_{\{f>y\}}]||(K)\, d\mathcal{L}^1 y \\ &\le \int_{-\infty}^{\infty} \left(||\partial[1_{\{g>y\}}]||(K) + \lambda \int |1_{\{f>y\}} - 1_{\{g>y\}}|\, d\mathcal{L}^n \right) d\mathcal{L}^1 y \\ &= ||\partial[g]||(K) + \lambda \int |f-g|\, d\mathcal{L}^n. \qquad \square \end{aligned}$$

Suppose $E \in \mathcal{C}_\lambda(\Omega)$. Evidently, $\{1_E > y\} \in \mathcal{C}_\lambda(\Omega)$ for all $y \in \mathbb{R}$ so, by the lemma, $1_E \in \mathcal{B}_\lambda(\Omega)$. It being trivial that $\{E : 1_E \in \mathcal{B}_\lambda(\Omega)\}$ is a subset of $\mathcal{C}_\lambda(\Omega)$, we find that (iv) holds.

Suppose $\mathcal{E}$ is a nonempty nested subfamily of $\mathcal{C}_\lambda(\Omega)$. Choose a nondecreasing sequence A and a nonincreasing sequence B in $\mathcal{E}$ such that $1_{A_\nu} \to 1_{\cup\mathcal{E}}$ and $1_{B_\nu} \to 1_{\cap\mathcal{E}}$ in $\mathbf{L}_1^{loc}(\Omega)$, as $\nu \to \infty$. From Theorem 5.2 we infer that the indicator functions of $\cup\mathcal{E}$ and $\cap\mathcal{E}$ belong to $\mathcal{B}_\lambda(\Omega)$, so (iii) now follows from (iv).

Suppose f and D are as in (ii). Since D is dense in $\mathbb{R}$ we have for any $y \in \mathbb{R}$ that

$$\{f > y\} = \cup_{z \in (y,\infty)\cap D}\{f > z\},$$

so $\{f > y\} \in \mathcal{C}_\lambda(\Omega)$ by (iii). The lemma now implies (ii).

Finally, suppose $f \in \mathcal{B}_\lambda(\Omega)$ and $y \in \mathbb{R}$. For each positive integer ν let

$$g_\nu = \nu\left(\left((f-y) \wedge \frac{1}{\nu}\right) \vee 0\right)$$

and note that $g_\nu \in \mathcal{B}_\lambda(\Omega)$ by Theorem 5.1. One readily verifies that $g_\nu \uparrow 1_{\{f>y\}}$ as $\nu \uparrow \infty$ so that, by Theorem 5.2, $1_{\{f>y\}} \in \mathcal{B}_\lambda(\Omega)$, so $\{f > y\} \in \mathcal{C}_\lambda(\Omega)$ by (iv), and thus (i) holds. $\square$

5.2. Generalized mean curvature.

PROPOSITION 5.1. *Suppose $\lambda \in [0, \infty)$, $D \in \mathcal{C}_\lambda(\Omega)$, and $X \in \mathcal{X}(\Omega)$. Then*

$$\int \operatorname{trace} P(x) \circ \partial X(x) \circ P(x)\, d||\partial[D]||x \le \lambda \int |X|\, d||\partial[D]||,$$

where, for each $x \in \mathbf{b}(D)$, we have let $P(x)$ be an orthogonal projection of $\mathbb{R}^n$ onto $\{v \in \mathbb{R}^n : v \bullet \mathbf{n}_D(x) = 0\}$.

Remark 5.2. We restate this theorem in the language of [AW1]. Let V be the $(n-1)$-dimensional varifold in Ω naturally associated to $\partial[D]$ as in [AW1, sect. 3.5]; the preceding theorem says that

$$||\delta V|| \le \lambda ||V||,$$

where δV is as in [AW1, sect. 4.2].

Proof. Let us adopt the notation of section 3. In particular, $A(t) = ||\partial[E_t]||(K)$ for $t \in I$. For any positive $t \in I$ we infer from Proposition 3.2 that

$$\frac{A(t) - A(0)}{t} \leq \frac{\lambda}{t}||[E_t] - [D]||(K) \leq \frac{1}{t}\lambda \int_0^t \left(\int |X|||\partial \dot{h}_\tau(x)||^{n-1} \, d||\partial[D]||x \right) d\mathcal{L}^1\tau.$$

The estimate to be proved now follows from Proposition 3.1. $\square$

5.3. Consequences of the monotonicity theorem.

THEOREM 5.4. *Suppose* $\lambda \in [0, \infty)$, $D \in \mathcal{C}_\lambda(\Omega)$, $a \in \Omega$, *and* $R = \mathbf{dist}\,(a, \mathbb{R}^n \sim \Omega)$. *Then*

(i) $(0, R) \ni r \mapsto e^{\lambda r}\Theta^{n-1}(||\partial[D]||, a, r)$ *is nondecreasing;*

(ii) $\Theta^{n-1}(||\partial[D]||, a)$ *exists and depends uppersemicontinuously on* a*;*

(iii) $\Theta^{n-1}(||\partial[D]||, a) \geq 1$ *if* $a \in \mathbf{spt}\,\partial[D]$*;*

if $a \in \mathbf{spt}\,[D]$, *we have*

(iv) $e^{-\lambda r}\alpha(n-1)r^{n-1} \leq ||\partial[D]||(\mathbf{U}^n(a, r))$ *whenever* $0 < r < R$*;*

(v) $e^{-\lambda r}\frac{\alpha(n-1)}{n}r^n \leq (1 + \lambda r)\mathcal{L}^n(D \cap \mathbf{U}^n(a, r))$ *whenever* $0 < r < R$.

Proof. In view of Remark 5.2, (i) follows from the monotonicity theorem of [AW1, sect. 5.1]. (i) clearly implies (ii). (iii) is a consequence of Theorem 2.1(ii) and (iii). (iv) follows directly from (i) and (iii).

Suppose $0 < r < R$. For each $\rho \in (0, r)$ let $E_\rho = D \cap \{u > \rho\}$, where we have set $u(x) = |x - a|$ for $x \in \Omega$ and note that $E_\rho \in \mathbf{k}(E, \{u \leq \rho\})$, so

$$\begin{aligned} e^{-\lambda r}\alpha(n-1)\rho^{n-1} &\leq e^{-\lambda\rho}\alpha(n-1)\rho^{n-1} \\ &\leq ||\partial[D]||(\{u \leq \rho\}) \\ &\leq ||\partial[E_\rho]||(\{u \leq \rho\}) + \lambda\Sigma_\Omega(E_\rho, D). \end{aligned}$$

Now integrate this inequality over $(0, r)$ and make use of (2.10), with f and g there equal to 1_E and 0, respectively. $\square$

Remark 5.3. It follows from (iv) that if $\Omega = \mathbb{R}^n$ and $\mathcal{L}^n(D) < \infty$, then $\mathbf{spt}\,[D]$ is compact.

COROLLARY 5.1. *Suppose* $0 < R < \infty$, $0 < r < \infty$, $a \in \Omega$, $R + r \leq \mathbf{dist}\,(a, \mathbb{R}^n \sim \Omega)$, $f \in \mathcal{B}_\lambda(\Omega)$, *and*

$$Y = \{y \in \mathbb{R} : ||\partial[\{f > y\}]||(\mathbf{U}^n(a, R)) > 0\}.$$

Then

$$\mathcal{L}^1(Y)e^{-\lambda r}\alpha(n-1)r^{n-1} \leq ||\partial[f]||(\mathbf{U}^n(a, R + r))$$

and

$$\mathcal{L}^1(Y)e^{-\lambda r}\frac{\alpha(n-1)}{n}r^n \leq (1 + \lambda r)\int_{\mathbf{U}^n(a, R+r)} |f| \, d\mathcal{L}^n.$$

Proof. For each $y \in Y \sim \{0\}$ we apply Theorem 5.4 with D there equal to $\{f > y\}$ to a ball of radius r with center at a point b, where $\Theta^{n-1}(||\partial[\{f > y\}]||, b) = 1$. $\square$

5.4. Proof of the regularity theorem for $\mathcal{C}_\lambda(\Omega)$. In view of the regularity theorem of [AW1, sect. 8] the present regularity theorem, Theorem 1.2, will follow from the following lemma.

LEMMA 5.3. *Suppose*

$$1 < \zeta < \infty.$$

There exists $\eta \in (0,1)$ *such that if* $0 \le \lambda < \infty$, $a \in \mathbb{R}^n$, $0 < R < \infty$,

$$\lambda R \le \eta, \quad E \in \mathcal{C}_\lambda(\mathbf{U}^n(a,R)), \quad \textit{and} \quad a \in \mathbf{spt}\,\partial[E],$$

then

$$\Theta^{n-1}(||\partial[E]||, a, \eta R) \le \zeta.$$

Proof. Due to the way the various entities in the lemma change under application of homotheties and translations, we find that we may assume without loss of generality that $a = 0$ and $R = 1$.

Suppose the lemma were false. Then there would exist $\zeta \in (1,\infty)$; a sequence η in $(0,1)$ with limit zero; and sequences E, λ such that, for each positive integer ν,

$$\lambda_\nu \le \eta_\nu, \quad E_\nu \in \mathcal{C}_{\lambda_\nu}(\mathbf{U}^n(0,1)), \quad \text{and} \quad 0 \in \mathbf{spt}\,\partial[E_\nu]$$

but such that

$$\Theta^{n-1}(||\partial[E_\nu]||, 0, \eta_\nu) > \zeta. \tag{5.2}$$

From the monotonicity theorem we have

$$(0,1) \ni t \mapsto e^{\lambda_\nu t}\Theta^{n-1}(||\partial[E_\nu]||, 0, t) \quad \text{is nondecreasing} \tag{5.3}$$

for each positive integer ν.

Replacing E by a subsequence if necessary we may use Theorems 2.2 and 5.2 to obtain a Lebesgue measurable subset F of $\mathbf{U}^n(0,1)$ such that $E_\nu \to F$ in $\mathbf{L}_1^{loc}(\mathbf{U}^n(0,1))$ as $\nu \to \infty$,

$$F \in \bigcap_{\nu=1}^{\infty} \mathcal{C}_{\lambda_\nu}(\mathbf{U}^n(0,1)) = \mathcal{C}_0(\mathbf{U}^n(0,1)), \tag{5.4}$$

and

$$||\partial[E_\nu]|| \to ||\partial[F]|| \quad \text{weakly as } \nu \to \infty. \tag{5.5}$$

Letting B equal the set of $t \in (0,1)$ such that $||\partial[F]||(\{x \in \mathbb{R}^n : |x| = t\})$ is positive we observed that B is countable and infer from (5.5) and (2.1) that

$$\lim_{\nu\to\infty} \Theta^{n-1}(||\partial[E_\nu], 0, t) = \Theta^{n-1}(||\partial[F], 0, t) \quad \text{for any } t \in (0,1) \sim B.$$

This together with (5.2), Theorem 5.4, and the fact that $\lambda_\nu \to 0$ as $\nu \to \infty$ implies

$$\Theta^{n-1}(||\partial[F]||, 0, t) \ge \zeta \quad \text{whenever } t \in (0,1) \sim B. \tag{5.6}$$

As $F \in \mathcal{C}_0(\mathbf{U}^n(0,1))$ we find that $\partial[F]$ is an absolutely area minimizing integral current of dimension $n-1$ in $\mathbf{U}^n(0,1)$. As Theorem 2.1 implies that

$$\Theta^{n-1}(||\partial[F]||, x) = 1 \quad \text{for } ||\partial[F]|| \text{ almost all } x$$

it follows from the regularity theorem of [FE, sect. 5.4.15] that $\partial[F]$ is integration over an oriented $(n-1)$-dimensional real analytic hypersurface M of $\mathbf{U}^n(0,1)$. Consequently, $\Theta^{n-1}(||\partial[F]||, 0) = 1$, which is incompatible with (5.6). $\square$

5.4.1. The case $n = 2$. One can do a little better than the preceding theorem if $n = 2$, as follows. Let $\mathbf{w}(m) = \sqrt{1+m^2}$ for $m \in \mathbf{R}$. Suppose V and W are nonempty open intervals, $g : V \to W$ is continuously differentiable, $0 \le \lambda < \infty$, and

$$D = \{(v,w) \in V \times W : w < g(v)\} \in \mathcal{C}_\lambda(V \times W).$$

We will show that

$$\mathbf{Lip}(\mathbf{w}' \circ g') \le \lambda. \tag{5.7}$$

Note that if g is twice differentiable, then $(\mathbf{w}' \circ g')'$ is the curvature function of the graph of g.

We prove (5.7) as follows. Suppose $\phi \in \mathcal{D}(V)$ and for each $t \in \mathbb{R}$ let $g_t = g + t\phi$. Let I be an open interval containing 0 such that if $t \in I$, then $g_t(v) \in W$ whenever $v \in V$. Let $E_t = \{(v,w) \in V \times W : w < g_t(v)\}$ for $t \in I$. We have

$$(||\partial[D]|| - ||\partial[E_t]||)(V \times W) \le \lambda \Sigma_{V\times W}(D, E_t).$$

Now

$$\lim_{t\downarrow 0} \frac{1}{t}(||\partial[D]|| - ||\partial[E_t]||)(V \times W) = -\int_I \frac{d}{dt}\mathbf{w}(g' + t\phi')\big|_{t=0}\, d\mathcal{L}^1 = -\int_I (\mathbf{w}' \circ g')\phi'\, d\mathcal{L}^1$$

and

$$\lim_{t\downarrow 0} \frac{1}{t}\Sigma_{V\times W}(D, E_t) = \int_V |\phi|\, d\mathcal{L}^1.$$

To obtain (5.7) we let ϕ approximate plus or minus one times the indicator function of a compact subinterval of V.

6. Locality. Suppose M and μ are as in Proposition 1.3.

6.1. Proof of Proposition 1.3. If M has a representation as in (1.2) where m is a bounded Borel function, it is trivial that M is local.

Suppose M is local. Then $||\mu||_{\mathbf{L}_\infty(\Omega)} \le \mathbf{l}(M)$, which implies there is a unique Radon measure on Ω whose restriction to $\mathcal{M}(\Omega)$ equals $\hat{M}$. That (1.2) holds with $m = \mu$ follows from the theory of symmetrical derivation; see, for example, [FE, sect. 2.9].

6.2. Proof of Proposition 1.4. If F has a representation as in (1.3) where k satisfies (i) and (ii) of Proposition 1.4, it is trivial that M is local, and it follows from the theory of symmetrical derivation that for $0 < y < \infty$ we have $k(x,y) = \kappa(x,y)$ for $\mathcal{L}^n$ almost all $x \in \Omega$.

Suppose F is local. For any $y \in (0,\infty)$ we have that $\mathcal{M}(\Omega) \ni E \mapsto \hat{F}(y1_E)$ is local so that, by Proposition 1.3,

$$\hat{F}(y1_E) = \int_E \kappa(x,y)\, d\mathcal{L}^n x \quad \text{for } E \in \mathcal{M}(\Omega).$$

Given $f \in \mathcal{F}(\Omega)$ and $0 = y_0 < y_1 < y_2 < \cdots < y_N < \infty$ we infer from the locality of F that

$$\hat{F}\left(\sum_{i=1}^N y_i 1_{\{y_{i-1}<f\le y_i\}}\right) = \sum_{i=1}^N \hat{F}\left(y_i 1_{\{y_{i-1}<f\le y_i\}}\right) = \sum_{i=1}^N \int_{\{y_{i-1}<f\le y_i\}} \kappa(x,y_i)\, d\mathcal{L}^n x,$$

from which the representation for $F(f)$ in (1.3) easily follows using the admissibility of F, and (i) and (ii) of Proposition 1.4 hold with $k = \kappa$.

6.3. Proof of Proposition 1.5. (i), (ii), and (iii) are immediate.

For any $x \in \Omega$ we have that $(0,\infty) \ni y \mapsto \kappa(x,y)$ is absolutely continuous so that

$$\kappa(x, f(x)) = \int_0^{f(x)} u(x,z)\, d\mathcal{L}^1 z = \int_0^{\infty} u(x,z) 1_{\{f \geq z\}}(x)\, d\mathcal{L}^1 z.$$

Integrating this equation over Ω and invoking Fubini's theorem, we infer that

$$\int_\Omega \kappa(x, f(x))\, d\mathcal{L}^n x = \int_0^\infty \left(\int_{\{f \geq z\}} u(x,z)\, d\mathcal{L}^n x \right) d\mathcal{L}^1 z. \tag{6.1}$$

For each $E \in \mathcal{M}(\Omega)$ let $\zeta(E)$ be the set of $y \in (0,\infty)$ such that

$$U_y(E) = \int_E u(x,y)\, d\mathcal{L}^n x.$$

Since $(0,\infty) \ni y \mapsto \hat{F}(y 1_E)$ is absolutely continuous we find that

$$\hat{F}(y 1_E) = \int_0^y U_y(E)\, d\mathcal{L}^1 y \quad \text{whenever } y \in (0,\infty).$$

Now assume F is local. Applying Proposition 1.4 together with (6.1) with f equal to $y1_E$, we find that

$$\int_0^y \left(\int_E u(x,z)\, d\mathcal{L}^n x \right) d\mathcal{L}^1 z = \int_0^y U_z(E)\, d\mathcal{L}^1 z \quad \text{for } y \in (0,\infty),$$

which implies that $\mathcal{L}^1((0,\infty) \sim \zeta(E)) = 0$. Let $\mathcal{E}$ be a countable subfamily of $\mathcal{M}(\Omega)$ which is dense with respect to $\Sigma_\Omega(\cdot,\cdot)$ and let $Z = \cap\{\zeta(E) : E \in \mathcal{E}\}$. Since $\mathcal{M}(\Omega) \ni E \mapsto U_y(E)$ and $\mathcal{M}(\Omega) \ni E \mapsto \int_E u(x,y)\, d\mathcal{L}^n x$ are Lipschitzian with respect to $\Sigma_\Omega(\cdot,\cdot)$, we find that

$$U_y(E) = \int_E u(x,y)\, d\mathcal{L}^n x \quad \text{whenever } y \in Z \text{ and } E \in \mathcal{M}(\Omega). \tag{6.2}$$

Since $\mathcal{L}^1((0,\infty) \sim Z) = 0$ we find that (iv) of Proposition 1.5 holds.

Suppose $f \in \mathcal{F}(\Omega)$. Use (1.3) with $k = \kappa$ to represent $F(f)$. (v) now follows from (6.1) and (6.2).

6.4. Proof of Proposition 1.6. That (i) implies (ii) is immediate. That (ii) implies (iii) is a direct consequence of the subadditivity of $\limsup$. That (iii) implies (i) follows directly from (v) of Proposition 1.5. Thus (i), (ii), and (iii) are equivalent.

We leave the proof of the following elementary lemma to the reader.

LEMMA 6.1. *Suppose $g : \mathbb{R} \to \mathbb{R}$, g is absolutely continuous, and*

$$h(y) = \liminf_{z \to y} \frac{g(z) - g(y)}{z - y} \quad \textit{for } y \in \mathbb{R}.$$

Then g is convex if and only if h is nondecreasing. Moreover, if g is convex, then h is right continuous.

The lemma implies that (iii) and (v) are equivalent. Since the admissibility of F implies that $\mathbb{R} \ni y \mapsto \hat{F}(y_E)$ is locally Lipschitzian for any $E \in \mathcal{M}(\Omega)$, the lemma implies that (ii) and (iv) are equivalent.

The final assertion follows from the right continuity assertion of the lemma.

6.5. The class $\mathcal{G}(\Omega)$. Let

$$p : \Omega \times (0, \infty) \to \Omega \quad \text{and} \quad q : \Omega \times (0, \infty) \to (0, \infty)$$

carry $(x, y) \in \Omega \times (0, \infty)$ to x and y, respectively.

Whenever G is an $\mathcal{L}^n \times \mathcal{L}^1$ measurable subset of $\Omega \times (0, \infty)$ we let

$$[G] \in \mathcal{D}_{n+1}(\Omega \times (0, \infty))$$

be as in (2.2), with $\mathbf{V}^n$ there replaced by $(p^{\#}\mathbf{V}^n) \wedge dq$; that is,

$$[G](\psi(p^{\#}\mathbf{V}^n) \wedge dq) = \int_G \psi \, d(\mathcal{L}^n \times \mathcal{L}^1) \quad \text{whenever } \psi \in \mathcal{D}(\Omega \times (0, \infty)).$$

DEFINITION 6.1. *We let*

$$\mathcal{G}(\Omega)$$

be the family of Lebesgue measurable subsets G of $\Omega \times (0, \infty)$ such that

$$(\mathcal{L}^n \times \mathcal{L}^1)(G) < \infty \quad \textit{and} \quad q[\mathbf{spt}\,[G]] \textit{ is bounded.}$$

Note that if $G \in \mathcal{G}(\Omega)$, then for $\mathcal{L}^1$ almost all $y \in (0, \infty)$ we have $\{x : (x, y) \in G\} \in \mathcal{M}(\Omega)$.

DEFINITION 6.2. *Whenever $G \in \mathcal{G}(\Omega)$ we let*

$$G^{\downarrow} : \Omega \to \mathbb{R}$$

be such that

$$G^{\downarrow}(x) = \begin{cases} \mathcal{L}^1(\{y : (x, y) \in G\}) & \textit{if } \{y : (x, y) \in G\} \in \mathcal{M}((0, \infty)), \\ 0 & \textit{otherwise.} \end{cases}$$

Note that $G^{\downarrow} \in \mathcal{F}(\Omega)$ and $\int_\Omega G^{\downarrow} \, d\mathcal{L}^n = (\mathcal{L}^n \times \mathcal{L}^1)(G)$.

DEFINITION 6.3. *Whenever $f : \Omega \to [0, \infty)$ we let*

$$f^{\uparrow} = \{(x, y) \in \Omega \times (0, \infty) : f(x) > y\}.$$

Suppose $f : \Omega \to [0, \infty)$. Evidently,

$$f \in \mathcal{F}(\Omega) \iff f^{\uparrow} \in \mathcal{G}(\Omega).$$

Tonelli's theorem implies that

$$[(f^{\uparrow})^{\downarrow}] = [f] \quad \text{whenever } f \in \mathcal{F}(\Omega).$$

PROPOSITION 6.1. *Suppose $G \in \mathcal{G}(\Omega)$, $\phi \in \mathcal{D}(\Omega)$, and $\Psi \in \mathcal{E}((0, \infty))$. Then*

$$\begin{aligned} &p_{\#} \left(\partial[G] \llcorner \Psi \circ q\right) (\phi \mathbf{V}^{\mathbf{n}}) \\ &\quad = (-1)^n [G] \left(p^{\#}(\phi \mathbf{V}^n) \wedge (\Psi' \circ q) dq\right) \\ &\quad = (-1)^n \int_\Omega \phi(x) \left(\int_{\{y:(x,y) \in G\}} \Psi' \, d\mathcal{L}^1 \right) d\mathcal{L}^n x. \end{aligned} \tag{6.3}$$

Proof. The first equation follows from the fact that

$$d((\Psi \circ q) \wedge p^{\#}(\phi \mathbf{V}^n)) = (\Psi' \circ q) dq \wedge p^{\#}(\phi \mathbf{V}^n),$$

and the second follows from Fubini's theorem. ◻

COROLLARY 6.1. *Suppose $G \in \mathcal{G}(\Omega)$. Then*

$$[G^{\downarrow}] = (-1)^n p_{\#}\left((\partial[G]) \llcorner q\right) \quad \textit{and} \quad \partial[G^{\downarrow}] = (-1)^{n+1} p_{\#}((\partial[G]) \llcorner dq).$$

Proof. Letting $\Psi(y) = y$ for $y \in \mathbb{R}$ in the preceding proposition we deduce the first equation; the second equation is an immediate consequence of the first. ◻

PROPOSITION 6.2. *Suppose $G \in \mathcal{G}(\Omega)$ and $\partial[G]$ is representable by integration. Then*

$$||\partial[G^{\downarrow}]||(B) \leq \int_0^{\infty} ||\partial[\{x : (x, y) \in G\}]||(B)\, d\mathcal{L}^1 y \quad \textit{for any Borel subset } B \textit{ of } \Omega.$$

Proof. Suppose U is an open subset of Ω, $\omega \in \mathcal{D}^{n-1}(\Omega)$, $\mathbf{spt}\, \omega \subset U$, and $|\omega| \leq 1$. For each $y \in (0, \infty)$ let $i_y(x) = (x, y)$ for $x \in \Omega$. From [FE, sect. 4.3.8] we have

$$\langle [G], q, y \rangle = i_{y\#}[\{x : (x, y) \in G\}] \quad \text{for } \mathcal{L}^1 \text{ almost all } y.$$

From Corollary 6.1, (2.8), and (2.7) we find that

$$\begin{aligned} (-1)^{n+1} \partial[G^{\downarrow}](\omega)| &= ((\partial[G]) \llcorner dq)(p^{\#}\omega) \\ &= \int_0^{\infty} \langle \partial[G], q, y \rangle (p^{\#}\omega)\, d\mathcal{L}^1 y \\ &= -\int_0^{\infty} \partial[\{x : (x, y) \in G\}](\omega)\, d\mathcal{L}^1 y \\ &\leq \int_0^{\infty} ||\partial[\{x : (x, y) \in G\}]||(U)\, d\mathcal{L}^1 y, \end{aligned}$$

from which the inequality to be proved immediately follows. ◻

6.6. Proof of Theorems 1.3 and 1.4. We now assume $F : \mathcal{F}(\Omega) \to \mathbb{R}$, F is local, and F is convex. In order to prove the fundamental theorems, Theorems 1.3 and 1.4, we will use F to define a functional $F^{\uparrow}$ on subsets of $\Omega \times \mathbb{R}$, which will be very useful in analyzing $\mathbf{n}_{\epsilon}^{loc}(F)$. This is one of the main new ideas of the paper.

We leave to the reader the elementary proof of the following proposition.

PROPOSITION 6.3. *Suppose $G \in \mathcal{G}(\Omega)$. Then*

$$(0, \infty) \ni y \mapsto U_y(\{x : (x, y) \in G\}) \textit{ is } \mathcal{L}^1 \textit{ summable.}$$

DEFINITION 6.4. *Let*

$$F^{\uparrow} : \mathcal{G}(\Omega) \to \mathbb{R}$$

be such that

$$F^{\uparrow}(G) = F(0) + \int_0^{\infty} U_y(\{x : (x, y) \in G\})\, d\mathcal{L}^1 y \quad \textit{whenever } G \in \mathcal{G}(\Omega).$$

We have a useful comparison principle.

THEOREM 6.1. *We have*

$$F(G^{\downarrow}) \leq F^{\uparrow}(G) \quad \textit{whenever } G \in \mathcal{G}(\Omega).$$

Proof. As we shall see, the theorem will follow rather directly from the following lemma.

LEMMA 6.2. *Suppose $a \in \Omega$ and $E \in \mathcal{M}((0,\infty))$. Then*

$$\kappa(a, \mathcal{L}^1(E)) \leq \int_E u(a,y)\, d\mathcal{L}^1 y.$$

Proof. Suppose $\phi \in \mathcal{D}((0,\infty))$ and $0 \leq \phi \leq 1$. Let $\Phi \in \mathcal{E}((0,\infty))$ be such that $\Phi' = \phi$ and $\lim_{y \downarrow 0} \Phi(y) = 0$. Then

$$0 \leq \Phi(y) \leq y \quad \text{if } 0 < y < \infty. \tag{6.4}$$

Thus, as $(0,\infty) \ni y \mapsto \kappa(a,y)$ is absolutely continuous and $(0,\infty) \ni y \mapsto u(a,y)$ is nondecreasing, we have

$$\kappa(a, \Phi(y)) = \int_0^y u(a, \Phi(y))\phi(y)\, d\mathcal{L}^1 y \leq \int_0^y u(a,y)\phi(y)\, d\mathcal{L}^1 y \quad \text{for } 0 < y < \infty.$$

We complete the proof by letting ϕ approximate the indicator function of E. □

From the lemma we infer that

$$\kappa(x, G^{\downarrow}(x)) \leq \int_{\{y:(x,y)\in G\}} u(x,y)\, d\mathcal{L}^1 y \quad \text{for } \mathcal{L}^n \text{ almost all } x \in \Omega.$$

Integrating this inequality over Ω we use (iv) and (v) of Proposition 1.5 to obtain

$$F(G^{\downarrow}) - F(0) \leq \int_0^{\infty} U_y(\{x : (x,y) \in G\})\, d\mathcal{L}^1 y = F^{\uparrow}(G) - F(0),$$

as desired. □

6.7. Proof of Theorem 1.3. We may assume without loss of generality that $F = \hat{F}$. For each $y \in (0,\infty)$ we let $D_y = \{f > y\}$.

Suppose $0 < b < \infty$, K is a compact subset of Ω, and $E \in \mathbf{k}(D_b, K)$. We need to show that

$$\epsilon||\partial[D_b]||(K) + U_b(D_b) \leq \epsilon||\partial[E]||(K) + U_b(E). \tag{6.5}$$

Let $u(x) = \mathbf{dist}\,(x,K)$ for $x \in \Omega$ and let R be the supremum of the set of $r \in (0,\infty)$ such that $\{v \leq r\} \subset \Omega$. For each $(y,r) \in (0,\infty) \times (0,R)$ let

$$\begin{aligned} C_{y,r} &= (E \cap \{v \leq r\}) \cup (D_y \cap \{v > r\}) \in \mathcal{M}(\Omega), \\ a(y,r) &= \epsilon||\partial[D_y]||(\{v \leq r\}) + U_y(D_y), \\ b(y,r) &= \epsilon||\partial[C_{r,y}]||(\{v \leq r\}) + U_y(C_{r,y}). \end{aligned}$$

Let

$$W = \{(y,r) \in (0,\infty) \times (0,R) : a(y,r) \leq b(y,r)\}.$$

LEMMA 6.3. *For $\mathcal{L}^1$ almost all $y \in (0,\infty)$ we have*

$$a(y,r) \leq b(y,r)\} \quad \textit{for } \mathcal{L}^1 \textit{ almost all } r \in (0,R).$$

Proof. Suppose $r \in (0, R)$, B is a Borel subset of $(0, \infty)$, and

$$G = \{(x, y) \in \Omega \times ((0, \infty) \sim B) : x \in D_y\} \cup \{(x, y) \in \Omega \times B : x \in C_{y,r}\}.$$

Evidently, $G^{\downarrow}(x) = f(x)$ for $\mathcal{L}^n$ almost all $x \in \{v > r\}$, from which it follows that

$$\epsilon ||\partial[f]||(\{v \le r\}) + F(f) \le \epsilon ||\partial[G^{\downarrow}]||(\{v \le r\}) + F(G^{\downarrow}).$$

Let

$$P = \int_{(0,\infty)\sim B} ||\partial[D_y]||(\{v \le r\})\, d\mathcal{L}^1 y \quad \text{and} \quad Q = \int_{(0,\infty)\sim B} U_y(D_y)\, d\mathcal{L}^1 y.$$

We have

$$||\partial[f]||(\{v \le r\}) = P + \int_B ||\partial[D_y||(\{v \le r\})\, d\mathcal{L}^1 y \quad \text{and} \quad F(f) = Q + \int_B U_y(D_y)\, d\mathcal{L}^1 y.$$

From Corollary 6.1 and Proposition 6.2 we obtain

$$\begin{aligned} ||\partial[G^{\downarrow}]||(\{v \le r\}) &\le ||\partial[G] \llcorner dq||(\{v \le r\} \times (0, \infty)) \\ &= \int ||\partial[\{x : (x, y) \in G\}]||(\{v \le r\})\, d\mathcal{L}^1 y \\ &= P + \int_B ||\partial[C_{y,r}]||(\{v \le r\}). \end{aligned}$$

From (6.1) we obtain

$$F(G^{\downarrow}) \le F^{\uparrow}(G) = Q + \int_B U_y(C_{r,y})\, d\mathcal{L}^1 y,$$

which implies

$$\int_B a(y, r)\, d\mathcal{L}^1 y \le \int_B b(y, r)\, d\mathcal{L}^1 y.$$

Owing to the arbitrariness of B we infer that $a(y, r) \le b(y, r)$ for $\mathcal{L}^1$ almost all $y \in (0, \infty)$, so the lemma follows from Tonelli's theorem. $\square$

We have $(D_y \sim D_b) \cup (D_b \sim D_y) = \{b < f \le y\}$ whenever $b < y < \infty$, so that

$$\lim_{y \downarrow b} \Sigma_\Omega(D_y, D_b) = 0. \tag{6.6}$$

This implies that

$$\lim_{y \downarrow b} \Sigma_{\Omega \sim K}(D_y, E) = \lim_{y \downarrow b} \Sigma_{\Omega \sim K}(D_y, D_b) = 0. \tag{6.7}$$

By Proposition 1.5(iv) we have

$$\begin{aligned} |U_y(D_y) - U_b(D_b)| &\le |U_y(D_y) - U_y(D_b)| + |U_b(D_y) - U_b(D_b)| \\ &\le \mathbf{l}(F, Y)\Sigma_\Omega(D_y, D_b) + |U_b(D_y) - U_b(D_b)| \end{aligned}$$

as well as

$$|U_y(C_{y,r}) - U_y(E)| \le \mathbf{l}(F, Y)\Sigma_{\{u > r\}}(D_y, E) \le \mathbf{l}(F, Y)\Sigma_{\Omega \sim K}(D_y, D_b) \tag{6.8}$$

whenever $0 < y < Y < \infty$. With the help of (6.6) and Proposition 1.6 we infer that

$$\lim_{y \downarrow b} U_y(D_y) = U_b(D_b). \tag{6.9}$$

Suppose $0 < r < R$. Since (6.6) and (2.1) imply that

$$||\partial[D_b]||(K) \le \liminf_{y \downarrow b} ||\partial[D_y]||(\{u \le \rho\}) \quad \text{for } 0 < \rho < R,$$

we infer from (6.9) that

$$r\,(||\partial[D_b]||(K) + U_b(D_b)) \le \liminf_{y \downarrow b} \int_0^r a(y,\rho)\, d\mathcal{L}^1\rho. \tag{6.10}$$

Applying (2.10), with f there equal to 1_{D_y} and g there equal to 1_E, and using (6.6), (6.7), (6.8), and (6.9), we find that

(6.11)
$$\int_0^r b(y,\rho)\, d\mathcal{L}^1\rho \le \Sigma_{\{u>r\}}(D_y, E) + \int_0^r ||\partial[E]||(\{u \le \rho\})\, d\mathcal{L}^1\rho + \int_0^r U_y(C_{y,\rho})\, d\mathcal{L}^1\rho$$
$$\to \int_0^r ||\partial[E]||(\{u \le \rho\}) + U_b(E)\, d\mathcal{L}^1\rho \quad \text{as } y \downarrow b.$$

Using Lemma 6.2 and Tonelli's theorem we may choose a sequence y in (b,∞) with limit b such that

$$\mathcal{L}^1(\{r \in (0,R) : (y_\nu, r) \notin W\}) = 0 \quad \text{for } \nu = 1,2,3,\ldots.$$

Thus

$$\int_0^r a(y_\nu,\rho)\, d\mathcal{L}^1\rho \le \int_0^r b(y_\nu,\rho)\, d\mathcal{L}^1\rho,$$

so (6.10) and (6.11) imply

$$r\,(||\partial[D_b]||(K) + U_b(D_b)) \le \int_0^r ||\partial[E]||(\{u \le \rho\}) + U_b(E)\, d\mathcal{L}^1\rho;$$

dividing by r and letting $r \downarrow 0$ we obtain (6.5).

We leave it to the reader to modify the proof just given in a straightforward way to show that $\{f \ge b\} \in \mathbf{n}_\epsilon^{loc}(L_b)$.

6.8. Proof of Theorem 1.4. Let K be a compact subset of Ω and let $g \in \mathcal{F}(\Omega)$ such that $\mathbf{spt}\,[G^\downarrow - g] \subset K$.

Suppose $y \in (0,\infty)$. Since $G^\downarrow(x) = g(x)$ for $\mathcal{L}^n$ almost all $x \in \Omega \sim K$ we find that

$$\mathbf{spt}\,[\{G^\downarrow > y\}] - [\{g > y\}] \subset K,$$

so that if $\{x : (x,y) \in G\} \in \mathbf{n}_\epsilon^{loc}(U_y)$, we have

$$||\partial[\{G^\downarrow > y\}]||(K) + U_y(\{G^\downarrow > y\}) \le ||\partial[\{g > y\}]||(K) + U_y(\{g > y\}).$$

Integrating over $y \in (0, \infty)$ with respect to $\mathcal{L}^1$ and using Proposition 6.2, Theorem 6.1, (2.6), and Proposition 1.5(v), we find that

$$
\begin{aligned}
&||\partial[G^{\downarrow}]||(K) + F(G^{\downarrow}) \\
&\qquad \leq \int_0^\infty ||\partial[\{G^{\downarrow} > y\}]||(K)\, d\mathcal{L}^1 y + F^{\uparrow}(G) \\
&\qquad = \int_0^\infty ||\partial[\{G^{\downarrow} > y\}]||(K) + U_y(\{G^{\downarrow} > y\})\, d\mathcal{L}^1 y \\
&\qquad \leq \int_0^\infty ||\partial[\{g > y\}]||(K) + U_y(\{g > y\})\, d\mathcal{L}^1 y \\
&\qquad = ||\partial[g]||(K) + F(g).
\end{aligned}
$$

It remains to deal with (1.6). For each $E \in \mathcal{M}(\Omega)$ let $C(E)$ be the set of $y \in (0, \infty)$ such that $L_y(E) \neq U_y(E)$. Since $(0, \infty) \ni \mapsto F(y 1_E)$ is convex we find that $C(E)$ is countable. Now choose a countable subfamily $\mathcal{E}$ of $\mathcal{M}(\Omega)$ which is dense with respect to the pseudometric $\Sigma_\Omega(\cdot, \cdot)$. By a straightforward approximation argument, which we leave to the reader, we find that $L_y(D) = U_y(D)$ whenever $D \in \mathcal{M}(\Omega)$ and $y \notin \cup\{C(E) : E \in \mathcal{E}\}$.

7. Proof of Theorem 1.6. Theorem 1.6 will be proved by calculating the appropriate first and second variations, invoking the regularity theorem for $\mathcal{C}_\lambda(\Omega)$, and then utilizing higher regularity results for the minimal surface equation.

For each $x \in \mathbf{b}(D)$ we let $P(x)$ equal the orthogonal projection of $\mathbb{R}^n$ onto $\{v \in \mathbb{R}^n : v \bullet \mathbf{n}_D(x) = 0\}$.

We may assume without loss of generality that $U = \Omega$. It follows from Proposition 1.2 and Theorem 1.2 that $\Sigma_\Omega(D, \Gamma) = 0$, so $[D] = [\Gamma]$.

Part 1. Suppose $a \in M$. From Proposition 1.2 and Theorem 1.2 there are Ψ, V, r, g such that Ψ carries $\mathbb{R}^{n-1} \times \mathbb{R}$ isometrically onto $\mathbb{R}^n$, $\Psi(0,0) = a$, V is an open subset of $\mathbb{R}^{n-1}$, $0 \in V$, $0 < r < \infty$, $g : V \to (-r, r)$ is of class $C^{1,\mu}$, $\Psi[V \times (-r, r)] \subset \Omega$, and

$$\Gamma \cap \Psi[V \times (-r, r)] = \Psi[\{(v, w) \in V \times (-r, r) : w < g(v)\}].$$

Suppose $\phi \in \mathcal{D}(V)$. Choose an open interval I such that $0 \in I$ and $g(v) + t\phi(v) \in (-r, r)$ whenever $(v, t) \in V \times I$. For each $t \in I$ let

$$\Phi(t) = \epsilon \int_V \sqrt{1 + |\nabla(g + t\phi)|^2}\, d\mathcal{L}^{n-1} + \int_V \left(\int_{-r}^{(g+t\phi)(v)} \zeta(\Psi(v, w))\, d\mathcal{L}^1 w \right) d\mathcal{L}^{n-1} v.$$

Then $\Phi(0) \leq \Phi(t)$ whenever $t \in I$ since $D \in \mathbf{n}_\epsilon^{loc}(Z)$. Thus

$$0 = \Phi'(0) = \epsilon \int_V \frac{\nabla g \bullet \nabla \phi}{\sqrt{1 + |\nabla g|^2}}\, d\mathcal{L}^n + \int_V \zeta(\Psi(v, g(v)))\phi(v)\, d\mathcal{L}^n.$$

That is, g is a weak solution of

$$-\epsilon \operatorname{div} J^{-1} \nabla g + \zeta \circ \Psi \circ G = 0,$$

where we have set $J = \sqrt{1 + |\nabla g|^2}$ and $G(v) = (v, g(v))$ for $v \in V$.

Inasmuch as ∂g is Hölder continuous, standard results on regularity of weak solutions of elliptic equations, as found, for example, in [GT, sect. 8.3], imply that g is of class $C^{k+2,\mu}$. Since a is an arbitrary point of M we conclude that M is of class $C^{k+2,\mu}$, so M has a second fundamental form. Since $H(a) \bullet \mathbf{n}_\Gamma(a) = -\operatorname{div} J^{-1}\nabla g(0)$ we find that (1.7) holds.

Part 2. We now suppose ζ is continuously differentiable. Let Π, Q, H be as in section 4.

Since M is of class C^2 by Part 1 there is a map $N : \Omega \to \mathbb{R}^n$ of class C^1 such that $N|M = \mathbf{n}_\Gamma|M$.

Suppose $\phi \in \mathcal{D}(\Omega)$. Let $X = \phi N$, and let K; I; h_t, $t \in I$; E_t, $t \in I$; P; and a_1 and a_2 be as in section 3. Let Y be as in Proposition 3.3. Since $\partial X = (\partial\phi)N + \phi(\partial N)$ we find that

$$
\begin{aligned}
l_1 &= P \circ \partial(\phi N) \circ P = \phi\,(P \circ \partial N \circ P),\\
l_2 &= P^\perp \circ \partial(\phi N) \circ P = ((\partial\phi) \circ P)\,N,\\
\operatorname{trace} l_1 &= \phi(H \bullet N),\\
\operatorname{trace} l_2^* \circ l_2 &= |(\partial\phi) \circ P|^2,\\
\operatorname{trace} l_1 \circ l_1 &= \phi^2 Q.
\end{aligned} \tag{7.1}
$$

For each $t \in I$ let

$$\Phi(t) = \epsilon||\partial[E_t]||(K) + Z(E_t).$$

Let A and B be as in Propositions 3.1 and 3.3, respectively, so $\Phi(t) = \epsilon A(t) + B(t)$ for $t \in I$. Since $\Phi(0) \le \Phi(t)$ for $t \in I$ we have

$$0 \le \epsilon A''(0) + B''(0). \tag{7.2}$$

We have

$$
\begin{aligned}
a_2 &= (\operatorname{trace} l_1)^2 + \operatorname{trace}(l_2^* \circ l_2 - l_1 \circ l_1)\\
&= \phi^2(H \bullet N)^2 + |\partial\phi \circ P|^2 - \phi^2 Q^2\\
&= \phi^2 \frac{\zeta^2}{\epsilon^2} + |\partial\phi \circ P|^2 - \phi^2 Q^2.
\end{aligned}
$$

Making use of (1.7) we obtain

$$
\begin{aligned}
&\big(\zeta Y + (\nabla\zeta \bullet X)X\big) \bullet N\\
&\quad = (\zeta(\phi(H \bullet N)\phi N) - \nabla\zeta \bullet (\phi N)\phi N) \bullet N\\
&\quad = -\frac{\zeta^2}{\epsilon}\phi^2 + \phi^2(\nabla\zeta \bullet N).
\end{aligned}
$$

So (1.8) now follows from (7.2) and Propositions 3.1 and 3.3.

8. The denoising case revisited. Suppose

(i) s, γ, and F are as in section 1.8;

(ii) γ is convex and β is as in section 1.8;

(iii) U is an open subset of Ω, $z \in \mathbb{R}$, and

$$s(x) = z \quad \text{for } x \in U;$$

(iv) $0 < y < \infty$ and β is continuously differentiable near $y - z$;

(v) $0 < \epsilon < \infty$ and $f \in \mathbf{m}_\epsilon^{loc}(F)$;

(vi) Γ is the intersection of U with the interior of the support of $[\{f > y\}]$, and M is the intersection of U with the boundary of Γ;

(vii) H is the mean curvature vector of M, and Q is the square of the length of the second fundamental form of M.

From Theorems 1.3 and 1.6 we find that $[\Gamma] = [U \cap \{f > y\}]$, that

$$H(x) = -\frac{1}{\epsilon}\beta(y-z)\mathbf{n}_\Gamma(x) \quad \text{whenever } x \in M,$$

and that

$$\int_M |\nabla_M \phi(x)|^2 - \phi(x)^2 Q(x)\, d\mathcal{H}^{n-1}x \geq 0$$

for any $\phi \in \mathcal{D}(\Omega)$, where, for each $x \in M$, $\nabla_M \phi(x)$ is the orthogonal projection of $\nabla\phi(x)$ on $\mathbf{Tan}(M, x)$.

Now suppose $n = 2$, let $a \in M$, and let A be the connected component of a in M. If $\beta(y-z) = 0$, then A is a subset of a straight line. Suppose $\beta(y-z) \neq 0$ and let

$$R = \frac{\epsilon}{|\beta(y-z)|}.$$

Then A is an arc of a circle of radius R. Let c be the center of this circle. Then for each $a \in A$, there is an open subset G of U containing a such that

$$\Gamma \cap G = \begin{cases} \mathbf{U}^2(c, R) \cap G & \text{if } \beta(y-z) < 0, \\ (\mathbb{R}^n \sim \mathbf{U}^2(c, R)) \cap G & \text{if } \beta(y-z) > 0. \end{cases}$$

Finally, let L be the length of A. Since $Q(x) = 1/R^2$ for $x \in M$ we find that

$$\int_0^L \phi'(\sigma)^2 - \frac{1}{R^2}\phi(\sigma)^2\, d\mathcal{L}^1\sigma \geq 0$$

for all continuously differentiable $\phi : [0, L] \to \mathbb{R}$ which are differentiable on $(0, L)$ and which vanish at 0 and L. Letting $\phi(\sigma) = \sin(\pi\sigma/L)$ for $\sigma \in [0, L]$ we infer that

$$L \leq \pi R.$$

9. Some results for functionals on sets.

9.1. Proof of Theorem 1.7. We begin with a simple lemma.

LEMMA 9.1. *Suppose A is a nested sequence in $\mathbf{n}_\epsilon^{loc}(N_S)$. Then $\cap_{\nu=1}^\infty A_\nu \in \mathbf{n}_\epsilon^{loc}(N_S)$ and, provided $\mathcal{L}^n(\cup_{\nu=1}^\infty A_\nu) < \infty$, $\cup_{\nu=1}^\infty \in \mathbf{n}_\epsilon^{loc}(N_S)$.*

We leave to the reader the straightforward proof making use of (2.1) and cutoff arguments like those used in the proof of Theorem 5.2.

Let

$$F(f) = \int_\Omega |f - 1_S|\, d\mathcal{L}^n \quad \text{for } f \in \mathcal{F}(\Omega).$$

For each $y \in \mathbb{R}$ let U_y be as in Theorem 1.5. Recall from section 1.9 that

$$U_y = \begin{cases} 0 & \text{if } 1 \leq y < \infty, \\ \widehat{N_S} & \text{if } 0 < y < 1. \end{cases}$$

Suppose $A, B \in \mathcal{A}$ and $0 < a < b < c < 1$. Let

$$G = (A \times (0,b)) \cup (B \times (b,1)) \in \mathcal{G}(\Omega).$$

Then $\{x : (x,y) \in G\} \in \mathbf{n}_\epsilon^{loc}(U_y)$ whenever $0 < y < \infty$, so $G^\downarrow \in \mathbf{m}_\epsilon^{loc}(F)$ by Theorem 1.4. From Theorem 1.3 we infer that $A \cup B = \{G^\downarrow > a\} \in \mathbf{n}_\epsilon^{loc}(U_a)$ and $A \cap B = \{G^\downarrow > c\} \in \mathbf{n}_\epsilon^{loc}(U_c)$, so $A \cup B$ and $A \cap B$ belong to $\mathbf{n}_\epsilon^{loc}(N_S)$.

It follows that if $\mathcal{F}$ is a finite subfamily of $\mathcal{A}$, then $\cup\mathcal{F}$ and $\cap\mathcal{F}$ belong to $\mathbf{n}_\epsilon^{loc}(N_S)$.

Let B be a sequence in $\mathcal{A}$ such that

$$\lim_{\nu\to\infty} \mathcal{L}^n(B_\nu) = \inf\{\mathcal{L}^n(A) : A \in \mathcal{A}\}.$$

Since each of $\cap_{\nu=1}^N B_\nu$ belongs to $\mathbf{n}_\epsilon^{loc}(N_S)$ we infer from the preceding lemma that $C = \cap_{\nu=1}^\infty B_\nu \in \mathbf{n}_\epsilon^{loc}(N_S)$. It is clear that $\mathcal{L}^n(C \sim \cap\mathcal{A}) = 0$, so $\cap\mathcal{A} \in \mathbf{n}_\epsilon^{loc}(N_S)$.

Let us now assume $\mathcal{L}^n(\cup\mathcal{A}) < \infty$. Let D be a sequence in $\mathcal{A}$ such that

$$\lim_{\nu\to\infty} \mathcal{L}^n(D_\nu) = \sup\{\mathcal{L}^n(A) : A \in \mathcal{A}\}.$$

Since each of $\cup_{\nu=1}^N D_\nu$ belongs to $\mathbf{n}_\epsilon^{loc}(N_S)$ we infer from the preceding lemma that $E = \cup_{\nu=1}^\infty D_\nu \in \mathbf{n}_\epsilon^{loc}(N_S)$. It is clear that $\mathcal{L}^n(\cup\mathcal{A} \sim E) = 0$, so $\cup\mathcal{A} \in \mathbf{n}_\epsilon^{loc}(N_S)$.

9.2. A comparison principle. The following proposition and its proof were suggested by a similar result found in [CA1] in a different context.

PROPOSITION 9.1. *Suppose $M, N \in \mathbf{M}(\Omega)$, M and N are local, $0 < \epsilon < \infty$, $D \in \mathbf{n}_\epsilon^{loc}(M)$, $E \in \mathbf{n}_\epsilon^{loc}(N)$, and $\mathbf{spt}\,[D \cup E]$ is compact. Then*

$$\hat{N}(E \sim D) \le \hat{M}(E \sim D).$$

In particular, if

$$\hat{M}(G) < \hat{N}(G) \quad \textit{whenever } G \in \mathcal{M}(\Omega) \textit{ and } \mathcal{L}^n(G) > 0,$$

then

$$\mathcal{L}^n(E \sim D) = 0.$$

Proof. Without loss of generality we may assume $M = \hat{M}$ and $N = \hat{N}$. Since $\mathbf{spt}\,[D] \cup \mathbf{spt}\,[E] \subset \mathbf{spt}\,[D \cup E]$ we have

$$\epsilon\mathbf{M}(\partial[D]) + M(D) \le \epsilon\mathbf{M}(\partial[D \cup E]) + M(D \cup E)$$

and

$$\epsilon\mathbf{M}(\partial[E]) + N(E) \le \epsilon\mathbf{M}(\partial[D \cap E]) + N(D \cap E).$$

Also,

$$\begin{aligned}
&\mathbf{M}(\partial[D \cup E]) + \mathbf{M}(\partial[D \cap E]) \\
&\quad = \int_0^1 \mathbf{M}(\partial[\{1_D + 1_E > y\}])\, d\mathcal{L}^1 y + \int_1^2 \mathbf{M}(\partial[\{1_D + 1_E > y\}])\, d\mathcal{L}^1 y \\
&\quad = \mathbf{M}(\partial[1_D + 1_E]) \\
&\quad \le \mathbf{M}(\partial[D]) + \mathbf{M}(\partial[E]).
\end{aligned}$$

Since M and N are local it follows that

$$
\begin{aligned}
&\epsilon(\mathbf{M}(\partial[D]) + \mathbf{M}(\partial[E])) + M(D \sim E) + M(D \cap E) + N(E \sim D) + N(E \cap D)\\
&= \epsilon(\mathbf{M}(\partial[D]) + \mathbf{M}(\partial[E])) + M(D) + N(E)\\
&\leq \epsilon(\mathbf{M}(\partial[D \cup E]) + \mathbf{M}(\partial[D \cap E])) + M(D \cup E) + N(D \cap E)\\
&\leq \epsilon(\mathbf{M}(\partial[D]) + \mathbf{M}(\partial[E])) + M(D \cup E) + N(D \cap E)\\
&= \epsilon(\mathbf{M}(\partial[D]) + \mathbf{M}(\partial[E])) + M(D \sim E) + M(D \cap E) + M(E \sim D) + N(E \cap D). \qquad \square
\end{aligned}
$$

9.3. Proof of Theorem 1.5. Suppose $0 < y < z < \infty$. Since F is strictly convex we have

$$\beta(x, y) < \beta(x, z) \quad \text{for } \mathcal{L}^n \text{ almost all } x \in \Omega.$$

This in turn implies that $U_y(D) < U_z(D)$ whenever $D \in \mathcal{M}(\Omega)$. Applying Proposition 9.1 with M, N there equal to U_y, U_z and D, E there equal to $\{f > y\}, \{x : (x, z) \in G\}$ and $\{x : (x, y) \in G\}, \{f > z\}$, respectively, we infer that

$$\text{(i)} \qquad \mathcal{L}^n(\{x : (x, z) \in G\} \sim \{f > y\}) = 0$$

and

$$\text{(ii)} \qquad \mathcal{L}^n(\{f > z\} \sim \{x : (x, y) \in G\}) = 0.$$

Suppose $0 < w < \infty$. Letting $z = w$ and $y \uparrow w$ in (i) we find that

$$\mathcal{L}^n(\{x : (x, w) \in G\} \sim \{f \geq w\}) = 0.$$

Letting $y = w$ and $z \downarrow w$ in (ii) we find that

$$\mathcal{L}^n(\{f > w\} \sim \{x : (x, w) \in G\}) = 0.$$

Since $\mathcal{L}^n(\{f = w\}) = 0$ for all but countably many $w \in (0, \infty)$ we may use Tonelli's theorem to complete the proof.

10. Two useful theorems in the denoising case. We suppose throughout this subsection that $\gamma : \mathbb{R} \to \mathbb{R}$, γ is locally Lipschitzian, γ is decreasing on $(-\infty, 0)$, and γ is increasing on $(0, \infty)$. We let

$$F(f) = \int_\Omega \gamma(f(x) - s(x))\, d\mathcal{L}^n x \quad \text{whenever } f \in \mathcal{F}(\Omega).$$

10.1. A simple maximum principle.

Proposition 10.1. *Suppose* $0 < \epsilon < \infty$, $f \in \mathbf{m}_\epsilon^{loc}(F)$, *and*

$$u = \inf\{||1_{\Omega \sim K} f||_{\mathbf{L}_\infty(\Omega)} : K \textit{ is a compact subset of } \Omega\}.$$

Then $||f||_{\mathbf{L}_\infty(\Omega)} \leq u \vee ||s||_{\mathbf{L}_\infty(\Omega)}$.

Remark 10.1. It follows from Corollary 5.1 that $u = 0$ if $\Omega = \mathbb{R}^n$.

Proof. Suppose $u \vee ||s||_{\mathbf{L}_\infty(\Omega)} < M < \infty$. Then $K = \mathbf{spt}\,[f - f \wedge M]$ is a compact subset of Ω, so

$$
\begin{aligned}
&\int_{\{f > M\}} \gamma(f(x) - s(x)) - \gamma(M - s(x))\, d\mathcal{L}^n x\\
&\qquad = F(f) - F(f \wedge M)\\
&\qquad \leq \epsilon(||\partial[f \wedge M]||(K) - ||\partial[f]||(K))\\
&\qquad = -\int_M^\infty ||\partial[\{f > y\}]||(K)\, d\mathcal{L}^1 y\\
&\qquad \leq 0.
\end{aligned}
$$

If $f(x) > M > s(x)$, then $f(x) - s(x) > M - s(x) > 0$, so that $\gamma(f(x) - s(x)) - \gamma(M - s(x)) > 0$. Owing to the arbitrariness of M we find that the proposition holds. $\square$

THEOREM 10.1. *Suppose $\Omega = \mathbb{R}^n$, $0 < \epsilon < \infty$, $f \in \mathbf{m}_\epsilon^{loc}(F)$, and, for each $y \in (0, \infty)$,*

$$C(y) \text{ equals the closed convex hull of } \mathbf{spt}\,[\{s > y\}].$$

Then

$$\mathbf{spt}\,[\{f > y\}] \subset C(y) \quad \text{whenever } 0 < y < \infty.$$

Proof. Suppose $0 < b < \infty$. Let $g_b = f1_{\{f<b\}} + b1_{\{f \geq b\} \sim C(b)} + f1_{\{f \geq b\} \cap C(b)}$ and note that

$$\{g_b > y\} = \begin{cases} \{f > y\} & \text{if } y \leq b, \\ \{f > y\} \cap C(b) & \text{if } y > b \end{cases} \quad \text{whenever } y \in \mathbb{R}.$$

It follows from (2.12) that $\mathbf{M}(\partial[\{g_b > y\}]) \leq \mathbf{M}(\partial[\{f > y\}])$ whenever $y \in \mathbb{R}$.

Let $K_b = \mathbf{spt}\,[f - g_b]$. Since $\{f - g_b \neq 0\} \subset \{f > b\}$ we infer from Theorems 1.1 and 5.4(v) that K_b is compact. Since $f \in \mathbf{m}_\epsilon^{loc}(F)$ we infer with the help of (5.1) that

$$\begin{aligned}
\int_{\{f>b\} \sim C(b)} & \gamma(f(x) - s(x)) - \gamma(b - s(x))\, d\mathcal{L}^n x \\
&= F(f) - F(g_b) \\
&\leq \epsilon(||\partial[g_b]||(K_b) - ||\partial[f]||(K_b)) \\
&= \epsilon \int_b^\infty ||\partial[\{g_b > y\}]||(K_b) - ||\partial[\{f > y\}]||(K_b)\, d\mathcal{L}^1 y \\
&\leq 0,
\end{aligned}$$

which implies $\mathcal{L}^n(\{f > b\} \sim C(b)) = 0$. $\square$

10.1.1. Convex containment.

PROPOSITION 10.2. *Suppose $M \in \mathbf{M}(\mathbb{R}^n)$, M is local, C is a closed convex subset of $\mathbb{R}^n$, and*

$$M(E) \geq M(\emptyset) \quad \text{whenever } E \in \mathcal{M}(\mathbb{R}^n) \text{ and } \quad \mathcal{L}^n(E \cap C) = 0. \tag{10.1}$$

Then $\mathbf{spt}\,[D]$ is a compact subset of C whenever $D \in \mathbf{n}_\epsilon^{loc}(M)$.

Remark 10.2. Evidently, (10.1) is equivalent to the statement that $\mu(x) \geq 0$ for $\mathcal{L}^n$ almost all $x \in \mathbb{R}^n \sim C$, where μ is as in Proposition 1.3.

Proof. Suppose $D \in \mathbf{n}_\epsilon^{loc}(M)$. It follows from Proposition 1.2 and Theorem 5.4(iv) that $\mathbf{spt}\,[D]$ is compact. From (2.12) we find that

$$\mathbf{M}(\partial[C \cap D]) \leq \mathbf{M}(\partial[D]).$$

Moreover, as M is local and $D \in \mathbf{n}_\epsilon^{loc}(M)$,

$$\epsilon(\mathbf{M}(\partial[D]) - \mathbf{M}(\partial[D \cap C])) \leq M(D \cap C) - M(D) = M(\emptyset) - M(D \sim C) \leq 0.$$

Thus $\mathbf{M}(\partial[C \cap D]) = \mathbf{M}(\partial[D])$, so the theorem now follows from (2.12). $\square$

11. Some examples. Let

$$S = [-1,1] \times [-1,1] \in \mathcal{M}(\mathbb{R}^2),$$

suppose $1 \le p < \infty$, and let

$$F(g) = \frac{1}{p} \int |g - 1_S|^p \, d\mathcal{L}^2 \quad \text{whenever } g \in \mathcal{M}(\mathbb{R}^2).$$

We will determine $\mathbf{m}_\epsilon^{loc}(F)$, $0 < \epsilon < \infty$.

11.1. The sets $A_{i,r}$. For each $r \in (0,1]$ let

$$A_{0,r} = \{(1-r, 1-r) + r\{(\cos\theta, \sin\theta) : 0 \le \theta \le \pi/2\},$$

let $A_{i,r}$, $i = 1,2,3$, be a counterclockwise rotation about the origin of $A_{0,r}$ by $i\pi/2$, and let

$$C(r)$$

be the convex hull of $\cup_{i=0}^3 A_{i,r}$.

THEOREM 11.1. *Suppose* $0 < \epsilon < \infty$ *and*

$$T = \{[f] : f \in \mathbf{m}_\epsilon^{loc}(F)\}.$$

If $(1 + \sqrt{\pi}/2)\epsilon > 1$, *then*

$$T = \{0\}.$$

If $(1 + \sqrt{\pi}/2)\epsilon = 1$ *and* $p = 1$, *then*

$$T = \{t[1_{C(\epsilon)}] : 0 \le t \le 1\}.$$

If $(1 + \sqrt{\pi}/2)\epsilon < 1$ *and* $p = 1$, *then*

$$T = \{[1_{C(\epsilon)]}\}.$$

If $(1 + \sqrt{\pi}/2)\epsilon = 1$ *and* $p > 1$, *then*

$$T = \{0\}.$$

If $(1 + \sqrt{\pi}/2)\epsilon < 1$ *and* $p > 1$, *then*

$$T = \{[G^\downarrow]\},$$

where

$$Y = 1 - \left(1 + \sqrt{\pi}/2)\epsilon\right)^{1/(p-1)}$$

and

$$G = \left\{(x,y) : 0 < y < Y \text{ and } x \in C\left(\frac{\epsilon}{(1-y)^{p-1}}\right)\right\} \in \mathcal{G}(\mathbb{R}^2).$$

Proof. For each $y \in (0,\infty)$ let

$$Q_y = \{[D] : D \in \mathbf{n}_\epsilon^{loc}(U_y)\},$$

where U_y is as in Theorem 1.5.

Using (1.9) we find that $U_y(E) > 0$ whenever $1 < y < \infty$, $E \in \mathcal{M}(\mathbb{R}^n)$, and $\mathcal{L}^2(E) > 0$; since $U_y(\emptyset) = 0$ we find that

$$Q_y = \{0\} \quad \text{if } 1 < y < \infty.$$

Suppose $0 < y < 1$, let

$$Z = \begin{cases} 1 & \text{if } p = 1, \\ (1-y)^{p-1} & \text{if } p > 1, \end{cases} \quad \text{and let} \quad R = \frac{\epsilon}{Z}.$$

Suppose $R \leq 1$ and let

$$I = (U_y)_\epsilon \,(C(R)) = \epsilon \mathbf{M}(\partial[C(R)]) + U_y(C(R)).$$

We have

$$\epsilon \mathbf{M}(\partial[C(R)]) = \epsilon(4(2-2R) + 2\pi R)$$

and

$$U_y(C(R)) = -Z\mathcal{L}^2(C(R)) = -Z(4 - (4-\pi)R^2),$$

so

$$\begin{aligned} I &= \epsilon(4(2-2R) + 2\pi R) - Z(4 - (4-\pi)R^2) \\ &= \frac{-4Z^2 + 8\epsilon Z + (\pi - 4)\epsilon^2}{Z} \\ &= -4\frac{(Z - (1 + \sqrt{\pi}/2)\epsilon)(Z - (1 - \sqrt{\pi}/2)\epsilon)}{Z}. \end{aligned}$$

Since $R \leq 1$ we have

$$Z = \epsilon/R \geq \epsilon > (1 - \sqrt{\pi}/2)\epsilon.$$

Thus

$$I \begin{cases} < 0 = U_y(\emptyset) & \Leftrightarrow Z > (1 + \sqrt{\pi}/2)\epsilon, \\ = 0 = U_y(\emptyset) & \Leftrightarrow Z = (1 + \sqrt{\pi}/2)\epsilon, \\ > 0 = U_y(\emptyset) & \Leftrightarrow Z < (1 + \sqrt{\pi}/2)\epsilon. \end{cases}$$

Suppose $D \in \mathbf{n}_\epsilon^{loc}(U_y)$, $[D] \neq 0$, and $D = \mathbf{spt}\,[D]$. We claim that

$$R \leq 1 \quad \text{and} \quad D = C(R). \tag{11.1}$$

From Proposition 10.2 we infer that $D \subset S$. Let U equal the interior of S and let $M = U \cap \mathbf{bdry}\, D$. Then $U \cap M \neq \emptyset$ since otherwise we would have $D = S$, in which case M would have corners, which is incompatible with Theorem 1.2. Let A be a connected component of M. We infer from section 8 that A is an arc of a circle of radius R, the length of which does not exceed πR. Because D can have no corners we find that A meets the interior of the boundary of S tangentially. Thus (11.1) holds.

The theorem now follows from Theorems 1.3 and 1.4. $\square$

Acknowledgments. It is a pleasure to thank Kevin Vixie for acquainting me with this area of research. In the course of carrying out this work I profited from conversations with Kevin Vixie and Selim Esedoḡlu and benefited from the support of Los Alamos National Laboratory.

REFERENCES

[Alli] S. S. Alliney, *Digital filters as absolute norm regularizers*, IEEE Trans. Signal Process., 40 (1992), pp. 1548–1562.

[AW1] W. K. Allard, *The first variation of a varifold*, Ann. Math., 95 (1972), pp. 417–491.

[AW2] W. K. Allard, *Total variation regularization for image denoising;* II. *Examples*, in preparation.

[BDG] E. Bombieri, E. DeGiorgi, and E. Giusti, *Minimal cones and the Bernstein problem*, Invent. Math., 7 (1969), pp. 243–268.

[CA1] A. Chambolle, *Total Variation Minimization and a Class Of Binary MRF Models*, preprint CMAP 578, Centre de Mathématiques Appliquées, École Polytechnique, Palaiseau, France; available online from http://www.cmap.polytechnique.fr/preprint/repository/578.pdf; this version varies a bit from that in Energy Minimization Methods in Computer Vision and Pattern Recognition: 5th International Workshop, EMMCVPR 2005, Lecture Notes in Comput. Sci. 3757, Springer-Verlag, Berlin, 2005, pp. 136–152.

[CA2] A. Chambolle, *An algorithm for mean curvature motion*, Interfaces Free Bound., 6 (2004), pp. 195–218.

[CE] T. F. Chan and S. Esedoḡlu, *Aspects of total variation regularized L^1 function approximation*, SIAM J. Appl. Math., 65 (2005), pp. 1817–1837.

[FE] H. Federer, *Geometric Measure Theory*, Grundlehren Math. Wiss. 153, Springer-Verlag, New York, 1969.

[GT] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Grundlehren Math. Wiss. 224, Springer-Verlag, New York, 1977.

[M] U. Massari, *Esistenza e regolaritá delle ipersuperficie di curvatura media assegnata in $\mathbb{R}^n$*, Arch. Ration. Mech. Anal., 55 (1974), pp. 357–382.

[ROF] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.

SIAM J. MATH. ANAL.
Vol. 39, No. 4, pp. 1191–1207

UNIQUENESS IN THE FABER–KRAHN INEQUALITY FOR ROBIN PROBLEMS*

DANIEL DANERS† AND JAMES KENNEDY†

Abstract. We prove uniqueness in the Faber–Krahn inequality for the first eigenvalue of the Laplacian with Robin boundary conditions, asserting that among all sufficiently smooth domains of fixed volume, the ball is the unique minimizer for the first eigenvalue. The method of proof, which avoids the use of any symmetrization, also works in the case of Dirichlet boundary conditions. We also give a characterization of all symmetric elliptic operators in divergence form whose first eigenvalue is minimal.

AMS subject classifications. 35P15, 35J25

Key words. Faber–Krahn inequality, isoperimetric inequality, elliptic partial differential equation, Robin problem, elastically supported membrane, unique minimizer

DOI. 10.1137/060675629

1. Introduction.

In this paper, we prove that the ball is the unique minimizer in the Faber–Krahn inequality for the first eigenvalue of the Robin boundary value problem

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu} + \beta u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^N, N \geq 2$, is a bounded C^2-domain with outward pointing unit normal ν and $\beta > 0$ is a constant. We build on the work of Daners [9] and Bossel [6], who proved a Faber–Krahn inequality for Robin problems, stating that among all Lipschitz domains Ω of fixed measure, the ball minimizes the first eigenvalue $\lambda_1 = \lambda_1(\Omega)$ of (1.1). The main purpose of this paper is to prove that the ball is the *unique* minimizer. More precisely, we will prove the following theorem.

THEOREM 1.1. *Suppose $\Omega \subset \mathbb{R}^N$ $(N \geq 2)$ is a bounded domain of class C^2, $B \subset \mathbb{R}^N$ is a ball with the same measure as Ω, and $\beta \in (0, \infty)$. Then $\lambda_1(B) \leq \lambda_1(\Omega)$ with equality if and only if $\Omega = B$ up to translation.*

If $\beta = \infty$, then the boundary condition in (1.1) becomes $u = 0$ and we have the Dirichlet problem

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{1.2}$$

Denote the first eigenvalue by $\lambda_1(\Omega)$ as before. Our approach also allows us to recover uniqueness in the classical Faber–Krahn inequality for Dirichlet boundary conditions. Statements on uniqueness already appeared in [12, 18], although there were some unresolved issues about rigor at the time. A discussion about the difficulties arising in uniqueness theorems is given in [17, section II.8], where it is also noted that some

*Received by the editors November 21, 2006; accepted for publication (in revised form) March 30, 2007; published electronically November 21, 2007.
http://www.siam.org/journals/sima/39-4/67562.html

†School of Mathematics and Statistics, The University of Sydney, Sydney, NSW 2006, Australia (D.Daners@maths.usyd.edu.au, J.Kennedy@maths.usyd.edu.au).

works, including the seminal book of Pólya and Szegö [21], omit the issue altogether. Moreover, most sources stating a uniqueness result, including the recent book [15, Remark 3.2.2], do not make precise assumptions on the domains. We therefore include a proof of the following theorem.

THEOREM 1.2 (Dirichlet problem). *Suppose $\Omega \subset \mathbb{R}^N$ $(N \geq 2)$ is an arbitrary bounded domain and $B \subset \mathbb{R}^N$ a ball with the same measure as Ω. Then $\lambda_1(B) \leq \lambda_1(\Omega)$ with equality if and only if Ω is a ball up to a set of capacity zero.*

Our proof of Theorem 1.1 makes use of a functional of the level sets of the first eigenfunction ψ of (1.1). This functional was used in [6, 9] to obtain a lower bound for $\lambda_1(\Omega)$. In this paper (see section 2) we strengthen that bound, which is then the critical ingredient in obtaining the uniqueness of the minimizer as stated in Theorem 1.1. The main idea then is to use the uniqueness in the geometric isoperimetric inequality to complete the proof of uniqueness. This is done in section 3. We will set up the proof so that it works for both Robin and Dirichlet problems, providing a new proof of uniqueness of the minimizer in case of the Dirichlet problem, at least if Ω is Wiener regular, which is the case if and only if the first eigenfunction of (1.2) is continuous in $\overline{\Omega}$.

In section 4 we provide an alternative proof of uniqueness. We show that if one level set of the first eigenfunction is a ball, then the domain is already a ball. Then we use the fact that $\lambda_1(\Omega) = \lambda_1(B)$ implies that most level sets are balls. We also include a proof of Theorem 1.2 in full generality at the end of section 4 using results on symmetrization from [7].

The final section is concerned with a classification of all elliptic operators which have the same first eigenfunction as the Laplace operator with the corresponding boundary conditions. We use this to classify all operators for which the first eigenvalue is minimal. This rectifies an error in [5, Theorem 3.8] and complements a counterexample in [16, section 3] by giving a complete classification (see Theorem 5.4).

2. A lower bound for the first eigenvalue. We will start by considering a problem with mixed Dirichlet and Robin boundary conditions, as was done in [9]. This is more general than what will ultimately be needed. We look at the mixed boundary value problem

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_0, \\ \dfrac{\partial u}{\partial \nu} + \beta u = 0 & \text{on } \Gamma_1 \end{cases} \tag{2.1}$$

on the C^2-domain Ω, where Γ_0, Γ_1 are disjoint open and closed subsets of $\partial\Omega$ with $\Gamma_0 \cup \Gamma_1 = \partial\Omega$. We assume that $\beta \in C^1(\Gamma_1)$, with $\beta(x) > 0$ for all $x \in \Gamma_1$. If $\Gamma_0 = \emptyset$, then we have a pure Robin problem, and if $\Gamma_1 = \emptyset$ we have a pure Dirichlet problem.

For open sets $U \subset \Omega$, we define the interior and exterior boundaries of U with respect to Ω by

$$\partial_i U := \partial U \cap \Omega \qquad \text{and} \qquad \partial_e U := \partial U \cap \partial\Omega,$$

respectively, so that $\partial U = \partial_i U \cup \partial_e U$, and $\partial_i U \cap \partial_e U = \emptyset$.

If $U \subset \Omega$ is open with $\overline{U} \cap \Gamma_0 = \emptyset$, and $\varphi \in C(\Omega)$ is nonnegative, then as in [9, section 2], we define the functional

$$H_\Omega(U, \varphi) := \frac{1}{|U|} \left(\int_{\partial_i U} \varphi \, d\sigma + \int_{\partial_e U} \beta \, d\sigma - \int_U |\varphi|^2 \, dx \right), \tag{2.2}$$

where $|U|$ denotes the Lebesgue measure of U and σ the $(N-1)$-dimensional Hausdorff measure on ∂U. Since φ is continuous on $\partial_i U$, these integrals are all well defined, although the first and last could be infinite for some choices of U and φ.

In the case of a pure Robin problem, we restrict our choice of test functions φ to a subset $M_\beta := M_\beta(\Omega)$ of $C(\Omega)$, namely

$$M_\beta := \left\{ u \in C(\Omega)\colon u \geq 0 \text{ in } \Omega \text{ and } \limsup_{x \to z} u(x) \leq \beta(z) \text{ for all } z \in \partial\Omega \right\}.$$

Observe that this restriction for Robin problems is a natural analogue of the unrestricted class $\{u \in C(\Omega)\colon u \geq 0 \text{ in } \Omega\}$ in the Dirichlet case, since in the latter case where $\beta = \infty$, we have $M_\infty = C(\Omega)$.

We are interested in the case where the subsets U are the level sets of the first eigenfunction ψ of (2.1). The first eigenvalue is simple, and the first eigenfunction can be chosen to be positive. We denote the level sets by

$$U_t := \{x \in \Omega\colon \psi(x) > t\} \tag{2.3}$$

and the level surfaces by

$$S_t := \{x \in \Omega\colon \psi(x) = t\}. \tag{2.4}$$

Then S_t coincides with the interior boundary $\partial_i U_t$ of U_t for almost all t. Now it is well known (see [1, Theorem 4.2]) that

$$\psi \in W_p^2(\Omega) \cap C^\infty(\Omega)$$

for all $p \in (1, \infty)$, and in particular, the level sets U_t are open. Moreover, by standard embedding theorems, $\psi \in C^1(\overline{\Omega})$. We can choose $\psi \geq 0$ and normalize it so that $\|\psi\|_\infty = 1$. We set

$$m := \min_{x \in \overline{\Omega}} \psi(x).$$

By the Hopf maximum principle, $\psi(x) > 0$ for all $\Omega \in \Gamma_1$, and ψ attains its minimum m on $\partial\Omega$. If $\Gamma_0 = \emptyset$, then $m > 0$; otherwise $m = 0$. Finally, we observe that $S_t = \emptyset$ if $t \notin (m, 1]$, and $\overline{U}_t \cap \Gamma_0 = \emptyset$ for all $t \in (m, 1)$.

Let $\varphi \in C(\Omega)$ be nonnegative and set

$$w := \varphi - \frac{|\nabla \psi|}{\psi}.$$

We will obtain a characterization of the functional H_Ω in terms of $\lambda_1(\Omega)$ and the function

$$F(t) := \int_t^1 \frac{1}{\tau} \int_{S_\tau} w \, d\sigma \, d\tau.$$

THEOREM 2.1. *Let w, F be as defined above. Then $F\colon (0,1] \to \mathbb{R}$ is locally absolutely continuous and*

$$H_\Omega(U_t, \varphi) = \lambda_1(\Omega) - \frac{1}{|U_t|} \left(\frac{1}{t} \frac{d}{dt} \big(t^2 F(t)\big) + \int_{U_t} |w|^2 \, dx \right) \tag{2.5}$$

for almost all $t \in (m, 1)$.

Proof. By [9, Lemma 3.3] the function F is absolutely continuous on $[\varepsilon, 1)$ for all $\varepsilon \in (0,1)$, and

$$F'(t) = -\frac{1}{t}\int_{S_t} w\,d\sigma \tag{2.6}$$

for almost all $t \in (0,1)$. Hence

$$\begin{aligned}\frac{1}{t}\frac{d}{dt}\bigl(t^2F(t)\bigr) &= \frac{1}{t}\bigl(2tF(t) + t^2F'(t)\bigr) = 2F(t) + tF'(t)\\ &= 2\int_t^1 \frac{1}{\tau}\int_{S_\tau} w\,d\sigma\,d\tau - \int_{S_t} w\,d\sigma\end{aligned}$$

for almost all $t \in (0,1)$. Therefore (2.5) is equivalent to the characterization

$$H_\Omega(U_t,\varphi) = \lambda_1(\Omega) + \frac{1}{|U_t|}\left(\int_{S_t} w\,d\sigma - 2\int_t^1 \frac{1}{\tau}\int_{S_\tau} w\,d\sigma\,d\tau - \int_{U_t}|w|^2\,dx\right) \tag{2.7}$$

for almost all $t \in (m,1)$, which was proved in [9, Theorem 2.2]. ∎

In the next two theorems we will use the characterization (2.5) of H_Ω to estimate the first eigenvalue from above. The theorems strengthen Theorems 3.1 and 3.5 of [9], respectively, giving a *strict inequality* for a *larger set* of $t \in (m,1)$ rather than just an inequality for some $t \in (m,1)$. That strict inequality is the key to proving the uniqueness of the minimizing domain.

The first of the theorems deals with the pure Robin problem, while the second deals with mixed or pure Dirichlet boundary conditions. Part of the proof of the latter requires a different method.

THEOREM 2.2. *Suppose that $\Gamma_0 = \emptyset$, and let $\varphi \in M_\beta$. If $\varphi \neq |\nabla\psi|/\psi$, then there exists a set $S \subset (m,1)$ of positive measure such that*

$$H_\Omega(U_t,\varphi) < \lambda_1(\Omega)$$

for all $t \in S$.

Proof. We give a proof by contradiction. Assume that $\varphi \neq |\nabla\psi|/\psi$ and that $H_\Omega(U_t,\varphi) \geq \lambda_1(\Omega)$ for almost all $t \in (m,1)$. By Theorem 2.1,

$$\frac{d}{dt}\bigl(t^2F(t)\bigr) = 2tF(t) + t^2F'(t) \leq -t\int_{U_t}|w|^2\,dx \leq 0 \tag{2.8}$$

for almost all $t \in (m,1)$. Using (2.6) we can also write the above inequality as

$$2F(t) + \int_{U_t}|w|^2\,dx \leq \int_{S_t} w\,d\sigma \tag{2.9}$$

for almost all $t \in (m,1)$. Now the fundamental theorem of calculus for absolutely continuous functions [22, Theorem 8.17] and (2.8) show that $t^2F(t)$ is nonincreasing on $(m,1)$. By assumption $\varphi \neq |\nabla\psi|/\psi$, so $w = \varphi - |\nabla\psi|/\psi \neq 0$. As $w \in C(\Omega)$ and $\bigcup_{t\in(m,1]} U_t = \Omega$, there exists $t \in (m,1)$ such that

$$\int_{U_t}|w|^2\,dx > 0. \tag{2.10}$$

Moreover, since the U_t are level sets, the map

$$t \mapsto \int_{U_t} |w|^2 \, dx$$

is nonincreasing, while

$$\int_{U_1} |w|^2 \, dx = 0$$

since $U_1 = \emptyset$ by definition. We now let

$$t^* = \sup \left\{ t \in (m,1)\colon \int_{U_t} |w|^2 \, dx > 0 \right\}.$$

From (2.10) we know that $t^* \in (m, 1]$, and thus

$$\frac{d}{dt}\bigl(t^2 F(t)\bigr) \leq -t \int_{U_t} |w|^2 \, dx < 0$$

for almost all $t \in (m, t^*)$. Hence $t^2 F(t)$ is strictly decreasing on (m, t^*). We showed earlier that $t^2 F(t)$ is nonincreasing on $[t^*, 1)$ and that F is continuous with $F(1) = 0$. It follows that there exist $\eta > 0$ and $t_0 \in (m, t^*)$ such that $F(t) > \eta$ for all $t \in (m, t_0]$. By [9, Lemma 2.3], there exist $t_1 \in (m, t_0]$ and $c > 0$ such that $\sigma(S_t) < c\sigma(\partial\Omega)$ for all $t \in (m, t_1]$. If we set

$$\varepsilon := \frac{\eta}{c\sigma(\partial\Omega)},$$

then by [9, Lemma 3.4], there exists $\delta > 0$ such that $w(x) \leq \varepsilon$ for all $x \in \Omega$ with $\operatorname{dist}(x, \partial\Omega) < \delta$. Since ψ attains a strict minimum on $\partial\Omega$, there exists $t \in (m, t_1]$ such that $\operatorname{dist}(x, \partial\Omega) < \delta$ for all $x \in S_t$. For such t, using (2.9), and by our choice of ε, η,

$$0 < 2\eta < 2F(t) \leq \int_{S_t} w \, d\sigma \leq \varepsilon\sigma(S_t) \leq \eta,$$

which is obviously a contradiction. Hence the assertion of the theorem follows. ∎

We next prove a similar result for problems involving Dirichlet boundary conditions on at least part of the boundary.

THEOREM 2.3. *Suppose $\Gamma_0 \neq \emptyset$, and suppose $\varphi \in C(\Omega)$ is nonnegative. If $\varphi \neq |\nabla\psi|/\psi$, then there exists a set $S \subset (0,1)$ of positive measure such that*

$$H_\Omega(U_t, \varphi) < \lambda_1(\Omega)$$

for all $t \in S$.

Proof. As with Theorem 2.2, we give a proof by contradiction. We assume that $\varphi \neq |\nabla\psi|/\psi$ and that $H_\Omega(U_t, \varphi) \geq \lambda_1(\Omega)$ for almost all $t \in (0, 1)$. As in the proof of Theorem 2.2, there exists $t^* \in (0, 1]$ satisfying

$$t^* = \sup \left\{ t \in (0,1)\colon \int_{U_t} |w|^2 \, dx > 0 \right\},$$

so that $G(t) := t^2 F(t)$ and $F(t)$ are positive and strictly decreasing on $(0, t^*)$ with $G(t^*) = G(1) = 0$. Hence

$$g(t) := \frac{1}{G(t)}$$

is well defined and strictly increasing on $(0, t^*)$. Since $F(t) \geq 0$ and

$$G'(t) = \frac{d}{dt}\big(t^2 F(t)\big) \leq -t \int_{U_t} |w|^2 \, dx \leq 0$$

as with (2.9), it follows that

$$\int_{S_t} w \, d\sigma \geq 2F(t) + \int_{U_t} |w|^2 \, dx \geq 0$$

for almost all $t \in (0,1)$. Hence, by the coarea formula (see [11, section 3.4.3] or [20, section 1.2.4]) and the Cauchy–Schwarz inequality,

$$\begin{aligned} G(t) = t^2 F(t) &= t \int_t^1 \frac{t}{\tau} \int_{S_\tau} w \, d\sigma \, d\tau \\ &\leq t \int_t^1 \int_{S_\tau} w \, d\sigma \, d\tau = t \int_{U_t} w |\nabla \psi| \, dx \\ &\leq t \left(\int_{U_t} |w|^2 \, dx \right)^{1/2} \left(\int_{U_t} |\nabla \psi|^2 \, dx \right)^{1/2} \end{aligned}$$

for all $t \in (0,1)$. By choice of t^*,

$$G'(t) = \frac{d}{dt}\big(t^2 F(t)\big) \leq -t \int_{U_t} |w|^2 \, dx < 0$$

for almost all $t \in (0, t^*)$. Combining these inequalities, we get

$$t g'(t) = -\frac{t G'(t)}{(G(t))^2} \geq \left(\int_{U_t} |\nabla \psi|^2 \, dx \right)^{-1} \geq c := \frac{1}{\|\nabla \psi\|^2}$$

for almost all $t \in (0, t^*)$. If we fix $t_1 \in (0, t^*)$, we therefore have $g'(t) > ct^{-1}$ for almost all $t \in (0, t_1]$. Since G is absolutely continuous and positive on $[\varepsilon, t^*)$ for all $\varepsilon \in (0, t^*)$, so is g. By the fundamental theorem of calculus for such functions (see [22, Theorem 8.17]),

$$g(t_1) \geq g(t_1) - g(\varepsilon) = \int_\varepsilon^{t_1} g'(\tau) \, d\tau \geq c \int_\varepsilon^{t_1} \frac{1}{\tau} \, d\tau = c(\log t_1 - \log \varepsilon)$$

for all $\varepsilon \in (0, t_1]$. Hence $-\log \varepsilon$ is bounded from above as $\varepsilon \to 0$, which is a contradiction. This completes the proof of the theorem. ∎

Remark 2.4. It was a consequence of Theorem 3.5 in [9] that for the Dirichlet case,

$$\lambda_1(\Omega) = \max_{\substack{\varphi \in C(\Omega) \\ \varphi \geq 0}} \left(\operatorname*{ess-inf}_{t \in (0,1)} H_\Omega(U_t, \varphi) \right) = \max_{\substack{\varphi \in C(\Omega) \\ \varphi \geq 0}} \left(\operatorname*{ess-inf}_{\substack{U \subset \Omega \text{ open} \\ \overline{U} \cap \Gamma_0 = \emptyset}} H_\Omega(U, \varphi) \right),$$

where the maximum is achieved by $\varphi := |\nabla \psi|/\psi$. Theorem 2.3 shows that in fact $|\nabla \psi|/\psi$ is the *unique* maximizer.

3. Proof of uniqueness. In this section we give a proof of the main uniqueness result stated in Theorem 1.1. We start with a brief discussion of (1.1) in the case where Ω is a ball. Assume that $\beta \in (0, \infty)$ is constant and that B is a ball of radius R centered at the origin. We will denote the first eigenvalue of (1.1) on B by $\lambda_1(B)$ and the corresponding eigenfunction by ψ^*. We choose $\psi^* \geq 0$ and normalize it so that $\|\psi^*\|_\infty = 1$. Since B is a ball, the eigenfunction is radially symmetric; that is, there exists a function $v \in C^1([0, R])$ satisfying $\psi^*(x) = v(|x|)$. Set

$$\varphi^* := \frac{|\nabla \psi^*|}{\psi^*}.$$

In fact, the eigenfunction is explicitly given in terms of Bessel functions by

$$\psi^*(x) = c|x|^{-(N/2-1)} J_{N/2-1}\left(\sqrt{\lambda_1(B)}|x|\right) \tag{3.1}$$

for some normalizing constant c. Since ψ^* is radially symmetric, it is constant on $\partial\Omega$, and hence by [9, Remark 3.2] we have $\varphi^* \in M_\beta$. Also by the radial symmetry of ψ^*,

$$g(|x|) := \varphi^*(x),$$

where $x \in B$. It is shown in [9, Lemma 4.1] that the function $g\colon (0, R) \to (0, \infty)$ is strictly increasing. Similar statements hold for Dirichlet boundary conditions, that is, (1.2).

In what follows, we will deal only with pure Robin boundary conditions. The proof for pure Dirichlet boundary conditions is very similar, with obvious modifications. The principal difference is that Theorem 2.3 is used in place of Theorem 2.2.

Suppose that Ω is a bounded C^2-domain for the problem (1.1), and let B be the ball having the same measure as Ω. As in [6, 9], we define a function $\varphi \in M_\beta$ by constructing a rearrangement of φ^*. Continuing with the notation of the previous section, we consider the level sets U_t and S_t as defined by (2.3) and (2.4), respectively. We will denote by B_r the ball of radius r centered at the origin, and let $r(t)$ be the radius of the ball with the same measure as U_t, that is, so that $|B_{r(t)}| = |U_t|$. Since Ω and B have the same measure and $U_m = \Omega$, we have $r(m) = R$. Given $t \in (m, 1]$ we define

$$\varphi(x) := g(r(t))$$

for all $x \in S_t$. Since Ω is a disjoint union of the sets S_t, $t \in (m, 1]$, the function $\varphi\colon \Omega \to (0, \infty)$ is well defined. From [9, Lemma 4.2] we have the following result.

LEMMA 3.1. *The function φ constructed above lies in $M_\beta(\Omega)$ and*

$$\lambda_1(B) = H_B(B_{r(t)}, \varphi^*) \leq H_\Omega(U_t, \varphi) \tag{3.2}$$

for all $t \in (m, 1)$.

The next two lemmas will allow us to conclude that almost all level sets U_t are concentric balls if $\lambda_1(\Omega) = \lambda_1(B)$. Theorem 2.2 will be essential for that.

LEMMA 3.2. *Suppose that Ω is a bounded C^2-domain such that $\lambda_1(\Omega) = \lambda_1(B)$. If we let φ denote the rearrangement of $\varphi^* = |\nabla\psi^*|/\psi^*$ as defined above, then $\varphi = |\nabla\psi|/\psi$, and $H_\Omega(U_t, \varphi) = \lambda_1(B)$ for almost all $t \in (m, 1)$.*

Proof. If $\lambda_1(\Omega) = \lambda_1(B)$, then by Lemma 3.1, $\lambda_1(\Omega) = \lambda_1(B) \le H_\Omega(U_t, \varphi)$ for all $t \in (m, 1)$. Hence by Theorem 2.2, $\varphi = |\nabla\psi|/\psi$, and hence by Theorem 2.1,

$$H_\Omega(U_t, \varphi) = \lambda_1(\Omega) = \lambda_1(B) \tag{3.3}$$

for almost all $t \in (m, 1)$. ◻

The following lemma makes the link to the uniqueness of the ball in the geometric isoperimetric inequality. Since it shows how the uniqueness in the geometric isoperimetric inequality is used, we repeat the proof given in [9, Remark 4.3].

LEMMA 3.3. *Suppose U_t is a Lipschitz domain for some $t \in (m, 1)$, and let φ be the rearranged function as above. Then $H_\Omega(U_t, \varphi) = \lambda_1(B)$ if and only if U_t is a ball and $\sigma(\partial_e U_t) = 0$.*

Proof. We know that $\lambda_1(B) = H_B(B_{r(t)}, \varphi^*)$ for all $t \in (m, 1)$. Also, since by construction the level sets of φ^* and φ have the same measure,

$$\int_{U_t} |\varphi|^2\, dx = \int_{B_{r(t)}} |\varphi^*|^2\, dx$$

for all $t \in (m, 1]$ (see [20, section 1.2.3]). Hence, using the definition of H_Ω and φ,

$$\begin{aligned} H_\Omega(U_t, \varphi) &= \frac{1}{|U_t|}\left(\int_{S_t} \varphi\, d\sigma + \int_{\partial_e U_t} \beta\, d\sigma - \int_{U_t} |\varphi|^2\, dx\right) \\ &= \frac{1}{|B_{r(t)}|}\left(g(r(t))\sigma(S_t) + \beta\sigma(\partial_e U_t) - \int_{B_{r(t)}} |\varphi^*|^2\, dx\right). \end{aligned}$$

If U_t is a ball with $\sigma(\partial_e U_t) = 0$, then $\sigma(S_t) = \sigma(\partial B_{r(t)})$, and

$$\begin{aligned} H_\Omega(U_t, \varphi) &= \frac{1}{|B_{r(t)}|}\left(g(r(t))\sigma(\partial B_{r(t)}) - \int_{B_{r(t)}} |\varphi^*|^2\, dx\right) \\ &= H_B(B_{r(t)}, \varphi^*) = \lambda_1(B). \end{aligned}$$

Conversely, if $H_\Omega(U_t, \varphi) = \lambda_1(B)$, then for this t,

$$g(r(t))\sigma(S_t) + \beta\sigma(\partial_e U_t) = g(r(t))\sigma(\partial B_{r(t)}).$$

Since $t \in (m, 1)$ and $0 < g(r(t)) < \beta$, this is only possible if

$$\int_{\partial_e U_t} \beta\, d\sigma = 0$$

and $\sigma(S_t) = \sigma(\partial B_{r(t)})$. As $\beta > 0$ on $\partial\Omega$ by assumption, we conclude that $\sigma(\partial_e U_t) = 0$. Now since ∂U_t is the disjoint union of S_t and $\partial_e U_t$, if $\sigma(\partial_e U_t) = 0$, then we get $\sigma(\partial U_t) = \sigma(\partial B_{r(t)})$. But we know that the ball is the unique minimizer of the isoperimetric inequality, at least among Lipschitz domains (see [8, Theorem 10.2.1]). Hence $U_t = B_{r(t)} + z$ for some $z \in \mathbb{R}^N$. ◻

The last ingredient we need in the proof of the main result is the following lemma, which will allow us to conclude that if U_t is a ball for some $t \in (m, 1)$, then all level sets interior to U_t are concentric balls.

LEMMA 3.4. *Assume $-\Delta u = \lambda u$ in Ω for some $\lambda > 0$. Suppose that for some $t \ge 0$, the level set $\{x \in \Omega : u(x) > t\} = B(x_0, r)$ for some $x_0 \in \Omega$ and $r > 0$. If*

$u \in C(\overline{B(x_0,r)})$ and $\sigma(\partial_e B(x_0,r)) = 0$, then u is radially symmetric with respect to x_0 in $B(x_0,r)$.

Proof. Set $v(x) := u(x) - t$. Then

$$-\Delta v = -\Delta u = \lambda u = \lambda(u-t) + \lambda t = \lambda v + \lambda t$$

in $B(x_0,r)$. Since $u \in C(\overline{B(x_0,r)})$ and $\sigma(\partial_e B(x_0,r)) = 0$, we get $u(x) = t$ for all $x \in \partial B(x_0,r)$, so $v = 0$ on $\partial B(x_0,r)$. Hence if we set $f(v) := \lambda v + \lambda t$, then v is a positive solution of the Dirichlet problem

$$\begin{cases} -\Delta v = f(v) & \text{in } B(x_0,r), \\ \quad v = 0 & \text{on } \partial B(x_0,r). \end{cases}$$

Clearly $f(v) \geq 0$ if $v \geq 0$, and by assumption, in $B(x_0,r)$, $u > t$, so $v = u - t > 0$. Hence by a result of Gidas, Ni, and Nirenberg (see [14, Corollary 3.5]), v is radially symmetric on $B(x_0,r)$ with respect to x_0. ☐

Remark 3.5. We could state Lemma 3.4 in greater generality. If the level set U_t is Steiner symmetric (see [14, Definition 3.2]), then the solution u will have the same symmetry. The proof is the same as the one we have given above, except that we refer to [14, Corollary 3.4].

We are now in a position to prove the main uniqueness theorem stated in the introduction.

Proof of Theorem 1.1. By Lemma 3.2, $H_\Omega(U_t,\varphi) = \lambda_1(B)$ for almost all $t \in (m,1)$, and so by Lemma 3.3, U_t is a ball for almost all $t \in (m,1)$. (Here we also use the property that U_t is Lipschitz for almost all t as shown in [9, Lemma 2.3].) For every such t, Lemma 3.3 also tells us that $\sigma(\partial_e U_t) = 0$, and so Lemma 3.4 applies. Hence the eigenfunction ψ corresponding to $\lambda_1(\Omega)$ is radially symmetric inside U_t, and all interior level sets U_τ, $\tau \in (t,1]$, are concentric balls. In particular, for all $t \in (m,1]$, the level sets U_t are concentric balls, and so $\Omega = \bigcup_{t\in(m,1]} U_t$ is a ball. ☐

The proof of equality in the case of Dirichlet boundary conditions in Theorem 1.2 is very similar to the one given above. We note that $\psi \in C(\overline{\Omega})$ if and only if Ω is Wiener (Dirichlet) regular (see [2, Corollary 2.5]). Hence in that case the level surfaces S_t are contained in Ω for all $t > 0$, and therefore the assertion of Theorem 2.3 is valid as is Lemma 3.3.

4. An alternative approach to uniqueness. In this section we show that if one of the level sets of the first eigenfunction of (1.1) is a ball, then Ω is already a ball. Combining that result with Lemma 3.3, we get an alternative proof for Theorem 1.1. Throughout this section we use the same notation as in the previous section.

PROPOSITION 4.1. *Suppose that $U_t \subset \Omega$ is a ball for some t with $\sigma(\partial_e U_t) = 0$ and that B is a ball concentric to U_t. If $x_0 \in \partial\Omega \cap \partial B$ is such that the outward pointing unit normals to B and Ω coincide at x_0, then $\lambda_1(\Omega) = \lambda_1(B)$.*

Proof. Suppose that U_t is a ball with $\sigma(\partial_e U_t) = 0$. Without loss of generality, we can assume that it is centered at $x = 0$. By Lemma 3.4 the function ψ is radially symmetric in U_t. Similarly to (3.1), ψ is therefore given by

$$\psi(x) = \tilde{\psi}(x) := c|x|^{-(N/2-1)} J_{N/2-1}\left(\sqrt{\lambda_1(\Omega)}|x|\right)$$

in U_t for some normalizing constant $c > 0$. By the unique continuation property of solutions to elliptic equations (see [4]), $\psi(x) = \tilde{\psi}(x)$ for all $x \in \Omega$. Moreover,

$$-\Delta\tilde{\psi} = \lambda_1(\Omega)\tilde{\psi}$$

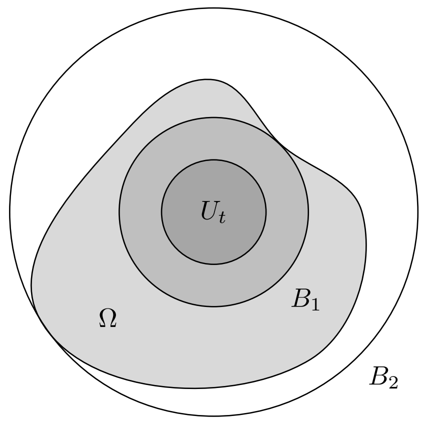


FIG. 4.1. *The balls B_1 and B_2.*

on $\mathbb{R}^N$. Now let B be a ball concentric to U_t and $x_0 \in \partial\Omega \cap \partial B$ such that the unit normals to B and Ω coincide at x_0. Note that $\psi(x_0) > 0$, so, if ν_B is the outer unit normal to B, then by the boundary conditions in (1.1),

$$\beta = -\frac{1}{\psi(x_0)}\frac{\partial\psi}{\partial\nu}(x_0) = -\frac{1}{\tilde{\psi}(x_0)}\frac{\partial\tilde{\psi}}{\partial\nu_B}(x_0).$$

Since $\tilde{\psi}$ is radially symmetric, $\tilde{\psi}$ satisfies the eigenvalue problem

$$\begin{cases} -\Delta u = \lambda_1(\Omega)u & \text{in } B,\\ \dfrac{\partial u}{\partial\nu_B} + \beta u = 0 & \text{on } \partial B. \end{cases}$$

As $\psi > 0$ and $\psi = \tilde{\psi}$ on Ω and $x_0 \in \partial\Omega \cap \partial B$, it follows that $\tilde{\psi} > 0$ on B by the radial symmetry. As $\lambda_1(B)$ is the only eigenvalue with a positive eigenfunction, we get $\lambda_1(\Omega) = \lambda_1(B)$. ∎

From the above we get as a corollary that if U_t is a ball for some t, then Ω is already a ball.

THEOREM 4.2. *Suppose that $U_t \subset \Omega$ is a ball for some t with $\sigma(\partial_e U_t) = 0$. Then Ω is a ball.*

Proof. Suppose that $U_t \subset \Omega$ is a ball with $\sigma(\partial_e U_t) = 0$ for some t. Assume without loss of generality that its center is at $x = 0$. Denote by B_r the open ball with radius r. We set

$$r_1 := \sup\{r > 0 \colon B_r \subset \Omega\} \quad \text{and} \quad r_2 := \inf\{r > 0 \colon \Omega \subset B_r\}$$

and let $B_i := B_{r_i}$ for $i = 1, 2$. Then

$$U_t \subset B_1 \subset \Omega \subset B_2 \tag{4.1}$$

as shown in Figure 4.1. Since Ω is smooth, ∂B_i is tangential to $\partial\Omega$ at some point on $\partial\Omega \cap \partial B_i$, so the outward pointing unit normals to B_i and Ω coincide at that point. Now Proposition 4.1 implies that $\lambda_1(B_1) = \lambda_1(\Omega) = \lambda_1(B_2)$. But $\lambda_1(B_r)$ is strictly decreasing in r, so $B_1 = B_2$. By (4.1) we conclude that Ω is a ball. ∎

Remark 4.3. The above also works for Dirichlet boundary conditions. We need only assume that $\Omega \subset \mathbb{R}^N$ is a bounded open connected set and conclude that Ω is a ball up to a set of capacity zero. To see this, suppose that U_t is a ball for some $t > 0$. As before, we assume its center is $x = 0$. We know that the eigenfunction ψ is discontinuous only at the irregular points, so $U_t \cap \partial\Omega$ consists of irregular points if nonempty. The set of irregular points has capacity zero (see [19, page 232]), so in particular it has surface measure zero on ∂U_t. Since U_t is a ball it follows that ψ is continuous on $\overline{U}_t$, so by the arguments in the proof of Lemma 3.4 the eigenfunction ψ is radially symmetric in U_t. Then as in the proof of Proposition 4.1, $\psi = \tilde{\psi}$ on Ω, and therefore ψ is continuous on $\overline{\Omega}$. Now let B be the smallest ball with center $x = 0$ containing Ω. Then since $\psi > 0$ in Ω and $\tilde{\psi}$ is radially symmetric, $\tilde{\psi} > 0$ on B. Hence $\lambda_1(\Omega) = \lambda_1(B)$, and thus by [3, Theorem 3.1] the set Ω coincides with the ball B up to a set of capacity zero.

We can now give an alternative proof of Theorem 1.1.

Alternative proof of Theorem 1.1. Suppose that $\lambda_1(\Omega) = \lambda_1(B)$, where B is a ball of the same measure as Ω. By Lemma 3.3 there is a level set U_t, which is a ball, and also $\overline{U}_t \subset \Omega$. Then Theorem 4.2 implies that Ω is a ball as claimed. $\square$

The above proof also applies in the case of Dirichlet boundary conditions, provided ψ is continuous on $\overline{\Omega}$. The reason that we use continuity is that otherwise we do not know whether Lemma 3.2 applies. There is, however, another proof based on symmetrization that works for arbitrary domains.

Proof of Theorem 1.2. Let Ω be a bounded domain and ψ the first eigenvalue of (1.2) normalized such that $\psi(x) > 0$ for all $x \in \Omega$ and $\|\nabla\psi\|_2 = 1$. We know that $\psi \in H_0^1(\Omega)$, so if we extend ψ by zero outside Ω, we may consider ψ as an element of $H^1(\mathbb{R}^N)$. Denote by ψ^* the spherical symmetric rearrangement of ψ (see [7, 21]). By the results in [7] we have $\psi^* \in H^1(\mathbb{R}^N)$ with $\|\psi^*\|_2 = 1$ and $\|\nabla\psi^*\|_2 \leq \|\nabla\psi\|_2$. By construction, ψ^* has support in the closure of a ball B of the same measure as Ω. Here we use that a function in $H_0^1(\Omega)$ is zero on Ω^c almost everywhere. As B has a smooth boundary, $\psi^* \in H_0^1(B)$, and thus by the variational characterization of the first eigenvalue,

$$\lambda_1(B) \leq \|\nabla\psi^*\|_2^2 \leq \|\nabla\psi\|_2^2 = \lambda_1(\Omega).$$

If $\lambda_1(B) = \lambda_1(\Omega)$, then from the above, $\|\nabla\psi^*\|_2^2 = \|\nabla\psi\|_2^2$. Also note that $\nabla\psi \neq 0$ almost everywhere in Ω by Remark 5.2(b) below. Hence [7, Theorem 1.1] implies that $\psi = \psi^*$ up to translation almost everywhere. Therefore, possibly after a translation, $\Omega = B$ up to a set of measure zero. We claim that $\Omega \subset B$. If not, then there exists a point $x_0 \in \Omega \cap B^c$. This is not possible since Ω is open, and therefore $\Omega \cap B^c$ has non-zero measure. Hence $\Omega \subset B$. But then $\lambda_1(\Omega) = \lambda_1(B)$ if and only if $\operatorname{cap}(B \setminus \Omega) = 0$ by [3, Theorem 3.1], completing the proof of Theorem 1.2. $\square$

5. Operators other than the Laplace operator.

For the remainder of the paper, we will consider the eigenvalue problem

$$\begin{cases} -\operatorname{div}\big(A(x)\nabla u(x)\big) = \lambda u & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_0, \\ \dfrac{\partial u}{\partial \nu_A} + \beta u = 0 & \text{on } \Gamma_1, \end{cases} \tag{5.1}$$

where $\Omega \subset \mathbb{R}^N$ is an open set, Γ_0 a closed subset of $\partial\Omega$ and $\Gamma_1 = \partial\Omega \setminus \Gamma_0$ Lipschitz, and

$$\frac{\partial u}{\partial \nu_A} := \big(A(x)\nabla u(x)\big) \cdot \nu$$

is the conormal derivative of u associated with the coefficient matrix A. Further assume that $A \in L_\infty(\Omega, \mathbb{R}^{N\times N})$ is such that $A(x)$ is symmetric and uniformly positive definite, that is, there exists $\alpha > 0$ such that

$$0 < \alpha \leq \inf_{|\xi|=1} \xi^T A(x)\xi$$

for almost all $x \in \Omega$. By replacing A with $\alpha^{-1}A$, we can assume without loss of generality that $\alpha = 1$, that is,

$$1 \leq \inf_{|\xi|=1} \xi^T A(x)\xi \tag{5.2}$$

for almost all $x \in \Omega$. Finally, assume that $\beta \in L_\infty(\Gamma_1)$ is nonnegative. Note that here we allow $\beta = 0$, that is, Neumann boundary conditions. The form associated with (5.1) is given by

$$a(u,v) = \int_\Omega \big(A(x)\nabla u(x)\big) \cdot \nabla v(x)\,dx + \int_{\Gamma_1} \beta(x)u(x)v(x)\,d\sigma.$$

It is well known that the above problem has a first eigenvalue, which is given by the variational characterization

$$\mu_1 = \inf_{\substack{u\in V\\ u\neq 0}} \frac{a(u,u)}{\|u\|_2^2},$$

where V is the closure of $C_c^\infty(\Omega \cup \Gamma_1)$ in $H^1(\Omega)$. The eigenvalue is simple and the corresponding eigenfunctions the only minimizers for the above quotient. We want to compare this eigenvalue to the first eigenvalue of

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega,\\ u = 0 & \text{on } \Gamma_0,\\ \dfrac{\partial u}{\partial \nu} + \beta u = 0 & \text{on } \Gamma_1, \end{cases} \tag{5.3}$$

that is, the case where $A(x) = I$. The corresponding form is

$$b(u,v) := \int_\Omega \nabla u(x) \cdot \nabla v(x)\,dx + \int_{\Gamma_1} \beta(x)u(x)v(x)\,d\sigma.$$

Let λ_1 be the first eigenvalue of (5.3) and μ_1 the first eigenvalue of (5.1). By (5.2),

$$\frac{b(v,v)}{\|v\|_2^2} \leq \frac{a(v,v)}{\|v\|_2^2}$$

for all $v \in V$, and thus by the variational characterization of λ_1 and μ_1,

$$\lambda_1 \leq \mu_1. \tag{5.4}$$

We now characterize those matrices A for which the first eigenvalues of (5.1) and (5.3) are equal.

THEOREM 5.1. *Suppose $\Omega \subset \mathbb{R}^N$ is a bounded domain. Let λ_1 and ψ be the first eigenvalue and eigenfunction of* (5.3)*, and let μ_1 and φ be the first eigenvalue and eigenfunction of* (5.1)*. Then the following assertions are equivalent:*

(1) $\lambda_1 = \mu_1$*;*

(2) $\lambda_1 = \mu_1$ *and* $\varphi = c\psi$ *for some* $c \in \mathbb{R}$*;*

(3) $A(x)\nabla\psi(x) = \nabla\psi(x)$ *almost everywhere in* Ω*.*

Proof. Throughout the proof we normalize the eigenfunctions such that they are positive and that $\|\varphi\|_2 = \|\psi\|_2 = 1$. Suppose that $\lambda_1 = \mu_1$. Then by the variational characterization of the eigenvalues and (5.2),

$$\lambda_1 \leq b(\varphi, \varphi) \leq a(\varphi, \varphi) = \mu_1 = \lambda_1.$$

Hence $\lambda_1 = b(\varphi, \varphi)$, and thus by the simplicity and normalization of the eigenvalue λ_1, it follows that $\varphi = \psi$. Hence (1) implies (2).

If (2) holds, then by the variational characterization of λ_1 and μ_1,

$$b(\psi, \psi) = \lambda_1 = \mu_1 = a(\psi, \psi)$$

or equivalently, using the definitions of the form $a(\cdot\,,\cdot)$ and $b(\cdot\,,\cdot)$,

$$\int_\Omega \big(A(x)\nabla\psi(x) - \nabla\psi(x)\big) \cdot \nabla\psi(x)\, dx = 0.$$

Since $A(x) - I$ is positive semidefinite by (5.2), it follows that the above integrand is nonnegative, and thus

$$\big(A\nabla\psi - \nabla\psi\big) \cdot \nabla\psi = 0$$

almost everywhere. Since $A(x) - I$ is positive semidefinite and symmetric almost everywhere, we conclude that

$$A\nabla\psi - \nabla\psi = 0$$

almost everywhere in Ω, proving that (2) implies (3).

To prove that (3) implies (1), we again use the variational characterization of the first eigenvalue. By (3) we have

$$b(\psi, \psi) - a(\psi, \psi) = 0,$$

so by (5.4) and the variational characterization of the eigenvalues,

$$\mu_1 \geq \lambda_1 = b(\psi, \psi) = b(\psi, \psi) - a(\psi, \psi) + a(\psi, \psi) = a(\psi, \psi) \geq \mu_1,$$

and thus $\lambda_1 = \mu_1$ as claimed. $\square$

Remark 5.2. (a) If we have pure Neumann boundary conditions, the above theorem is, of course, trivial since then $\lambda_1 = \mu_1 = 0$ and the eigenfunctions $\varphi = \psi$ are constant. Hence every matrix A satisfies the conditions. In what follows we therefore assume that $\lambda_1 > 0$, and thus the eigenfunction is not constant.

(b) We also note that $\psi \in C^\infty(\Omega)$ always, so $\nabla\psi \in C^\infty(\Omega)$ as well. There is even a stronger statement, namely, that ψ is real analytic in Ω always (see [10, section V.4.2]). Moreover, if $\lambda_1 > 0$, then ψ and also $\nabla\psi(x)$ are nonconstant. By the real analyticity, $\nabla\psi \neq 0$ almost everywhere (see [13, section 3.1.24]).

Next we characterize those matrices A for which (3) of the above theorem is satisfied. For that purpose we look at a domain $\Omega \subset \mathbb{R}^2$ in the plane and let ψ denote the first eigenvalue of (5.3). Since $\nabla\psi \neq 0$ almost everywhere as shown above, $A_0(x) := A(x) - I$ must have rank strictly less than two. Note that the set of symmetric positive semidefinite (2×2)-matrices having a given nonzero vector (a, b) in their null space are precisely the matrices

$$t\begin{bmatrix} b^2 & -ab \\ -ab & a^2 \end{bmatrix}, \qquad t \geq 0.$$

Hence, as $\nabla\psi \neq 0$ almost everywhere,

$$A_0(x)\nabla\psi(x) = 0$$

if and only if

$$A_0(x) := m(x)\begin{bmatrix} \left(\dfrac{\partial\psi}{\partial y}\right)^2 & -\dfrac{\partial\psi}{\partial x}\dfrac{\partial\psi}{\partial y} \\ -\dfrac{\partial\psi}{\partial x}\dfrac{\partial\psi}{\partial y} & \left(\dfrac{\partial\psi}{\partial x}\right)^2 \end{bmatrix}$$

almost everywhere in Ω for some measurable function $m\colon \Omega \to [0, \infty)$. If we set $A(x) := I(x) + A_0(x)$, then $A(x)$ satisfies the condition in part (3) of Theorem 5.1. Hence there are many nontrivial operators having the same eigenvalue as the Laplace operator on Ω.

Remark 5.3. We can generalize the above to $\mathbb{R}^N$. We let M denote the set of real symmetric $N \times N$ matrices. Given a nonzero vector $\xi \in \mathbb{R}^N$, we are interested in the kernel of the map $S\colon M \to \mathbb{R}$ given by $A \to \xi^T A\xi$. Now $\dim M = (N+1)N/2$, and therefore

$$\dim(\ker S) = \frac{N(N+1)}{2} - 1 = \frac{N(N-1)}{2}.$$

Let $1 \leq k < \ell \leq N$ and $\xi = (\xi_1, \ldots, \xi_N)$. If ξ_k and ξ_ℓ are not both zero, define the matrices $A_{k\ell}$ with entries $a_{kk} = \xi_j^2$, $a_{jj} = \xi_k^2$, and $a_{kl} = a_{lk} = -\xi_k\xi_\ell$, and $a_{ij} = 0$ otherwise. If $\xi_k = \xi_\ell = 0$, let $a_{kk} = a_{\ell\ell} = 1$ and $a_{kl} = a_{lk} = -1$, and $a_{ij} = 0$ otherwise. Then the set of matrices $A_{k\ell}$, $1 \leq k < \ell \leq N$, is a linearly independent subset of $\ker S$. Since there are precisely $N(N-1)/2$ such matrices, they form a basis of $\ker S$.

Note that the matrices $A_{k\ell}$ are positive semidefinite, and by construction $A_{k\ell}\xi = 0$. It follows that every positive semidefinite matrix A with $A\xi = 0$ can be written in the form

$$A = \sum_{1 \leq k < \ell \leq N} m_{k\ell}A_{k\ell}$$

for $m_{k\ell} \geq 0$.

We now use this to characterize the operators for which $\lambda_1 = \mu_1$. To do so, we set $\xi = \nabla\psi(x)$ and look at the corresponding set of matrices $A_{k\ell}(x)$ as constructed above. This can be done since $\nabla\psi \neq 0$ almost everywhere in Ω as shown in Remark 5.2. Hence, by Theorem 5.1 we have $\lambda_1 = \mu_1$ if and only if there exist measurable functions $m_{k\ell}\colon \Omega \to [0,\infty)$ such that

$$A(x) = I + \sum_{1\leq k<\ell\leq N} m_{k\ell}(x)A_{k\ell}(x)$$

for almost all $x \in \Omega$.

We next apply the above to the special case where Ω is a ball, complementing a result in [16, section 3], giving a complete characterization of all symmetric operators in divergence form for Robin and Dirichlet problems having the same first eigenvalue as the Laplacian with the corresponding boundary conditions on a domain of the same measure.

THEOREM 5.4. *Suppose that $\beta \in (0,\infty)$ is constant and that μ_1 is the first eigenvalue of*

$$\begin{cases} -\operatorname{div}\bigl(A\nabla u\bigr) = \lambda u & \textit{in } \Omega, \\ \dfrac{\partial u}{\partial \nu_A} + \beta u = 0 & \textit{on } \partial\Omega, \end{cases} \tag{5.5}$$

where Ω is a domain of class C^2 and A satisfies (5.2)*. Let B be a ball of the same measure as Ω. Finally, denote the first eigenvalue of* (5.5) *by $\mu_1(\Omega)$ and a first eigenvalue of $\lambda_1(B)$ by ψ^*. Then $\mu_1(\Omega) = \lambda_1(B)$ if and only if $\Omega = B + z$ for some $z \in \mathbb{R}^N$ and $A(x-z)\nabla\psi^*(x) = \nabla\psi^*(x)$ almost everywhere.*

Proof. First suppose that $\lambda_1(B) = \mu_1(\Omega)$. By Theorem 1.1 and (5.4) we have

$$\lambda_1(B) \leq \lambda_1(\Omega) \leq \mu_1(\Omega) = \lambda_1(B),$$

implying that $\lambda_1(\Omega) = \mu_1(\Omega) = \lambda_1(B)$. From Theorem 1.1 we conclude that $\Omega = B+z$ for some $z \in \mathbb{R}^n$, and then Theorem 5.1 implies that $A(x-z)\nabla\psi^*(x) = \psi^*(x)$ almost everywhere in B. To prove the converse we again use Theorem 5.1 to conclude that $\lambda_1(B) = \mu_1(B)$ as claimed. □

Remark 5.5. Statements similar to the ones in Theorem 5.4 hold for Dirichlet boundary conditions. The only difference is that we can assume that Ω is an arbitrary bounded domain. Also a rather precise characterization of those operators for which $\mu_1(\Omega) = \lambda_1(B)$ follows from Remark 5.3.

Remark 5.6. We could look at even more general operators and characterize those for which $\mu_1(\Omega) = \lambda_1(B)$. We could look at

$$\begin{cases} -\operatorname{div}\bigl(A(x)\nabla u(x)\bigr) + b(x)u = \lambda u & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_0, \\ \dfrac{\partial u}{\partial \nu_A} + b_0 u = 0 & \text{on } \Gamma_1, \end{cases}$$

with $b \in L_\infty(\Omega)$ nonnegative and $b_0 \in L_\infty(\partial\Omega)$ such that $b_0 \geq \beta > 0$ ($\beta \in (0,\infty)$ constant). If μ_1 denotes the first eigenvalue of the above problem, then $\mu_1(\Omega) = \lambda_1(\Omega)$ implies that $b = 0$ and $b_0 = \beta$ almost everywhere. Clearly $\mu_1(\Omega) \geq \lambda_1(\Omega)$

by the variational characterization. Suppose now that $b > 0$ or $b_0 > \beta$ on a set of positive measure, and let φ denote the first eigenvalue of the above problem. The form corresponding to the above is

$$a_1(u,v) := \int_\Omega \big(A\nabla u\big) \cdot \nabla v + buv\,dx + \int_{\Gamma_1} b_0 uv\,d\sigma.$$

By the variational characterization

$$\lambda_1(\Omega) \leq \frac{b(\varphi,\varphi)}{\|\varphi\|_2^2} \leq \frac{a(\varphi,\varphi)}{\|\varphi\|_2^2} < \frac{a_1(\varphi,\varphi)}{\|\varphi\|_2^2} = \mu_1(\Omega)$$

as claimed. Here we use that $\varphi > 0$ on $\Omega \cup \Gamma_1$.

Acknowledgment. We thank E. N. Dancer for helpful discussions on section 5.

REFERENCES

[1] S. Agmon, *On the eigenfunctions and on the eigenvalues of general elliptic boundary value problems*, Comm. Pure Appl. Math., 15 (1962), pp. 119–147.

[2] W. Arendt and Ph. Bénilan, *Wiener regularity and heat semigroups on spaces of continuous functions*, in Topics in Nonlinear Analysis, J. Escher and G. Simonett, eds., Progr. Nonlinear Differential Equations Appl. 35, Birkhäuser, Basel, 1999, pp. 29–49.

[3] W. Arendt and S. Monniaux, *Domain perturbation for the first eigenvalue of the Dirichlet Schrödinger operator*, in Partial Differential Equations and Mathematical Physics, M. Demuth and B. Schulze, eds., Oper. Theory Adv. Appl. 78, Birkhäuser, Basel, 1995, pp. 9–19.

[4] N. Aronszajn, *A unique continuation theorem for solutions of elliptic partial differential equations or inequalities of second order*, J. Math. Pures Appl. (9), 36 (1957), pp. 235–249.

[5] C. Bandle, *Isoperimetric Inequalities and Applications*, Monogr. Stud. Math. 7, Pitman, Boston, MA, 1980.

[6] M.-H. Bossel, *Membranes élastiquement liées: Extension du théorème de Rayleigh-Faber–Krahn et de l'inégalité de Cheeger*, C. R. Acad. Sci. Paris Sér. I Math., 302 (1986), pp. 47–50.

[7] J. E. Brothers and W. P. Ziemer, *Minimal rearrangements of Sobolev functions*, J. Reine Angew. Math., 384 (1988), pp. 153–179.

[8] Y. D. Burago and V. A. Zalgaller, *Geometric Inequalities*, Grund. Math. Wiss. 285, Springer-Verlag, Berlin, 1988.

[9] D. Daners, *A Faber–Krahn inequality for Robin problems in any space dimension*, Math. Ann., 335 (2006), pp. 767–785.

[10] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*. Vol. 2, Springer-Verlag, Berlin, 1988.

[11] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.

[12] G. Faber, *Beweis, dass unter allen homogenen Membranen von gleicher Fläche und gleicher Spannung die kreisförmige den tiefsten Grundton gibt*, Sitzungsberichte der Mathematisch-Physikalischen Klasse der Bayerischen Akademie der Wissenschaften zu München, Die Akademie, München, 1923, pp. 169–172.

[13] H. Federer, *Geometric Measure Theory*, Grund. Math. Wiss. 153, Springer-Verlag, New York, 1969.

[14] L. E. Fraenkel, *An Introduction to Maximum Principles and Symmetry in Elliptic Problems*, Cambridge Tracts in Math. 128, Cambridge University Press, Cambridge, UK, 2000.

[15] A. Henrot, *Extremum Problems for Eigenvalues of Elliptic Operators*, Front. Math., Birkhäuser Verlag, Basel, 2006.

[16] P. Hess, *An isoperimetric inequality for the principal eigenvalue of a periodic-parabolic problem*, Math. Z., 194 (1987), pp. 121–125.

[17] B. Kawohl, *Rearrangements and Convexity of Level Sets in PDE*, Lecture Notes in Math. 1150, Springer-Verlag, Berlin, 1985.

[18] E. Krahn, *Über eine von Rayleigh formulierte Minimaleigenschaft des Kreises*, Math. Ann., 94 (1925), pp. 97–100.

[19] N. S. Landkof, *Foundations of Modern Potential Theory*, Grund. Math. Wiss. 180, Springer-Verlag, New York, 1972.

[20] V. G. Maz'ja, *Sobolev Spaces*, translated from the Russian by T. O. Shaposhnikova, Springer Ser. Soviet Math., Springer-Verlag, Berlin, 1985.
[21] G. Pólya and G. Szegö, *Isoperimetric Inequalities in Mathematical Physics*, Ann. of Math. Stud. 27, Princeton University Press, Princeton, NJ, 1951.
[22] W. Rudin, *Real and Complex Analysis*, 2nd ed., McGraw-Hill Inc., New York, 1974.

SIAM J. MATH. ANAL.
Vol. 39, No. 4, pp. 1208–1230

DIFFUSION MEDIATED TRANSPORT IN MULTIPLE STATE SYSTEMS*

STUART HASTINGS†, DAVID KINDERLEHRER‡, AND J. BRYCE MCLEOD†

Abstract. Intracellular transport in eukarya is attributed to motor proteins that transduce chemical energy into directed mechanical motion. Nanoscale motors like kinesins tow organelles and other cargo on microtubules or filaments, have a role separating the mitotic spindle during the cell cycle, and perform many other functions. The simplest description gives rise to a weakly coupled system of evolution equations. The transport process, to the mind's eye, is analogous to a biased coin toss. We describe how this intuition may be confirmed by a careful analysis of the cooperative effect among the conformational changes and potentials present in the equations.

Key words. Fokker–Planck, weakly coupled systems, molecular motor, transport

AMS subject classifications. 34D23, 35K50, 35K57, 92C37, 92C45

DOI. 10.1137/060650994

1. Introduction.

Motion in small live systems has many challenges, as famously discussed in Purcell [25]. Prominent environmental conditions are high viscosity and warmth. Not only is it difficult to move, but maintaining a course is rendered difficult by immersion in a highly fluctuating bath. Intracellular transport in eukarya is attributed to motor proteins that transduce chemical energy into directed mechanical motion. Proteins like kinesin function as nanoscale motors, towing organelles and other cargo on microtubules or filaments, have a role separating the mitotic spindle during the cell cycle, and perform many other functions. Because of the presence of significant diffusion, they are sometimes referred to as Brownian motors. Since a specific type tends to move in a single direction, for example, anterograde or retrograde to the cell periphery, these proteins are sometimes referred to as molecular ratchets. How do they overcome the issues posed by Purcell to provide the transport necessary for the activities of the cell?

There are many descriptions of the function of these proteins, or aspects of their thermodynamical behavior, beginning with Ajdari and Prost [1], Astumian [2], Doering, Ermentrout, and Oster [6], and Peskin, Ermentrout, and Oster [23]. For more recent work, note the review paper [26] and [27, Chapter 8]. Additional description is given in [8], [13], [17], [18], [21], [28]. The descriptions consist either in discussions of stochastic differential equations, which give rise to the distribution functions via the Chapman–Kolmogorov equation, or of distribution functions directly. In [5], we have suggested a dissipation principle approach for motor proteins like conventional kinesin, motivated by Howard [12]. The dissipation principle, which involves a Kantorovich–Wasserstein metric, identifies the environment of the system and gives rise to an implicit scheme from which evolution equations follow; see [3], [14], [15], [16], [19], [20], and more generally [29]. Most of these formulations consist, in the end,

*Received by the editors January 26, 2006; accepted for publication (in revised form) May 14, 2007; published electronically November 21, 2007.
http://www.siam.org/journals/sima/39-4/65099.html

†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (sph+@pitt.edu, mcleod@pitt.edu).

‡Center for Nonlinear Analysis and Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (davidk@cmu.edu). The research of this author was supported by NSF grants DMS 0305794 and DMS 0405343.

of Fokker–Planck-type equations coupled via conformational change factors, typically known as weakly coupled parabolic systems. Our own is also distinguished because it has natural boundary conditions. To investigate transport properties, our attention is directed toward the stationary solution of such a system, as we explain below.

A special collaboration among the potentials and the conformational changes in the system must be present for transport to occur. To investigate this, we introduce the n-state system. Suppose that $\rho_1, \dots, \rho_n$ are partial probability densities defined on the unit interval $\Omega = (0, 1)$ satisfying

$$\begin{aligned} &\frac{d}{dx}\left(\sigma \frac{d\rho_i}{dx} + \psi_i' \rho_i\right) + \sum_{j=1,\dots,n} a_{ij}\rho_j = 0 \text{ in } \Omega, \\ &\sigma \frac{d\rho_i}{dx} + \psi_i' \rho_i = 0 \text{ on } \partial\Omega,\ i = 1, \dots n, \\ &\rho_i \geqq 0 \text{ in } \Omega,\ \int_\Omega (\rho_1 + \cdots + \rho_n) dx = 1. \end{aligned} \tag{1.1}$$

Here $\sigma > 0, \psi_1, \dots, \psi_n$ are smooth nonnegative functions of period $1/N$, and $A = (a_{ij})$ is a smooth rate matrix of period $1/N$, that is,

$$\begin{aligned} &a_{ii} \leqq 0,\ a_{ij} \geqq 0 \text{ for } i \neq j, \\ &\textstyle\sum_{i=1,\dots,n} a_{ij} = 0,\ j = 1, \dots, n. \end{aligned} \tag{1.2}$$

We shall also have occasion to enforce a nondegeneracy condition,

$$a_{ij} \not\equiv 0 \text{ in } \Omega,\ i, j = 1, \dots, n. \tag{1.3}$$

The conditions (1.2) mean that $P = \mathbf{1} + \tau A$, for $\tau > 0$ small enough, is a probability matrix. The condition (1.3), we shall see, ensures that none of the components of ρ are identically zero passive placeholders in the system. In this context, the potentials $\psi_1, \dots, \psi_n$ describe interactions among the states, the elements of the protein's structure, and the microtubule track, and the matrix A describes interchange of activity among the states. The system (1.1) are the stationary equations of the evolution system

$$\begin{aligned} &\frac{\partial\rho_i}{\partial t} = \frac{\partial}{\partial x}\left(\sigma \frac{\partial\rho_i}{\partial x} + \psi_i' \rho_i\right) + \sum_{j=1,\dots,n} a_{ij}\rho_j \quad \text{in } \Omega,\ t > 0, \\ &\sigma \frac{\partial\rho_i}{\partial x} + \psi_i' \rho_i = 0 \quad \text{on } \partial\Omega,\ t > 0,\ i = 1, \dots n, \\ &\rho_i \geqq 0 \text{ in } \Omega,\ \int_\Omega (\rho_1 + \cdots + \rho_n) dx = 1, \quad t > 0. \end{aligned} \tag{1.4}$$

Before proceeding further, let us discuss what we intend by transport. In a chemical or conformational change process, a reaction coordinate (or coordinates) must be specified. This is the independent variable. In a mechanical system, it is usually evident what this coordinate must be. In our situation, even though both conformational change and mechanical effects are present, it is natural to specify the distance along the motor track, the microtubule, here the interval Ω, as the independent variable. We interpret the migration of density to one end of the track during the evolution as evidence of transport.

We shall show in section 4 that the stationary solution of the system (1.1), which we denote by $\rho^\sharp$, is globally stable: given any solution $\rho(x,t)$ of (1.4),

$$\rho(x,t) \to \rho^\sharp(x) \text{ as } t \to \infty. \tag{1.5}$$

So the migration of density we referred to previously may be ascertained by inspection of $\rho^\sharp$. In what follows, we simply set $\rho = \rho^\sharp$.

If the preponderance of mass of ρ is distributed at one end of the track, our domain Ω, then transport is present. Our main result, stated precisely later in section 3, is that with suitable potentials $\psi_1, \ldots, \psi_n$ and with favorable coupling between them and the rate matrix A, there are constants K and M, independent of σ, such that

$$\sum_{i=1}^{n} \rho_i \left(x + \frac{1}{N} \right) \leqq K e^{-\frac{M}{\sigma}} \sum_{i=1}^{n} \rho_i(x), \; x \in \Omega, \; x < 1 - \frac{1}{N}, \tag{1.6}$$

for sufficiently small $\sigma > 0$. So from one period to the next, total mass decays exponentially as in Bernoulli trials with a biased coin.

In summary, transport results from functional relationships in the system (1.1) or (1.4). The actual proof is technical, as most proofs are these days, and a detailed explanation of the conditions we impose is given after Theorem 3.2, when the notation has been introduced. Here we briefly consider the main elements of our thinking. We also refer to [10] for a more extended discussion.

To favor transport, we wish to avoid circumstances that permit decoupling in (1.1), for example,

$$A\rho = 0, \text{ where } \rho = (\rho_1, \ldots, \rho_n),$$

since in this case the solution vector is periodic. Such circumstances may be related to various types of detailed balance conditions. In detailed balance, we would find that

$$a_{ij}\rho_j = a_{ji}\rho_i \quad \text{for all } i, j = 1, \ldots, n,$$

implying that $A\rho = 0$. For example, if it is possible to find a solution ρ of (1.1) that minimizes the free energy of the system

$$F(\eta) = \sum_{i=1\cdots n} \int_\Omega \left\{ \psi_i \eta_i + \sigma \eta_i \log \eta_i \right\} dx,$$

then $A\rho = 0$.

But avoiding these situations is not nearly sufficient. First we require that the potentials ψ_i have some asymmetry property. Roughly speaking, to favor transport to the left, toward $x = 0$, a period interval must have some subinterval where all the potentials ψ_j are increasing. In addition every point must have a neighborhood where at least one ψ_i is increasing. Some coupling among the n states must take place.

Now we explain the nature of the coupling we impose, the properties of the matrix A. As mentioned above, in any neighborhood in Ω, at least one ψ_i should be increasing to promote transport toward $x = 0$. Density tends to accumulate near the minima of the potentials, which correspond to attachment sites of the motor to the microtubule and its availability for conformational change. This typically would be where the matrix A is supported. In a neighborhood of such a minimum, states which are not favored for left transport should have the opportunity to switch to state i, so we impose $a_{ij} > 0$ for all of these states. The weaker assumption, insisting only that the state associated with potential achieving the minimum have this switching opportunity, is insufficient. This is a type of ergodic hypothesis saying that there must be mixing between at least one potential which transports left and all the ones which may not. Our hypothesis is not optimal, but some condition is necessary. One may consider, for example, simply adding new states to the system which are uncoupled to the original states. In fact, it is possible to construct situations where there is actually transport to the right by inauspicious choice of the supports of the a_{ij}, as we show in section 5.

Here we consider only (1.1), although many other and more complex situations are possible. One example is a system where there are many conformational changes, not all related to movement. For example, one may consider the system whose stationary state is

$$\begin{aligned} \frac{d}{dx}\left(\sigma\frac{d\rho_i}{dx} + \psi_i'\rho_i\right) + \sum_{j=1,\ldots,n} a_{ij}\rho_j = 0 \text{ in } \Omega, \quad i = 1,\ldots m,\\ \sum_{j=1,\ldots,n} a_{ij}\rho_j = 0 \text{ in } \Omega, \quad i = m+1,\ldots n,\\ \sigma\frac{d\rho_i}{dx} + \psi_i'\rho_i = 0 \text{ on } \partial\Omega,\ i = 1,\ldots m,\\ \rho_i \geqq 0 \text{ in } \Omega,\ \int_\Omega(\rho_1 + \cdots + \rho_n)dx = 1. \end{aligned} \tag{1.7}$$

One such example is in [10]. We leave additional such explorations to the interested reader. In Chipot, Hastings, and Kinderlehrer [4] (cf. also [9]), the two-component system was analyzed. As well as being valid for an arbitrary number of active components, our proof here is based on a completely different and more widely applicable approach.

2. Existence. There are several ways to approach the existence question for (1.1). In [4], we gave existence results based on the Schauder fixed point theorem and a second proof based on an ordinary differential equations shooting method. The Schauder proof extends to the current situation, and to higher dimensions, but the shooting method was limited to the two-state case. Here we offer a new ordinary differential equations method proof which is of interest because it separates existence from uniqueness and positivity, showing that existence is a purely algebraic property depending only on the second line in (1.2),

$$\sum_{i=1,\ldots,n} a_{ij} = 0,\ j = 1,\ldots,n, \tag{2.1}$$

while positivity and uniqueness rely on the more geometric nature of the inequalities. We shall prove Theorem 2.1 below, followed by a brief discussion of a stronger result whose proof is essentially the same. Recall that $\Omega = (0, 1)$.

THEOREM 2.1. *Assume that $\psi_i, a_{ij} \in C^2(\overline{\Omega}), i, j = 1,\ldots,n$, and that (2.1) holds. Then there exists a (nontrivial) solution $\rho = (\rho_1,\ldots,\rho_n)$ to the equations and boundary conditions in (1.1). Assume furthermore that (1.2) and (1.3) hold. Then ρ is unique and*

$$\rho_i(x) > 0 \textit{ in } \Omega \quad \textit{and} \quad \rho_i \in C^2(\overline{\Omega}),\ i = 1,\ldots,n.$$

Proof. Introduce

$$\phi_i = \sigma \frac{d\rho_i}{dx} + \psi_i' \rho_i \ \text{ in } \Omega,\ i = 1, \ldots, n.$$

Our system may be written as the system of $2n$ ordinary differential equations, where (2.1) holds:

$$\begin{aligned} \sigma \frac{d\rho_i}{dx} &= \phi_i - \psi_i' \rho_i, \quad i = 1, \ldots, n, \\ \frac{d\phi_i}{dx} &= -\textstyle\sum_{j=1,\ldots,n} a_{ij} \rho_j, \quad i = 1, \ldots, n. \end{aligned} \tag{2.2}$$

Let Φ denote the $2n \times 2n$ fundamental solution matrix of (2.2) with $\Phi(0) = \mathbf{1}$. Let Ψ be the $2n \times n$ matrix consisting of the first n columns of Φ. Then

$$\Psi = \begin{pmatrix} R \\ S \end{pmatrix},$$

where R and S are $n \times n$ matrix functions with $R(0) = \mathbf{1}$ and $S(0) = 0$. We wish to obtain a solution

$$\begin{pmatrix} \rho \\ \phi \end{pmatrix} = \Phi c$$

such that $\phi(0) = \phi(1) = 0$. To have $\phi(0) = 0$, we need the last n components of c to be zero, so

$$\begin{pmatrix} \rho \\ \phi \end{pmatrix} = \Psi d,$$

where d is the vector consisting of the first n components of c. We then need the last n components of $\Psi(1)d$ to be zero, namely,

$$S(1)d = 0. \tag{2.3}$$

Now in this setup, we have $\phi_i(0) = 0,\ i = 1, \ldots, n$, for each column of S and from (2.1),

$$\sum_{i=1,\ldots,n} \frac{d\phi_i}{dx}(x) = 0,\ x \in \Omega,$$

whence

$$\sum_{i=1,\ldots,n} \phi_i(x) = 0,\ x \in \Omega.$$

But this simply means that for each j

$$\sum_{i=1,\ldots,n} S_{ij}(x) = 0,$$

so the sum of the rows of S is zero for every $x \in \Omega$, i.e., det $S(x) = 0$, and so S is singular. Hence we can find a (nontrivial) solution to (2.3).

Now we assume (1.2) and (1.3). If the solution is positive, it is the unique solution. This follows a standard argument. Suppose that ρ is a positive solution and that ρ^* is a second solution. Then $\rho+\mu\rho^*$ is a solution for any constant μ and $\rho+\mu\rho^*>0$ in Ω for sufficiently small $|\mu|$. Increase $|\mu|$ until we reach the first value for which some ρ_i has a zero, say at $x_0 \in \Omega$. For this value of i we have that for $f=\rho+\mu\rho^*$, f_i has a minimum at x_0 and

$$-\frac{d}{dx}\left(\sigma\frac{df_i}{dx}+\psi_i' f_i\right)-a_{ii}f_i=\sum_{\substack{j=1,\dots,n\\ j\neq i}} a_{ij}f_j \geqq 0, \tag{2.4}$$

$$\sigma\frac{df_i}{dx}+\psi_i' f_i=0. \tag{2.5}$$

By an elementary maximum principle [24] (cf. also [4]), we have that $f_i \equiv 0$.

We now claim that $f\equiv 0$. Choose any f_j and assume that it does not vanish identically. Using the maximum principle as before, $f_j>0$. Now choose a point x_0 such that $a_{ij}(x_0)>0$. Substituting onto (2.4) we now have a contradiction because $f_i\equiv 0$. Thus there is at most one solution satisfying (1.1).

It now remains to show that there is a positive solution. We employ a continuation argument. Note that there is a particular case where $\psi_i'(x)\equiv 0$ for all i and $a_{ii}(x)=1-n$, and $a_{ij}(x)=1$ for $j\neq i$. The solution in this case is $\rho_i(x)=\frac{1}{n}$, with our normalization in (1.1). For the moment, it is convenient to use a different normalization in terms of the vector d found above: choose the unique $d=(d_1,\dots,d_n)^T$ satisfying $\max_i d_i=1$.

For the special case above with

$$\psi_i'=0,\quad a_{ii}=1-n,\quad \text{and}\quad a_{ij}=1,\ i\neq j,$$

we find that $d=(1,\dots,1)^T$. To abbreviate the system in vector notation, let ψ_0' and ψ' be the diagonal matrices of potentials $\psi_i'=0$ and ψ_i', respectively, and let A_0 and A denote the matrices of lower order coefficients. For each λ, $0\leqq\lambda\leqq 1$, we solve the problem

$$\begin{aligned}\sigma\frac{d^2\rho}{dx^2}+\frac{d}{dx}((\lambda\psi'+(1-\lambda)\psi_0')\rho)+(\lambda A+(1-\lambda)A_0)\rho=0 \text{ in } \Omega,\\ \sigma\frac{d\rho}{dx}+(\lambda\psi'+(1-\lambda)\psi_0')\rho=0 \text{ at } x=0,1.\end{aligned} \tag{2.6}$$

For $\lambda=0$, (2.6) has a unique solution satisfying $\max_i\rho_i(0)=1$, and this solution is positive. As long as the solution is positive, the argument given above shows that it is unique. As we increase λ from 0, the solution is continuous as a function of λ, since the vector d will be continuous as long as it is unique.

Let Λ denote the subset of $\lambda\in[0,1]$ for which there is a positive solution of (2.6). To show that $\Lambda\subset[0,1]$ is open, consider $\lambda_0\in\Lambda$ and a sequence of points in Λ^c, the complement of Λ, convergent to λ_0. For each of these there is a nonpositive solution of (2.6), and we may assume that the initial conditions d are bounded. Hence a subsequence converges to the initial condition for a nonpositive solution with $\lambda=\lambda_0$, which contradicts the uniqueness of the positive solution.

To show Λ is closed, again suppose the contrary and that $\hat\lambda$ is a limit point of Λ not in Λ. Now some component $\hat\rho_i$ must have a zero, and $\hat\rho_i\geqq 0$ in Ω. Then by the maximum principle used above, $\hat\rho_i\equiv 0$. We now repeat the argument above to conclude that $\hat{\rho}_j \equiv 0$ in Ω for all $j = 1, \dots, n$. But this is impossible because we have imposed the condition that $\max_i \hat{\rho}_i(0) = 1$. This implies that Λ is open, so $\Lambda = [0, 1]$.

Renormalizing to obtain total mass 1 completes the proof.

Condition (1.3) is more restrictive than necessary for uniqueness and positivity of the solution. For an improved result, recall that $P_\tau = \mathbf{1} + \tau A$, $\tau > 0$ small, is a probability matrix when (1.2) is assumed. A probability matrix P is ergodic if some power P^k has all positive entries. In this case it has an eigenvector with eigenvalue 1 whose entries are positive, corresponding to a unique stationary state of the Markov chain it determines and other well-known properties from the Perron–Frobenius theory. Such matrices are often called irreducible and sometimes even "regular." We may now state an improvement of Theorem 2.1

THEOREM 2.2. *In Theorem* 2.1 *replace condition* (1.3) *with*

$$\int_0^1 P_\tau(x)dx \quad \textit{is ergodic.}$$

Then the conclusions of Theorem 2.1 *hold.*

We outline the changes which must be made to prove this result. The previous proof relied on showing that if for some i, $\rho_i \equiv 0$, then $\rho_j \equiv 0$ for every j. This followed from the maximum principle and the feature of the equations that each constituent was nontrivially represented near at least one point $x_0 \in \Omega$. But suppose that $a_{ij} \equiv 0$ for some j. In this case we could have $\rho_j > 0$, and this has no effect on ρ_i.

Under the assumption that $\int_0^1 P_\tau(x)dx$ is ergodic, some nondiagonal element in the ith row of A is not identically zero. This means that there is a $\pi(i) \neq i$ such that $\rho_i \equiv 0$ implies that $\rho_{\pi(i)} \equiv 0$. We may repeat this argument since ergodicity implies that the permutation π can be chosen so that $\pi^m(i)$ cycles around the entire set of integers $1, \dots, n$.

This completes the proof of existence of a unique solution of the stationary system with $\max_i d_i = 1$. This solution is also positive. Renormalizing to obtain total mass 1 completes the proof.

3. Transport. As we observed in the existence proof of the last section, the condition (1.1) implies that

$$\sum_{i=1,\dots,n} \frac{d}{dx}\left(\sigma \frac{d\rho_i}{dx} + \psi_i' \rho_i\right) = 0,$$

so that

$$\sum_{i=1,\dots,n} \left(\sigma \frac{d\rho_i}{dx} + \psi_i' \rho_i\right) = \gamma = \text{ const.}$$

In the case of interest of kinesin-type models, the boundary condition of (1.1) implies that $\gamma = 0$. In other words,

$$\sum_{i=1,\dots,n} \left(\sigma \frac{d\rho_i}{dx} + \psi_i' \rho_i\right) = 0. \tag{3.1}$$

A simulation of typical behavior in a two-species system is given in Figure 3.1.

THEOREM 3.1. *Suppose that ρ is a positive solution of* (1.1), *where the coefficients a_{ij}, $i, j = 1, \dots, n$, and the ψ_i, $i = 1, \dots, n$, are smooth and $1/N$-periodic in $\overline{\Omega}$. Suppose that* (1.2) *holds and also that the following conditions are satisfied:*

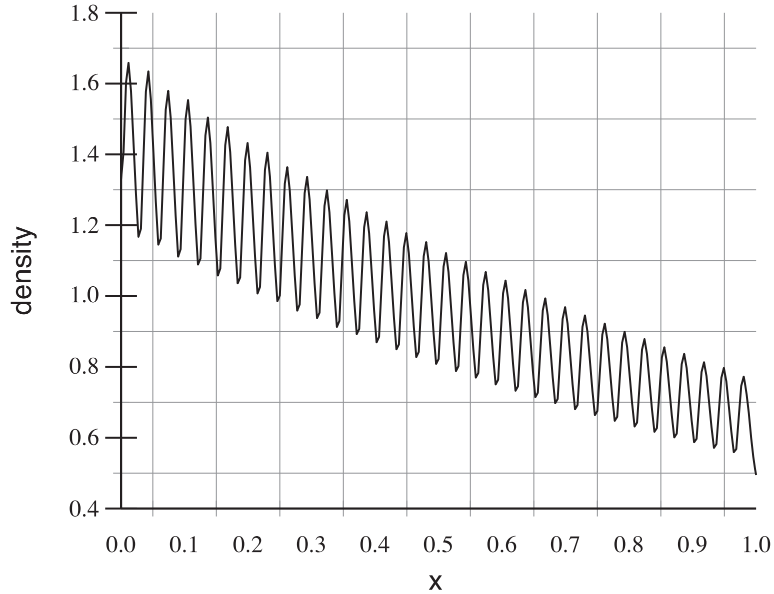


FIG. 3.1. *Transport in a two-species system with period* 16. *The abscissa shows the total density* $\rho_1 + \rho_2$. *In this computation,* ψ_2 *is a one-half period translate of* ψ_1, *and the support of the* $a_{ij}, i, j = 1, 2$, *is a neighborhood of the minima of the* $\psi_i, i = 1, 2$. *The simulation was executed with a semi-implicit scheme. Additional details are given at the conclusion of section* 5.

(i) *Each* ψ_i' *has only a finite number of zeros in* $\overline{\Omega}$.
(ii) *There is some interval in which* $\psi_i' > 0$ *for all* $i = 1, \dots, n$.
(iii) *In any interval in which no* ψ_i' *vanishes,* $\psi_j' > 0$ *in this interval for at least one* j.
(iv) *If* $I, |I| < 1/N$, *is an interval in which* $\psi_i' > 0$ *for* $i = 1, \dots, p$ *and* $\psi_i' < 0$ *for* $i = p+1, \dots, n$, *and* a *is a zero of at least one of the* ψ_k' *which lies within* ϵ *of the right-hand end of* I, *then for* ϵ *sufficiently small, there is at least one index* i, $i = 1, \dots, p$, *for which* $a_{ij} > 0$ *in* $(a - \eta, a)$ *for some* $\eta > 0$ *and for all* $j = p+1, \dots, n$.

Then there exist positive constants K, M *independent of* σ *and depending on the potentials* $\psi_i, i = 1, \dots, n$, *and the coefficients* $a_{ij}, i, j = i, \dots, n$, *such that*

$$\sum_{i=1}^{n} \rho_i \left(x + \frac{1}{N} \right) \leqq K e^{-\frac{M}{\sigma}} \sum_{i=1}^{n} \rho_i(x), \;\; x \in \Omega, \;\; x < 1 - \frac{1}{N}, \tag{3.2}$$

for sufficiently small σ.

Note that (3.1) holds under the hypotheses of the theorem. Also note that from (iv), where $a_{ij} > 0, j = p+1, \dots, n$, necessarily, $a_{ii} < 0$ according to (1.2). We shall prove Theorem 3.2 below. For this, it is convenient to consider a single-period interval rescaled to be $[0, 1]$. Theorem 3.1 then follows by rescaling and applying Theorem 3.2 to period intervals.

THEOREM 3.2. *Suppose that* ρ *is a positive solution of* (1.1), *where the coefficients* a_{ij}, $i, j = 1, \dots, n$, *and the* ψ_i, $i = 1, \dots, n$, *are smooth in* $[0, 1]$. *Suppose that* (1.2) *holds and also that the following conditions are satisfied:*

(i) *Each* ψ_i' *has only a finite number of zeros in* $[0, 1]$.

(ii) *There is some interval in which* $\psi_i' > 0$ *for all* $i = 1, \dots, n$.

(iii) *In any interval in which no* ψ_i' *vanishes,* $\psi_j' > 0$ *in this interval for at least one* j.

(iv) *If* I *is an interval in which* $\psi_i' > 0$ *for* $i = 1, \dots, p$ *and* $\psi_i' < 0$ *for* $i = p+1, \dots, n$, *and* a *is a zero of at least one of the* ψ_k' *which lies within* ϵ *of the right-hand end of* I, *then for* ϵ *sufficiently small, there is at least one* i, $i = 1, \dots, p$, *for which* $a_{ij} > 0$ *in* $(a - \eta, a)$ *for some* $\eta > 0$, $j = p+1, \dots, n$.

Then there exist positive constants K, M *independent of* σ *and depending on the potentials* $\psi_i, i = 1, \dots, n$, *and the coefficients* $a_{ij}, i, j = i, \dots, n$, *such that*

$$\sum_{i=1}^{n} \rho_i(1) \leqq K e^{-\frac{M}{\sigma}} \sum_{i=1}^{n} \rho_i(0) \tag{3.3}$$

for sufficiently small σ.

The conclusion of Theorem 3.2 is that the magnitude of the solution ρ, $\sum_{i=1}^{n} \rho_i$, is much smaller at $x = 1$ than at $x = 0$, or in terms of Theorem 3.1, that it is bounded above by an exponentially decreasing function for small σ. There is no suggestion that $\sum_{i=1}^{n} \rho_i$ is itself exponentially decreasing, and it is not. Indeed, the core of the mathematical argument is that $\sum \rho_i$ is exponentially decreasing on intervals where all ψ_i' are positive, while not significantly increasing in the remainder of [0,1]. The $\sum \rho_i$ may increase, even exponentially, in regions within δ of a zero of a ψ_i', but because the total length of these intervals is very small, the increase is outweighed by the decrease elsewhere. The argument in intervals where the signs of the ψ_i' are mixed is more delicate and relies on the coupling, as spelled out in (iv), the nonvanishing of some a_{ij} near the minima of ψ_i, and we briefly describe it below.

First let us assess how, essentially, the constants K, M depend on the parameters, especially the potentials ψ_i, by examining an interval where $\psi_i' > 0$ for all i. Such an interval exists by condition (ii) of the theorem. Let $[a, b]$ be such an interval and set

$$q(x) = \min_{i=1,\dots,n} \psi_i'(x).$$

From (3.1),

$$\frac{d}{dx}(\rho_1 + \cdots + \rho_n)(x) \leqq -\frac{1}{\sigma} q(x)(\rho_1 + \cdots + \rho_n)(x),$$

so that by integrating,

$$(\rho_1 + \cdots + \rho_n)(b) \leqq e^{-\frac{1}{\sigma}\int_a^b q\, ds}(\rho_1 + \cdots + \rho_n)(a).$$

If there are several such intervals, we just combine the effects, and this is the essence of how K, M (particularly M) depend on the ψ_i'. In other words, a Gronwall-type argument is successful here.

Now let us try to explain the role of the coupling. This comes into play when condition (iv) of the hypotheses holds. Suppose that $\nu \in \{1, \dots, p\}$ is a favorable index in the interval I and consider the νth equation

$$\sigma \rho_\nu'' + \psi_\nu' \rho_\nu' + \psi_\nu'' \rho_\nu + a_{\nu\nu}\rho_\nu + \sum_{j \neq \nu}^{n} a_{\nu j}\rho_j = 0. \tag{3.4}$$

Equation (3.4) represents a balance between ρ_ν and the other ρ_j. As seen in what follows, since the items in the Σ are nonnegative, they may be discarded and (3.4) can be employed to find an upper bound for ρ_ν on I, because ψ_ν is increasing. We can then exploit (3.4) to impede the growth of the unfavorable $\rho_j, j = p+1, \ldots, n$. Namely, $\{\rho_j\}$ cannot be too large without forcing ρ_ν to be negative. But this can be ensured only if the coupling is really there, namely, if $a_{\nu j} > 0$. This is the motivation for the ergodic-type hypothesis in (iv).

Proof of Theorem 3.2. Since each ψ_i' has only finitely many zeros, we can enclose these zeros with intervals of length 2δ, where $\delta > 0$ and small will be chosen later. The remainder of [0,1] consists of a finite number of closed intervals $J_m, m = 1, \ldots, M$, in which no ψ_i' vanishes, and so we have that $\psi_i' \geqq k(\delta) > 0$ or $\psi_i' \leqq -k(\delta) < 0$ for each i and some positive $k(\delta)$. From (iii), $k(\delta)$ may be chosen so that in at least one J_m, $\psi_i' \geqq k(\delta)$ for all i.

First we establish the exponential decay which governs the behavior of the solution. This will be a simple application of Gronwall's lemma. Consider an interval $I_0 = J_m$ for one of the m's where $\psi_i' \geq k(\delta)$ for all i. Suppose that

$$\sup_{I_0} \big\{ \inf_i \psi_i'(x) \big\} = K_0,$$

where, of course, K_0 is independent of δ for small δ. So there is a point $x_0 \in I_0$ where

$$\psi_i'(x_0) \geqq K_0, \quad i = 1, \ldots, n,$$

and

$$\psi_i'(x) \geqq \frac{1}{2} K_0, \quad |x - x_0| < L_0, \quad i = 1, \ldots, n,$$

for

$$L_0 = \frac{1}{2} \frac{K_0}{\max_{i=1\ldots n} \{\sup_{[0,1]} |\psi_i''|\}}.$$

Hence, from (3.1),

$$\frac{d}{dx} \sum_{i=1}^n \rho_i \leqq -\frac{K_0}{2\sigma} \sum_{i=1}^n \rho_i \quad \text{in } |x - x_0| < L_0,$$

so that

$$\sum_{i=1}^n \rho_i(x_0 + L_0) \leqq e^{-\frac{1}{\sigma} K_0 L_0} \sum_{i=1}^n \rho_i(x_0 - L_0).$$

Since $\sum_{i=1}^n \rho_i' \leqq 0$ in I_0, we have that

$$\left(\sum_{i=1}^n \rho_i \right)(\xi^*) \leqq e^{-\frac{1}{\sigma} K_0 L_0} \left(\sum_{i=1}^n \rho_i \right)(\xi), \quad \text{where } I_0 = [\xi, \xi^*]. \tag{3.5}$$

Indeed, we could extend I_0 to an interval in which we demand only that all $\psi_i' \geqq 0$.

Next consider an interval, say I_1, of length 2δ centered on a zero a of one of the ψ_i'. From (3.1) we have that

$$\left| \frac{d}{dx} \left(\sum_{i=1}^n \rho_i \right) \right| \leq \frac{K_1}{\sigma} \sum_{i=1}^n \rho_i \quad \text{in } I_1,$$

where

$$K_1 = \max_{i=1..n} \sup_{0 \leqq x \leqq 1} |\psi_i'|,$$

so that

$$\left(\sum_{i=1}^{n} \rho_i\right)(a+\delta) \leqq e^{\frac{1}{\sigma}2K_1\delta} \left(\sum_{i=1}^{n} \rho_i\right)(a-\delta). \tag{3.6}$$

There may be N such intervals, but over them all the exponential growth is only $\frac{2}{\sigma}NK_1\delta$, and we can choose δ sufficiently small, which does not affect K_0, L_0 so that

$$2NK_1\delta < K_0L_0.$$

Finally, with δ so chosen, we consider an interval $I_2 = [\alpha, \beta]$ where, say,

$$\begin{aligned} &\psi_i' \geqq k(\delta),\ \ i = 1, \dots, p, \\ &\psi_i' \leqq -k(\delta),\ \ i = p+1, \dots, n. \end{aligned} \tag{3.7}$$

We may assume that there is some overlap, that the endpoints α, β of I_2 are in 2δ intervals considered above. In the interval I_2, we shall bound $\rho_1, \dots, \rho_p$ on the basis of (3.7) above. We shall then argue that $\rho_{p+1}, \dots, \rho_n$ are necessarily bounded or, owing to the coupling of the equations, the positivity of $\rho_1, \dots, \rho_p$ would fail.

Write the equation for ρ_1 in the form

$$\sigma\rho_1'' + \psi_1'\rho_1' + \psi_1''\rho_1 + a_{11}\rho_1 + \sum_{j=2}^{n} a_{1j}\rho_j = 0, \tag{3.8}$$

so that

$$\frac{d}{dx}\left(\rho_1' e^{\frac{1}{\sigma}(\psi_1(x)-\psi_1(\alpha))}\right) = -\frac{1}{\sigma}\left\{(a_{11}+\psi_1'')\rho_1 + \sum_{j=2}^{n} a_{1j}\rho_j\right\} e^{\frac{1}{\sigma}(\psi_1(x)-\psi_1(\alpha))},$$

and carrying out the integration,

(3.9)

$$\rho_1'(x) = \rho_1'(\alpha)e^{\frac{1}{\sigma}(\psi_1(\alpha)-\psi_1(x)} - \frac{1}{\sigma}\int_\alpha^x \left\{(a_{11}+\psi_1'')\rho_1 + \sum_{j=2}^{n} a_{1j}\rho_j\right\} e^{\frac{1}{\sigma}(\psi_1(s)-\psi_1(x))} ds.$$

Now the $a_{1j}, j \geqq 2$, and the ρ_i are all nonnegative, so that we may neglect the large sum and find a constant K_2 for which

$$\rho_1'(x) \leqq \rho_1'(\alpha)e^{\frac{1}{\sigma}(\psi_1(\alpha)-\psi_1(x))} + \frac{K_2}{\sigma}\int_\alpha^x \rho_1(s)e^{\frac{1}{\sigma}(\psi_1(s)-\psi_1(x))} ds. \tag{3.10}$$

Note that for small σ,

$$\int_\alpha^x e^{\frac{1}{\sigma}(\psi_1(s)-\psi_1(x))} ds \leqq \int_\alpha^x e^{\frac{k(\delta)}{\sigma}(s-x)} ds \leqq \frac{\sigma}{k(\delta)}. \tag{3.11}$$

Integrating (3.10),

$$\begin{aligned}\rho_1(x) - \rho_1(\alpha) &\leqq \rho_1'(\alpha)\int_\alpha^x e^{\frac{1}{\sigma}(\psi_1(\alpha)-\psi_1(s))}ds + \frac{K_2}{\sigma}\int_\alpha^x\int_\alpha^t \rho_1(s)e^{\frac{1}{\sigma}(\psi_1(s)-\psi_1(t))}dsdt\\ &\leqq K(\delta)\sigma|\rho_1'(\alpha)| + K(\delta)\int_\alpha^x \max_{[\alpha,t]}\rho_1 dt,\end{aligned}$$

so that

$$\max_{[\alpha,x]}\rho_1 \leqq \rho_1(\alpha) + K(\delta)\sigma|\rho_1'(\alpha)| + K(\delta)\int_\alpha^x \max_{[\alpha,t]}\rho_1 dt.$$

We may now use Gronwall's lemma to obtain

$$\rho_1(x) \leqq K(\delta)\big\{\rho_1(\alpha) + \sigma|\rho_1'(\alpha)|\big\}, \quad \alpha \leqq x \leqq \beta. \tag{3.12}$$

If we insert this into (3.10), we obtain

$$\rho_1'(x) \leqq |\rho_1'(\alpha)| + K(\delta)\big\{\rho_1(\alpha) + \sigma|\rho_1'(\alpha)|\big\}, \quad \alpha \leqq x \leqq \beta. \tag{3.13}$$

Similar estimates hold for $\rho_2, \dots, \rho_p$.

Our attention is directed to $\rho_{p+1}, \dots, \rho_n$. Our first step is lower bounds for $\rho_1', \dots, \rho_p'$, for which it suffices to carry out the details for ρ_1'. We can use (3.12) to modify our formula (3.9). Using (3.11),

$$\begin{aligned}\rho_1'(x) \geqq \rho_1'(\alpha)e^{\frac{1}{\sigma}(\psi_1(\alpha)-\psi_1(x))} - \frac{1}{\sigma}\max_{I_2}\rho_1(\max_{I_2}|a_{11}+\psi_1''|)\int_\alpha^x e^{\frac{1}{\sigma}(\psi_1(s)-\psi_1(x))}ds\\ - \frac{1}{\sigma}\max_{[\alpha,x]}(\rho_2 + \cdots \rho_p + \rho_{p+1} + \cdots + \rho_n)\max_{1\leqq i\leqq n, i\neq 1}\max_{I_2} a_{1i}\int_\alpha^x e^{\frac{1}{\sigma}(\psi_1(s)-\psi_1(x))}ds.\end{aligned}$$

So, since by (3.12) we have bounds for $\rho_2, \dots, \rho_p$, we can write

$$\begin{aligned}\rho_1'(x) \geqq \rho_1'(\alpha)e^{\frac{1}{\sigma}(\psi_1(\alpha)-\psi_1(x))} - K(\delta)\big\{\rho_1(\alpha) + \sigma|\rho_1'(\alpha)|\big\}\\ - K(\delta)\max\nolimits_{[\alpha,x]}(\rho_{p+1} + \cdots + \rho_n), \quad \alpha \leqq x \leqq \beta.\end{aligned} \tag{3.14}$$

Similarly for $\rho_2', \dots, \rho_p'$.

With our technique we can handle only the sum $\rho_{p+1} + \cdots + \rho_n$ and not individual $\rho_i, p+1 \leqq i \leqq n$. From (3.1), and taking into account (3.12), (3.13), and the signs of the ψ_i',

$$\begin{aligned}\frac{d}{dx}(\rho_{p+1} + \cdots + \rho_n) &= -\frac{d}{dx}(\rho_1 + \cdots + \rho_p) - \frac{1}{\sigma}(\psi_1'\rho_1 \cdots + \psi_p'\rho_p)\\ &\quad - \frac{1}{\sigma}(\psi_{p+1}'\rho_{p+1} + \cdots + \psi_n'\rho_n)\\ &\geqq -\frac{K_1(\delta)}{\sigma}\sum_{i=1}^p\left(\rho_i(\alpha) + \sigma\left|\frac{d\rho_i}{dx}(\alpha)\right|\right)\\ &\quad + \frac{K_2(\delta)}{\sigma}\sum_{i=p+1}^n \rho_i \quad \text{in } I_2.\end{aligned} \tag{3.15}$$

Let

$$C(\alpha) = \sum_{i=1}^p\left(\rho_i(\alpha) + \left|\frac{d\rho_i}{dx}(\alpha)\right|\right),$$

which means (3.15) assumes a fortiori the form

$$\frac{d}{dx}(\rho_{p+1}+\cdots+\rho_n) \geqq -\frac{K_1(\delta)}{\sigma}C(\alpha)+\frac{K_2(\delta)}{\sigma}(\rho_{p+1}+\cdots+\rho_n). \tag{3.16}$$

We assert that

$$\rho_{p+1}+\cdots+\rho_n \leqq \frac{K_1(\delta)}{K_2(\delta)}C(\alpha), \quad \alpha \leqq x \leqq \beta-\delta. \tag{3.17}$$

Suppose the contrary and that

$$A > 1 \quad \text{and}$$

$$(\rho_{p+1}+\cdots+\rho_n)(x^*) > \frac{AK_1(\delta)C(\alpha)}{K_2(\delta)} \text{ for some } x^* \in [\alpha, \beta-\delta]. \tag{3.18}$$

This continues to hold in $[x^*, \beta]$, since at a first $x \in I_2$, where it fails, (3.16) would imply that $\rho_{p+1}+\cdots+\rho_n$ were increasing, which is not possible. Indeed, integrating (3.16) between points $x^*, x \in I_2$ with $x^* < x$, we have that

$$\begin{aligned}(\rho_{p+1}+\cdots+\rho_n)(x) &\geqq (\rho_{p+1}+\cdots+\rho_n)(x^*)e^{\frac{K_2(\delta)}{\sigma}(x-x^*)}\\ &\quad+\frac{K_1(\delta)}{K_2(\delta)}C(\alpha)(1-e^{\frac{K_2(\delta)}{\sigma}(x-x^*)})\\ &\geqq \frac{AK_1(\delta)}{K_2(\delta)}C(\alpha)e^{\frac{K_2(\delta)}{\sigma}(x-x^*)}+\frac{K_1(\delta)}{K_2(\delta)}C(\alpha)(1-e^{\frac{K_2(\delta)}{\sigma}(x-x^*)})\\ &= (A-1)\frac{K_1(\delta)}{K_2(\delta)}C(\alpha)e^{\frac{K_2(\delta)}{\sigma}(x-x^*)}+\frac{K_1(\delta)}{K_2(\delta)}C(\alpha), \quad x^*, x \in I_2. \end{aligned} \tag{3.19}$$

Now let us suppose, without loss of generality, that (iv) holds for $i = 1$, that is, we can find a $K_3(\delta) > 0$ such that

$$\sum_{i=p+1}^{n} a_{1j}\rho_j \geqq K_3(\delta)(\rho_{p+1}+\cdots+\rho_n) \quad \text{ in } [\beta-\delta, \beta]. \tag{3.20}$$

Keep in mind that

$$e^{\frac{K_2(\delta)}{\sigma}(x-x^*)} \geqq e^{\frac{K_2(\delta)}{\sigma}\frac{1}{4}\delta} = e^{\frac{K_4(\delta)}{\sigma}} \quad \text{for } \beta-\frac{1}{4}\delta \leqq x \leqq \beta.$$

Then we have from (3.9)

$$\begin{aligned}\frac{d\rho_1}{dx}(x) &\leqq \frac{d\rho_1}{dx}(\alpha)e^{-\frac{1}{\sigma}(\psi_1(x)-\psi_1(\alpha))}+\frac{K(\delta)}{\sigma}C(\alpha)\int_\alpha^x e^{\frac{1}{\sigma}(\psi_1(s)-\psi_1(x))}ds\\ &\quad-\frac{(A-1)K_1(\delta)K_3(\delta)C(\alpha)}{K_2(\delta)}e^{\frac{1}{\sigma}K_4(\delta)}\\ &\leqq \frac{d\rho_1}{dx}(\alpha)e^{-\frac{1}{\sigma}(\psi_1(x)-\psi_1(\alpha))}+K(\delta)C(\alpha)\\ &\quad-\frac{(A-1)K_1(\delta)K_3(\delta)C(\alpha)}{K_2(\delta)}e^{\frac{1}{\sigma}K_4(\delta)} \text{ for } \beta-\frac{1}{4}\delta \leqq x \leqq \beta.\end{aligned}$$

Above, $\psi_1(x) > \psi_1(\alpha)$, so the exponential in the first term on the right may be neglected. From the trivial inequality

$$\begin{aligned} 0 \leqq \rho_1(x) &= \rho_1(\alpha) + \int_\alpha^x \rho_1'(s)ds \\ &\leqq C(\alpha) + \int_\alpha^x \rho_1'(s)ds, \end{aligned}$$

we have that

$$\begin{aligned} 0 \leqq \frac{\rho_1(x)}{C(\alpha)} &\leqq 1 + \int_\alpha^x (1+K(\delta))ds - \int_{\beta-\frac{1}{4}\delta}^{\beta-\frac{1}{8}\delta} (A-1)\frac{K_1(\delta)}{K_2(\delta)}K_3(\delta)e^{\frac{K_4(\delta)}{\sigma}}ds \\ &\leqq 1 + (1+K(\delta))(\beta-\alpha) - (A-1)\frac{K_1(\delta)}{K_2(\delta)}K_3(\delta)e^{\frac{K_4(\delta)}{\sigma}}\frac{1}{8}\delta \qquad (3.21) \\ &\quad \text{for } \beta - \frac{1}{8}\delta \leqq x \leqq \beta. \end{aligned}$$

Since $A > 1$, the above cannot hold for arbitrarily small σ independent of δ because the extreme right-hand side of (3.21) becomes negatively infinite as $\sigma \to 0$. This proves (3.17). Note that the size of σ determined by (3.21) depends on the geometrical features of the potentials $\psi_i, i = 1, \dots, n$, but not on $C(\alpha)$, that is, the magnitude of the solution ρ.

The theorem now follows by concatenating the three cases.

4. Stability of the stationary solution. In this section we discuss the trend to stationarity of solutions of the time-dependent system (1.4). We have the following stability theorem.

THEOREM 4.1. *Let $\rho(x,t)$ denote a solution of* (1.4) *with initial data*

$$\rho(x,0) = f(x) \qquad (4.1)$$

satisfying

$$f_i(x) \geqq 0,\ i = 1,\dots,n,\ \textit{and} \sum_{i=1,\dots,n} \int_\Omega f_i dx = 1.$$

Then there are positive constants K and ω such that

$$|\rho(x,t) - \rho_0(x)| \leqq Ke^{-\omega t} \text{ as } t \to \infty, \qquad (4.2)$$

where ρ_0 is the stationary positive solution obtained in Theorem 2.1.

Thus the stationary positive solution is globally stable. One proof of this was given in [4] for $n = 2$, and this proof may be extended to general n. A proof based on monotonicity of an entropy function is given in [22]. A different type of monotonicity result showing that the solution operator is an L^1-contraction is given in [11]. Here we outline a different way of viewing the problem based on inspection of the semigroup generated by the operator, written in vector form,

$$S\rho = \sigma\frac{\partial^2\rho}{\partial x^2} + \frac{\partial}{\partial x}(\psi'\rho) + A\rho, \qquad (4.3)$$

with natural boundary conditions. All the methods known to us are based on ideas from positive operators via Perron–Frobenius–Krein–Rutman generalizations or on closely related monotonicity methods.

We need the result that (4.3) has a real eigenvalue λ_0, which is simple and has an associated positive eigenfunction, and that all other eigenvalues λ satisfy Re $\lambda < \lambda_0$.

This is a standard result (see, for example, Zeidler [30]) obtained using the ideas of positive operators, but to assist the reader and for completeness we give a sketch of the proof. We define e^S by writing the solution of (1.4) in terms of (4.3) as

$$\rho(x,t) = e^{tS} f(x). \tag{4.4}$$

This is consistent with the notions of exponent, since

$$e^{(t+s)S} = e^{tS} e^{sS}$$

just expresses the fact that the solution at time $t+s$ is just the solution after time s followed by a further time t. We note that e^S is a positive operator, since $f \geqq 0$ implies $\rho \geq 0$, making use of the maximum principle [24], and it is compact, since it is essentially an integration. And so e^S has a real eigenvalue, e^{λ_0}, which is simple and has a positive eigenfunction. Further, it is simple to see, for example, by solving the equation explicitly, that if S has an eigenvalue λ, then e^S has an eigenvalue e^λ, and vice versa. Thus S has a real eigenvalue λ_0, which is simple and has a positive eigenfunction. Further, the fact that all the other eigenvalues e^λ of e^S have $|e^\lambda| < e^{\lambda_0}$ implies that all other eigenvalues λ of S have Re $\lambda < \lambda_0$, as required.

We assume that the positive initial data f is normalized so that

$$\sum_{i=1,\ldots,n} \int_\Omega f_i dx = 1, \tag{4.5}$$

as in (1.1).

Now form the Laplace transform

$$\hat{\rho}\,(x,\lambda) = \int_0^\infty e^{-\lambda t} \rho\,(x,t)\,dt, \quad \mathrm{Re}\,\lambda > 0,$$

and (4.4) gives

$$\hat{\rho}\,(\cdot,\lambda) = (\lambda I - S)^{-1} f, \tag{4.6}$$

and $\hat{\rho}$ is analytic in λ for $\mathrm{Re}\,\lambda > 0$, but (4.6) allows us to extend this into the left half plane except for an isolated singularity at $\lambda = 0$, for in our problem the fact that we have a positive stationary solution implies that the real eigenvalue λ_0 is given by $\lambda_0 = 0$. The usual inversion formula gives

$$\rho(x,t) = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} e^{\lambda t} \hat{\rho}(x,\lambda) d\lambda, \ \gamma > 0. \tag{4.7}$$

Now for finite γ and ν, with ν large, and with $\lambda = \gamma + i\nu$,

$$\|(\lambda I - S)^{-1}\| = O\left(\frac{1}{\nu}\right). \tag{4.8}$$

For we can write

$$\begin{aligned} -(T\rho)_i &= \sigma \frac{d^2 \rho_i}{dx^2} + \frac{d}{dx}(\psi_i' \rho_i) + (A\rho)_i \\ &= \sigma \frac{d}{dx}(e^{-\frac{1}{\sigma}\psi_i} \phi_i') - (a'\phi)_i \\ &= -(M\phi)_i, \end{aligned}$$

where $\phi_i = e^{\frac{1}{\sigma}\psi_i}\rho_i$ and $a'_{ij} = a_{ij}e^{-\frac{1}{\sigma}\psi_i}$. The boundary conditions here are

$$\phi_i'(0) = \phi_i'(1) = 0,\ i = 1, \ldots, n.$$

Now consider

$$\lambda I - S = \lambda I + T = \lambda I + M$$

and

$$((M + \lambda I)\phi, \phi) = \sum_{i=1,\ldots,n} \int_\Omega [(M + \lambda I)\phi]_i \,\bar{\phi}_i dx. \tag{4.9}$$

The first term in M is self-adjoint, which gives a real contribution, and the second contribution in (4.9) is certainly

$$O\left\{\sum_{i=1,\ldots,n} \int_\Omega |\phi_i|^2 dx\right\}.$$

Thus if ν is sufficiently large,

$$im((M + \lambda I)\phi, \phi) \geqq \frac{1}{2}\nu \sum_{i=1,\ldots,n} \int_\Omega |\phi_i|^2 dx,$$

from which (4.8) follows easily.

Given the initial data f, we write

$$f = c\rho_0 + \rho^*,$$

where ρ^* is orthogonal to the positive eigenfunction, say ρ_0^*, of the adjoint operator of S. This determines c uniquely, since

$$c(\rho_0, \rho_0^*) = (f, \rho_0^*),$$

and $(\rho_0, \rho_0^*) \neq 0$ inasmuch as $\rho_{0,i} > 0, \rho_{0,i}^* > 0$. Then by the Fredholm alternative, we can solve, for any small λ,

$$(S - \lambda I)\phi = \rho^*,$$

uniquely if we insist that the solution is orthogonal to ρ_0. Then $(S - \lambda I)^{-1}\rho^*$ is bounded, and

$$\begin{aligned}(S - \lambda I)^{-1} f &= c\,(S - \lambda I)^{-1}\rho_0 + O\,(1)\\ &= -\lambda^{-1}c\rho_0 + O\,(1)\end{aligned}$$

as $\lambda \to 0$, showing that the pole of $(S - \lambda I)^{-1}$ at $\lambda = 0$ has residue $-c\rho_0$ so that we can now move the line of integration in (4.7) from $\gamma > 0$ to $\gamma < 0$. The contribution from the pole is $c\rho_0$, and the contribution from large λ is small by (4.8). Further, once this move is made, the contribution from the vertical line is of the form $O(e^{-\omega t})$. In all, therefore,

$$\rho\,(x, t) = c\rho_0 + O\left(e^{-\omega t}\right),$$

as required. Note that $c = 1$ since

$$\sum_{i=1,\ldots,n} \int_\Omega \rho_i dx = \sum_{i=1,\ldots,n} \int_\Omega \rho_{0,i} dx.$$

5. Discussion and some conditions for "reverse" transport. We now investigate what may happen if the conditions on the ψ_i in Theorem 3.1 are satisfied but those on the a_{ij} are not. In particular, we note that condition (iv) of Theorem 3.1 requires that if a is the minimum of one of the ψ_k, then some a_{ij} has support containing an interval $(a-\eta, a)$ to the left of a. We will show that without this condition, a_{ij} can be found such that the direction of transport is in the opposite direction from that described in Theorem 3.1. We remark that the necessity of some positivity condition on the a_{ij} to get transport is obvious, for if the a_{ij} are all identically zero, for example, or satisfy conditions that permit the functional F of the introduction to be minimized, then the solutions of (1.1) are periodic. What we look for in the following example is a situation in which there is transport, but in the opposite direction from that predicted by Theorem 3.1 even though the conditions on the ψ_i in that theorem are satisfied.

In constructing our example, we will specialize to $n = 2$, a two-state system. We also reverse direction. By this we mean that conditions (ii) and (iii) in Theorem 3.1, for $n = 2$, will be replaced by the following:

(ii′) *There is some interval in which $\psi_i' < 0$ for $i = 1, 2$.*

(iii′) *In any interval in which neither ψ_i' vanishes, $\psi_j' < 0$ in this interval for at least one j.*

For (iv) we substitute a simpler condition which could be used in Theorem 3.1 as well, as it implies (iv).

(iv′) *There is a neighborhood of each local minimum of ψ_1 or ψ_2 in which $a_{ij} \neq 0$ for all (i, j).*

Figure 5.1 shows potentials satisfying (ii′) and (iii′).

We then have the following corollary of Theorem 3.1.

COROLLARY 5.1. *If the hypotheses of Theorem* 3.1 *when $n = 2$ are satisfied, except that* (ii′), (iii′), *and* (iv′) *replace* (ii), (iii), *and* (iv), *then there exist constants K_1, K_2 independent of σ such that*

$$\sum_{i=1}^{2} \rho_i(x) \leq K_1 e^{-\frac{K_2}{\sigma}} \sum_{i=1}^{2} \rho_i\left(x + \frac{1}{N}\right) \text{ for } x \in \Omega,\ x \leq 1 - \frac{1}{N}.$$

We will now construct an example for the case $N = 1$ where conditions (i), (ii′), and (iii′) are satisfied, but not condition (iv′). Figure 5.1 shows conformation coefficients which do not satisfy (iv), and Figure 5.2 shows the resulting solution. Our example is constructed initially using δ-functions for the a_{ij}, and with this class of rate coefficients we are able to show that there is a $c > 0$ such that for sufficiently small σ,

$$\rho_1(1) + \rho_2(1) < e^{-\frac{c}{\sigma}}(\rho_1(0) + \rho_2(0)). \tag{5.1}$$

At the end we briefly discuss a slightly weaker form of reverse transport which we can then obtain for continuous coefficients.

Assume that ψ_1 has a minimum at $y_1 = 0$ followed by a maximum at $z_1 \in (0, 1)$ and then a second minimum at 1, with $\psi_1(0) = \psi_1(1)$. Further assume that ψ_2 has a minimum at $y_2 \in (z_1, 1)$ followed by a maximum at $z_2 \in (y_2, 1)$, and $\psi_2(0) = \psi_2(1)$. Finally assume that $\psi_i' \neq 0$ except at the minima and maxima specified above. Then $0 = y_1 < z_1 < y_2 < z_2 < 1$. There is no point where both $\psi_1' \geqq 0$ and $\psi_2' \geqq 0$, and so when the a_{ij} are all nonzero on $[0, 1]$, transport will be to the right as given in Corollary 5.1.

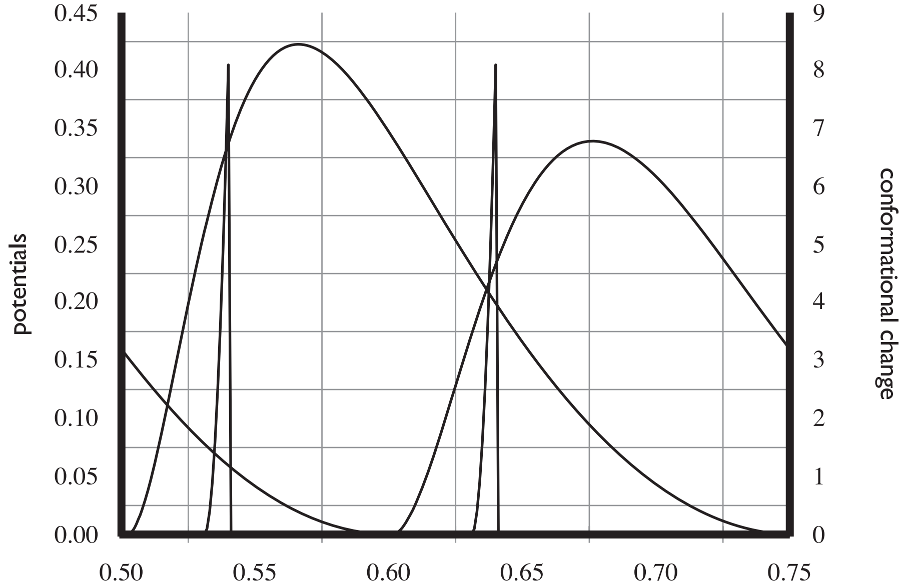


Fig. 5.1. *A period interval showing potentials and conformation coefficients which do not satisfy hypothesis* (iv′) *of Corollary* 3.2.

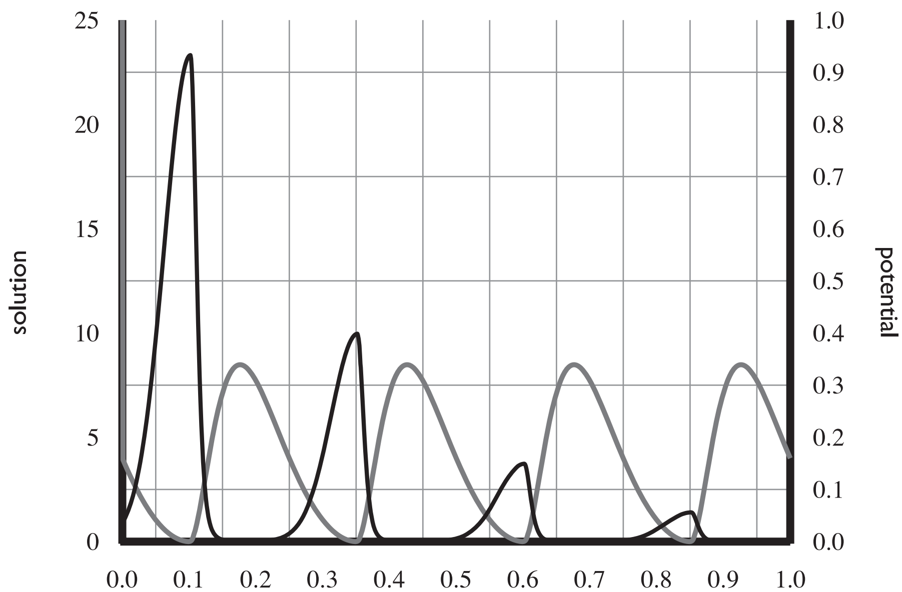


Fig. 5.2. *Reverse transport computed using the potentials and conformation coefficients shown in Figure* 5.1. *The simulation was done with XPP* [7].

But we will now choose new a_{ij} to give transport to the left.

Obviously, condition (iv′) must be violated. Choose a point $x_1 \in (y_1, z_1)$ and a point $x_2 \in (y_2, z_2)$. Then

$$0 = y_1 < x_1 < z_1 < y_2 < x_2 < z_2 < 1. \tag{5.2}$$

We consider the system

$$
\begin{aligned}
&(\sigma\rho_1' + \psi_1'\rho_1)' = (\delta(x-x_1) + \delta(x-x_2))(\rho_1 - \rho_2), \\
&(\sigma\rho_2' + \psi_2'\rho_2)' = (\delta(x-x_1) + \delta(x-x_2))(\rho_2 - \rho_1),
\end{aligned} \tag{5.3}
$$

with boundary conditions

$$
\sigma\rho_i' + \psi_i'\rho_i = 0 \text{ at } x = 0, 1 \text{ for } i = 1, 2. \tag{5.4}
$$

We wish to find further conditions which imply inequality (5.1) for some $c > 0$ and sufficiently small σ.

We follow the technique in [4] and let $\phi_i = \sigma\rho_i' + \psi_1'\rho_i$. Adding the equations in (5.3) shows that $\phi_1 + \phi_2$ is constant, and applying the boundary conditions shows that $\phi_1 + \phi_2 = 0$. This leads to the system

$$
\begin{aligned}
\sigma\rho_1' &= \phi - \psi_1'\rho_1, \\
\sigma\rho_2' &= -\phi - \psi_2'\rho_2, \\
\phi' &= (\delta(x-x_1) + \delta(x-x_2))(\rho_1 - \rho_2),
\end{aligned} \tag{5.5}
$$

where $\phi = \phi_1 = -\phi_2$.

Having obtained (5.5) under the conditions $\phi_i = 0$ at $x = 0, 1$, we now weaken these conditions, assuming only that $\phi_1 + \phi_2 = 0$. In this way, the same analysis applies to any period interval of the functions ψ_i, thus showing that if $N > 1$, decay occurs in each period interval. Therefore in (5.5) it is not assumed that $\phi(0)$ or $\phi(1)$ vanish. The only assumption made is that $\rho_i > 0$ on the entire interval, for $i = 1, 2$.

Observe that ϕ takes a jump of amount $\rho_1(x_j) - \rho_2(x_j)$ at each x_j. Further, ϕ is constant in the intervals $[0, x_1), (x_1, x_2), (x_2, 1]$. Let $\phi_j = \phi(y_j)$. Then

$$
\rho_i(x_j) = \rho_i(y_j)\, e^{\frac{\psi_i(y_j) - \psi_i(x_j)}{\sigma}} + (-1)^{i-1}\phi_j \int_{y_j}^{x_j} \frac{1}{\sigma} e^{\frac{\psi_i(s) - \psi_i(x_j)}{\sigma}} ds,
$$

$i = 1, 2$. Hence,

$$
\begin{aligned}
\rho_1(x_j) - \rho_2(x_j) = {} & \rho_1(y_j)\, e^{\frac{\psi_1(y_j) - \psi_1(x_j)}{\sigma}} - \rho_2(y_j)\, e^{\frac{\psi_2(y_j) - \psi_2(x_j)}{\sigma}} \\
& + \phi_j \int_{y_j}^{x_j} \frac{1}{\sigma} \left(e^{\frac{\psi_1(s) - \psi_1(x_j)}{\sigma}} + e^{\frac{\psi_2(s) - \psi_2(x_j)}{\sigma}} \right) ds.
\end{aligned}
$$

For $i = 1, 2$ let

$$
\begin{aligned}
&a_i = \frac{\psi_i(y_1)}{\sigma},\ b_i = \frac{\psi_i(x_1)}{\sigma},\ c_i = \frac{\psi_i(x_2)}{\sigma}, \\
&A_i = \int_0^{x_1} \frac{1}{\sigma} e^{\frac{\psi_i(s)}{\sigma}} ds,\ B_i = \int_{x_1}^{x_2} \frac{1}{\sigma} e^{\frac{\psi_i(s)}{\sigma}} ds,\ C_i = \int_{x_2}^{1} \frac{1}{\sigma} e^{\frac{\psi_i(s)}{\sigma}} ds.
\end{aligned} \tag{5.6}
$$

Since $\psi_i(0) = \psi_i(1)$, we eventually obtain (computation facilitated by Maple)

$$
\begin{aligned}
\rho_1(1) &= k_{11}\rho_1(0) - k_{12}\rho_2(0) + k_{13}\phi(0), \\
\rho_2(1) &= -k_{21}\rho_1(0) + k_{22}\rho_2(0) - k_{23}\phi(0),
\end{aligned}
$$

where

$$k_{11} = 1 + e^{-b_1}B_1 + e^{-b_1}C_1 + e^{-c_1}C_1 + e^{-b_1-c_1}B_1C_1 + e^{-b_1-c_2}B_2C_1$$

and $k_{12}, \ldots, k_{23}$ are similar expressions in terms of the constants defined in (5.6).

As in [4], we solve each of the inequalities $\rho_1(1) > 0$, $\rho_2(1) > 0$ for $\phi(0)$ and substitute the result into the other of these two relations. We find that

$$\rho_1(1) \le \frac{k_{11}k_{23} - k_{21}k_{13}}{k_{23}}\rho_1(0) + \frac{k_{13}k_{22} - k_{12}k_{23}}{k_{23}}\rho_2(0),$$

$$\rho_2(1) \le \frac{k_{11}k_{23} - k_{13}k_{21}}{k_{13}}\rho_1(0) + \frac{k_{13}k_{22} - k_{12}k_{23}}{k_{13}}\rho_2(0).$$

The desired decay relation (5.1) follows by showing that under certain additional conditions the four fractional coefficients

$$\frac{k_{11}k_{23} - k_{21}k_{13}}{k_{13}}, \quad \frac{k_{11}k_{23} - k_{21}k_{13}}{k_{23}}, \quad \frac{k_{13}k_{22} - k_{12}k_{23}}{k_{13}}, \quad \frac{k_{13}k_{22} - k_{12}k_{23}}{k_{23}} \tag{5.7}$$

tend to zero exponentially as $\sigma \to 0$.

Further Maple computation (checked with Scientific Workplace) shows, for example, that

$$\begin{aligned} k_{11}k_{23} - k_{21}k_{13} &= \frac{A_2}{e^{a_2}} + \frac{B_2}{e^{a_2}} + \frac{C_2}{e^{a_2}} + A_2\frac{B_1}{e^{a_2}e^{b_1}} + A_2\frac{C_1}{e^{a_2}e^{b_1}} + A_2\frac{B_2}{e^{a_2}e^{b_2}} + A_2\frac{C_1}{e^{a_2}e^{c_1}} \\ &+ A_2\frac{C_2}{e^{a_2}e^{b_2}} + B_2\frac{C_1}{e^{a_2}e^{c_1}} + A_2\frac{C_2}{e^{a_2}e^{c_2}} + B_2\frac{C_2}{e^{a_2}e^{c_2}} + A_2B_1\frac{C_1}{e^{a_2}e^{b_1}e^{c_1}} + A_2B_1\frac{C_2}{e^{a_2}e^{b_1}e^{c_2}} \\ &+ A_2B_2\frac{C_1}{e^{a_2}e^{b_2}e^{c_1}} + A_2B_2\frac{C_2}{e^{a_2}e^{b_2}e^{c_2}}. \end{aligned}$$

Many cancellations have occurred, eliminating terms in which four or five integrals are multiplied. Similar formulas are obtained for the other expressions in (5.7).

In estimating the integrals, first consider B_1. We will say that $f \propto g$ if there are positive numbers α and β such that for sufficiently small σ, $\alpha < \frac{f}{g} < \beta$. We then have

$$B_1 = \int_{x_1}^{x_2} e^{\frac{\psi_1(s)}{\sigma}}ds \propto \sigma^k e^{\frac{\psi_{1,\max}}{\sigma}}$$

for some $k > 0$ and with $\psi_{1,\max} = \psi_1(z_1) = \max_x \psi_1(x)$. Also, for possibly different values of k,

$$\begin{aligned} &A_1 \propto \sigma^k e^{b_1},\ A_2 \propto \sigma^k e^{a_2}, \\ &B_2 \propto \sigma^k \left(e^{b_2} + e^{c_2}\right),\ C_1 \propto \sigma^k \left(e^{c_1} + e^{a_1}\right),\ C_2 \propto \sigma^k e^{\frac{\psi_{2,\max}}{\sigma}}. \end{aligned}$$

From (5.6) and (5.2), we see that for small σ the two largest terms among $A_1, A_2, B_1, B_2, C_1, C_2$, $e^{a_1}, e^{a_2}, e^{b_1}, e^{b_2}, e^{c_1}$, and e^{c_2} are B_1 and C_2.

For the moment we let $d_i = \frac{\psi_{i,\max}}{\sigma}$ and set

$$\begin{aligned} &A_1 = e^{b_1},\ A_2 = e^{a_2}, \\ &B_1 = e^{d_1},\ B_2 = e^{b_2} + e^{c_2}, \\ &C_1 = e^{c_1} + e^{a_1},\ C_2 = e^{d_2}. \end{aligned}$$

We will find also that to get the desired backward transport, we need to take x_2 close to the maximum of ψ_2. Therefore for now we will set $x_2 = z_2$, so that $c_2 = d_2$. Finally, we can assume without loss of generality that $a_1 = 0$. The additional conditions we will give for backwards transport for small σ are that the inequalities (5.8) and (5.9) below hold and that x_2 is sufficiently close to y_2.

We then have

$$\begin{aligned} k_{11}k_{23} - k_{21}k_{13} = {}& \frac{1}{e^{b_1}} + \frac{2}{e^{c_1}} + \frac{3}{e^{a_2}}e^{b_2} + \frac{1}{e^{b_1}}e^{c_1} + \frac{3}{e^{b_1}}e^{d_1} + \frac{4}{e^{a_2}}e^{d_2} + \frac{4}{e^{b_2}}e^{d_2} \\ & + \frac{1}{e^{a_2}}\frac{e^{b_2}}{e^{c_1}} + \frac{1}{e^{b_1}e^{c_1}}e^{d_1} + \frac{1}{e^{a_2}e^{c_1}}e^{d_2} + \frac{1}{e^{b_2}e^{c_1}}e^{d_2} + 6 \end{aligned}$$

and similar expressions for $k_{22}k_{13} - k_{12}k_{23}$, k_{13}, and k_{23}.

We now assume that $d_1 > \max\{b_1, c_1\}$, $a_1 = 0 < \min\{b_1, c_1\}$, and $d_2 > \max\{a_2, b_2\}$. We compare terms pairwise wherever possible, eliminating the term which is necessarily smaller as $\sigma \to 0$. This results in the asymptotic relations

$$\begin{aligned} k_{11}k_{23} - k_{21}k_{13} &\propto \frac{3}{e^{b_1}}e^{d_1} + \frac{4}{e^{a_2}}e^{d_2} + \frac{4}{e^{b_2}}e^{d_2}, \\ k_{22}k_{13} - k_{12}k_{23} &\propto 6e^{d_1} + 4\frac{e^{b_1}}{e^{b_2}}e^{d_2}, \\ k_{13} &\propto 4e^{d_1} + 2\frac{e^{a_2}}{e^{b_2}}e^{d_1}, \\ k_{23} &\propto \frac{2}{e^{a_2}e^{c_1}}e^{d_1}e^{d_2} + \frac{1}{e^{b_2}e^{c_1}}e^{d_1}e^{d_2}. \end{aligned}$$

From these we conclude that the four fractions in question are exponentially small as $\sigma \to 0$ if in addition to the previous assumptions we have

$$d_2 - a_2 < d_1 - b_1 < d_2 - b_2 \tag{5.8}$$

and

$$d_1 > b_1 + c_1. \tag{5.9}$$

(If $a_1 \neq 0$, this becomes $d_1 + a_1 > b_1 + c_1$.)

By continuity we see that these inequalities will also suffice if c_2 is sufficiently close to d_2. The conclusions also hold with the factors σ^k included in the asymptotic expressions, since these don't affect the exponential limits.

Finally we wish to obtain a result with continuous functions for the a_{ij}. Here we don't have a limit result as $\sigma \to 0$. But suppose that $\varepsilon > 0$ is given, and for (5.3)–(5.4), we choose σ so small that for any positive solution,

$$\sum_{i=1}^{n} \rho_i(1) < \varepsilon \sum_{i=1}^{n} \rho_i(0).$$

Then for this σ, the same inequality will hold for continuous functions a_{ij} sufficiently close in the $\mathcal{L}_1$ norm to the δ-functions in (5.3)–(5.4).

In this paragraph we discuss the simulation parameters for Figure 3.1. Simulations, of which this is a sample, were executed with a semi-implicit scheme and run in Maple. In this case, for potentials we took $\psi_1(x) = \psi_1^0(2^4x)$ and $\psi_2(x) = \psi_2^0(2^4x)$ with $\psi_1^0(\xi) = \frac{1}{4}(\cos(\pi(\frac{2\xi}{\xi+1})))^2$ and $\psi_2^0(\xi) = \psi_1^0(\xi - \frac{1}{2})$. The matrix elements were $-a_{11} = a_{12} = a_{21} = -a_{22}$ with $a_{12}(x) = a^0(2^4x)$, where $a^0(\xi) = \frac{1}{2}(\cos 2\pi(\xi - \frac{1}{4}))^6$. The diffusion constant $\sigma = 2^{-7}$.

Acknowledgments. We take this opportunity to thank our collaborators Michel Chipot and Michal Kowalczyk for their help. We also thank Bard Ermentrout, Michael Grabe, William Troy, and Jonathan Rubin, and the Mathematical Biology Seminar of the University of Pittsburgh for generosity of spirit and their interest. We thank the referees for their comments.

REFERENCES

[1] A. Ajdari and J. Prost, *Mouvement induit par un potentiel périodique de basse symétrie: Dielectrophorese pulse*, C. R. Acad. Sci. Paris Sér. II, 315 (1992), pp. 1653–1657.

[2] R. D. Astumian, *Thermodynamics and kinetics of a Brownian motor*, Science, 276 (1997), pp. 917–922.

[3] J.-D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numer. Math., 84 (2000), pp. 375–393.

[4] M. Chipot, S. Hastings, and D. Kinderlehrer, *Transport in a molecular motor system*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 1011–1034.

[5] M. Chipot, D. Kinderlehrer, and M. Kowalczyk, *A variational principle for molecular motors, Meccanica*, 38 (2003), pp. 505–518.

[6] C. Doering, B. Ermentrout, and G. Oster, *Rotary DNA motors*, Biophys. J., 69 (1995), pp. 2256–2267.

[7] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, SIAM, Philadelphia, 2002.

[8] D. D. Hackney, *The kinetic cycles of myosin, kinesin, and dynein*, Ann. Rev. Physiol., 58 (1996), pp. 731–750.

[9] S. Hastings and D. Kinderlehrer, *Remarks about diffusion mediated transport: Thinking about motion in small systems*, in Variational Analysis and Applications, Nonconvex Optim. Appl. 79, Springer, New York, 2005, pp. 497–511.

[10] S. Hastings, D. Kinderlehrer, and J. B. McLeod, *Diffusion mediated transport with a look at motor proteins*, to appear.

[11] D. Hilhorst et al., to appear.

[12] J. Howard, *Mechanics of Motor Proteins and the Cytoskeleton*, Sinauer Associates, Sunderland, MA, 2001.

[13] A. F. Huxley, *Muscle structure and theories of contraction*, Prog. Biophys. Biophys. Chem., 7 (1957), pp. 255–318.

[14] R. Jordan, D. Kinderlehrer, and F. Otto, *The variational formulation of the Fokker–Planck equation*, SIAM J. Math. Anal., 29 (1998), pp. 1–17.

[15] D. Kinderlehrer and M. Kowalczyk, *Diffusion-mediated transport and the flashing ratchet*, Arch. Ration. Mech. Anal., 161 (2002), pp. 149–179.

[16] D. Kinderlehrer and N. Walkington, *Approximation of parabolic equations based upon Wasserstein's variational principle*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 837–852.

[17] Y. Okada and N. Hirokawa, *A processive single-headed motor: Kinesin superfamily protein KIF1A*, Science, 283 (1999), pp. 1152–1157.

[18] Y. Okada and N. Hirokawa, *Mechanism of the single headed processivity: Diffusional anchoring between the K-loop of kinesin and the C terminus of tubulin*, Proc. Natl. Acad. Sci., 7 (2000), pp. 640–645.

[19] F. Otto, *Dynamics of labyrinthine pattern formation: A mean field theory*, Arch. Ration. Mech. Anal., 141 (1998), pp. 63–103.

[20] F. Otto, *The geometry of dissipative evolution equations: The porous medium equation*, Comm. Partial Differential Equations, 26 (2001), pp. 101–174.

[21] A. Parmeggiani, F. Jülicher, A. Ajdari, and J. Prost, *Energy transduction of isothermal ratchets: Generic aspects and specific examples close and far from equilibrium*, Phys. Rev. E, 60 (1999), pp. 2127–2140.

[22] B. Perthame, *The general relative entropy principle—applications in Perron-Frobenius and Floquet theories and a parabolic system for biomotors*, Rend. Accad. Naz. Sci. XL Mem. Mat. Appl. (5), 29 (2005), pp. 307–325.

[23] C. S. Peskin, G. B. Ermentrout, and G. F. Oster, *The correlation ratchet: A novel mechanism for generating directed motion by ATP hydrolysis*, in Cell Mechanics and Cellular Engineering, V. C. Mow et al., eds., Springer, New York, 1994, pp. 479–489.

[24] M. Protter and H. Weinberger, *Maximum Principles in Differential Equations*, Prentice–

Hall, Englewood Cliffs, NJ, 1967.
[25] E. M. Purcell, *Life at low Reynolds' number*, Amer. J. Phys., 45 (1971), pp. 3–11.
[26] P. Reimann, *Brownian motors: Noisy transport far from equilibrium*, Phys. Rep., 361 (2002), pp. 57–265.
[27] M. Schliwa, ed., *Molecular Motors*, Wiley-VCH Verlag, Wennheim, 2003.
[28] R. D. Vale and R. A. Milligan, *The way things move: Looking under the hood of motor proteins*, Science, 288 (2000), pp. 88–95.
[29] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, AMS, Providence, RI, 2003.
[30] E. Zeidler, *Nonlinear Functional Analysis and Its Applications,* I, Springer, New York, 1986.

SIAM J. MATH. ANAL.
Vol. 39, No. 4, pp. 1231–1259

SYNCHRONIZATION OF DISCRETE-TIME DYNAMICAL NETWORKS WITH TIME-VARYING COUPLINGS*

WENLIAN LU[†‡], FATIHCAN M. ATAY[†], AND JÜRGEN JOST[†§]

Abstract. We study the local complete synchronization of discrete-time dynamical networks with time-varying couplings. Our conditions for the temporal variation of the couplings are rather general and include variations in both the network structure and the reaction dynamics; the reactions could, for example, be driven by a random dynamical system. A basic tool is the concept of the Hajnal diameter, which we extend to infinite Jacobian matrix sequences. The Hajnal diameter can be used to verify synchronization, and we show that it is equivalent to other quantities which have been extended to time-varying cases, such as the projection radius, projection Lyapunov exponents, and transverse Lyapunov exponents. Furthermore, these results are used to investigate the synchronization problem in coupled map networks with time-varying topologies and possibly directed and weighted edges. In this case, the Hajnal diameter of the infinite coupling matrices can be used to measure the synchronizability of the network process. As we show, the network is capable of synchronizing some chaotic map if and only if there exists an integer $T > 0$ such that for any time interval of length T, there exists a vertex which can access other vertices by directed paths in that time interval.

Key words. synchronization, dynamical networks, time-varying coupling, Hajnal diameter, projection joint spectral radius, Lyapunov exponents, spanning tree

AMS subject classifications. 37C60, 15A51, 94C15

DOI. 10.1137/060657935

1. Introduction. Synchronization of dynamical processes on networks is presently an active research topic. It represents a mathematical framework that on the one hand can elucidate—desired or undesired—synchronization phenomena in diverse applications. On the other hand, the synchronization paradigm is formulated in such a manner that powerful mathematical techniques from dynamical systems and graph theory can be utilized. A standard version is

$$x^i(t+1) = f^i(x^1(t), x^2(t), \ldots, x^m(t)), \quad i = 1, 2, \ldots, m, \tag{1.1}$$

where $t \in \mathbb{Z}^+ = \{0, 1, 2, \ldots\}$ denotes the discrete time, $x^i(t) \in \mathbb{R}$ denotes the state variable of unit (vertex) i, and for $i = 1, 2, \ldots, m$, $f^i : \mathbb{R}^m \to \mathbb{R}$ is a C^1 function. This dynamical systems formulation contains two aspects. One of them is the reaction dynamics at each node or vertex of the network. The other one is the coupling structure, that is, whether and how strongly the dynamics at one node is directly influenced by the states of the other nodes.

Equation (1.1) clearly is an abstraction and simplification of synchronization problems found in applications. On the basis of understanding the dynamics of (1.1), research should then move on to more realistic scenarios. Therefore, in the present work, we address the question of synchronization when the right-hand side of (1.1) is allowed to vary in time. Thus, not only the dynamics itself is a temporal process, but

*Received by the editors April 22, 2006; accepted for publication (in revised form) May 7, 2007; published electronically November 30, 2007.

http://www.siam.org/journals/sima/39-4/65793.html

†Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, 04103 Leipzig, Germany (wenlian@mis.mpg.de, wemlian.lu@gmail.com, atay@member.ams.org, jjost@mis.mpg.de).

‡Laboratory of Mathematics for Nonlinear Sciences, School of Mathematical Sciences, Fudan University, 200433, Shanghai, China.

§Institut des Hautes Études Scientifiques, 91440 Bures-sur-Yvette, France.

also the underlying structure changes in time, albeit in some applications that may occur on a slower time scale.

The essence of the hypotheses on $f = [f^1, \dots, f^m]$ needed for synchronization results (to be stated in precise terms shortly) is that synchronization is possible as an invariant state, that is, when the dynamics starts on the diagonal $[x, \dots, x]$, it will stay there, and that this diagonal possesses a stable attracting state. The question about synchronization then is whether this state is also attracting for dynamical states $[x^1, \dots, x^m]$ outside the diagonal, at least locally, that is, when the components x^i are not necessarily equal but close to each other. This can be translated into a question about transverse Lyapunov exponents, and one typically concludes that the existence of a synchronized attractor in the sense of Milnor. In our contribution, we can already strengthen this result by concluding (under appropriate assumptions) the existence of a synchronized attractor in the strong sense instead of only in the weaker sense of Milnor. (We shall call this local complete synchronization.) This comes about because we achieve a reformulation of the synchronization problem in terms of Hajnal diameters (a concept to be explained below).

Our work, however, goes beyond that. As already indicated, our main contribution is that we can study the local complete synchronization of general coupled networks with time-varying coupling functions, in which each unit is dynamically evolving according to

$$x^i(t+1) = f_t^i(x^1(t), x^2(t), \dots, x^m(t)), \quad i = 1, 2, \dots, m. \tag{1.2}$$

This formulation, in fact, covers both aspects described above, the reaction dynamics as well as the coupling structure. The main purpose of the present paper then is to identify general conditions under which we can prove synchronization of the dynamics (1.2). Thus, we can handle variations of the reaction dynamics as well as those of the underlying network topology. We shall mention below various applications where this is of interest.

Before that, however, we state our technical hypothesis on the right-hand side of (1.2): for each $t \in \mathbb{Z}^+$, $f_t^i : \mathbb{R}^m \to \mathbb{R}$ is a C^1 function with the following hypothesis.

(H_1) *There exists a* C^1 *function* $f(s) : \mathbb{R} \to \mathbb{R}$ *such that*

$$f_t^i(s, s, \dots, s) = f(s)$$

holds for all $s \in \mathbb{R}$, $t \in \mathbb{Z}^+$, *and* $i = 1, 2, \dots, m$. *Moreover, for any compact set* $K \subset \mathbb{R}^m$, f_t^i *and the Jacobian matrices* $[\partial f_t^i / \partial x^j]_{i,j=1}^m$ *are all equicontinuous in* K *with respect to* $t \in \mathbb{Z}^+$ *and the latter are all nonsingular in* K.

This hypothesis ensures that the diagonal synchronization manifold

$$\mathcal{S} = \left\{ [x^1, x^2, \dots, x^m]^\top \in \mathbb{R}^m : \; x^i = x^j, \; i, j = 1, 2, \dots, m \right\}$$

is an invariant manifold for the evolution (1.2). If $x^1(t) = x^2(t) = \cdots = x^m(t) = s(t)$ denotes the synchronized state, then

$$s(t+1) = f(s(t)). \tag{1.3}$$

For the synchronized state (1.3), we assume the existence of an attractor, as follows.

(H_2) *There exists a compact asymptotically stable attractor* A *for* (1.3). *That is*, (i) $A \subset \mathbb{R}$ *is a forward invariant set*; (ii) *for any neighborhood* U *of* A *there* *exists a neighborhood* V *of* A *such that* $f^n(V) \subset U$ *for all* $n \in \mathbb{Z}^+$; (iii) *for any sufficiently small neighborhood* U *of* A, $f^n(U)$ *converges to* A, *in the sense that for any neighborhood* V, *there exists* n_0 *such that* $f^n(U) \subset V$ *for* $n \geq n_0$; (iv) *there exists* $s^* \in A$ *for which the* ω*-limit set is* A.

Let A^m denote the Cartesian product $A \times \cdots \times A$ (m times). Local complete synchronization (synchronization for simplicity) is defined in the sense that the set $\mathcal{S} \cap A^m = \{[x, \dots, x] : x \in A\}$ is an asymptotically stable attractor in $\mathbb{R}^m$. That is, for the coupled dynamical system (1.2), differences between components converge to zero if the initial states are picked sufficiently near $\mathcal{S} \cap A^m$, i.e., if the components are all close to the attractor A and if their differences are sufficiently small. In order to show such a synchronization, one needs a third hypothesis (H_3) that in technical terms is about Lyapunov exponents transverse to the diagonal. That is, while the dynamics on the attractor may well be expanding (the attractor might be chaotic), the transverse directions need to be suitably contracting to ensure synchronization. The corresponding hypothesis (H_3) will be stated below (see (3.2)) because it requires the introduction of crucial technical concepts.

It is an important aspect of our work that we shall derive the attractivity here in the classical sense, and not in the sense of Milnor, i.e., not only some set of positive measure, but a full neighborhood is attracted. For details about the difference between Milnor attractors and asymptotically stable attractors, see [1, 2]. Usually, when studying synchronization, one derives only the existence of a Milnor attractor; see [3].

The motivation for studying (1.2) comes from the well-known coupled map lattices (CML) [4], which can be written as follows:

$$x^i(t+1) = f(x^i(t)) + \sum_{j=1}^{m} L_{ij} f(x^j(t)),\ i = 1, 2, \dots, m, \tag{1.4}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a differentiable map and $L = [L_{ij}]_{i,j=1}^m \in \mathbb{R}^{m \times m}$ is the diffusion matrix, which is determined by the topological structure of the network and satisfies $L_{ij} \geq 0$ for all $i \neq j$, and $\sum_{j=1}^m L_{ij} = 0$ for all $i = 1, 2, \dots, m$. Letting $x = [x^1, x^2, \dots, x^m]^\top \in \mathbb{R}^m$, $F(x) = [f(x^1), f(x^2), \dots, f(x^m)]^\top \in \mathbb{R}^m$, and $G = I_m + L$, where I_m denotes the identity matrix of dimension m, the CML (1.4) can be written in the matrix form

$$x(t+1) = GF(x(t)), \tag{1.5}$$

where $G = [G_{ij}]_{i,j=1}^m \in \mathbb{R}^{m \times m}$ denotes the coupling and satisfies $G_{ij} \geq 0$ for $i \neq j$ and $\sum_{j=1}^m G_{ij} = 1$ for all $i = 1, 2, \dots, m$. So, if $G_{ii} \geq 0$ holds for all $i = 1, 2, \dots, m$, then G is a stochastic matrix.

Recently, synchronization of CML has attracted increasing attention [3, 5, 6, 7, 8]. Linear stability analysis of the synchronization manifold was proposed and transverse Lyapunov exponents were used to analyze the influence of the topological structure of networks. In [1], conditions for generalized transverse stability were presented. If the transverse (normal) Lyapunov exponents are negative, a chaotic attractor on an invariant submanifold can be asymptotically stable over the manifold. In [9, 10] it was shown that chaos synchronization in a network of nonlinear continuous-time or discrete-time dynamical systems, respectively, is possible if and only if the corresponding graph has a spanning tree. However, synchronization analysis has so far been limited to autonomous systems, where the interactions between the vertices (state components) are static and do not vary through time.

In the social, natural, and engineering real world, the topology of the network often varies through time. In communication networks, for example, one must consider dynamical networks of moving agents. Since the agents are moving, some of the existing connections can fail simply due to occurrence of an obstacle between agents. Also, some new connections may be created when one agent enters the effective region of other agents [11]. On top of that, randomness may also occur. In a communication network, the information channel of two agents at each time may be random [12]. When an error occurs at some time, the connections in the system will vary. In [11, 12, 13], synchronization of multiagent networks was considered where the state of each vertex is adapted according to the states of its connected neighbors with switching connecting topologies. This multiagent dynamical network can be written in discrete-time form as

$$x^i(t+1) = \sum_{j=1}^{m} G_{ij}(t)x^j(t),\ i = 1, 2, \dots, m, \tag{1.6}$$

where $x^j(t) \in \mathbb{R}$ is the state variable of vertex j and $[G_{ij}(t)]_{i,j=1}^m$, $t \in \mathbb{Z}^+$, are stochastic matrices. In [14] a convexity-conserving coupling function was considered that is equivalent to the linear coupling function in (1.6). It was found that the connectivity of the switching graphs plays a key role in the synchronization of multiagent networks with switching topologies. Also, in the recent literature [15, 16, 17], synchronization of continuous-time dynamical networks with time-varying topologies was studied. The time-varying couplings investigated, however, are specific, with either symmetry [15], node balance [16], or fixed time average [17].

Therefore, it is natural to investigate the synchronization of CML with general time-varying connections as

$$x(t+1) = G(t)F(x(t)), \tag{1.7}$$

where $G(t) = [G_{ij}(t)]_{i,j=1}^m \in \mathbb{R}^{m\times m}$ denotes the coupling matrix at time t and $F(x) = [f(x_1), \dots, f(x_n)]^\top$ is a differentiable function. We shall address this problem in the context of the general coupled system (1.2).

Let

$$x(t) = \begin{bmatrix} x^1(t) \\ x^2(t) \\ \vdots \\ x^m(t) \end{bmatrix} \quad \text{and} \quad F_t(x(t)) = \begin{bmatrix} f_t^1(x^1(t), \cdots, x^m(t)) \\ f_t^2(x^1(t), \cdots, x^m(t)) \\ \vdots \\ f_t^m(x^1(t), \cdots, x^m(t)) \end{bmatrix}.$$

Equation (1.2) can be rewritten in matrix form:

$$x(t+1) = F_t(x(t)). \tag{1.8}$$

The time-varying coupling can have a special form and may be driven by some other dynamical system. Let $\mathcal{Y} = \{\Omega, \mathcal{F}, P, \theta^{(t)}\}$ denote a metric dynamical system (MDS), where Ω is the metric state space, $\mathcal{F}$ is the σ-algebra, P is the probability measure, and $\theta^{(t)}$ is a semiflow satisfying $\theta^{(t+s)} = \theta^{(t)} \circ \theta^{(s)}$ and $\theta^{(0)} = \mathrm{id}$, where id denotes the identity map. Then the coupled system can be regarded as a random dynamical system (RDS) driven by $\mathcal{Y}$:

$$x(t+1) = F(x(t), \theta^{(t)}\omega),\ t \in \mathbb{Z}^+,\ \omega \in \Omega. \tag{1.9}$$

In fact, one can regard the dynamical system (1.9) as a skew product semiflow,

$$\begin{aligned}&\Theta : \mathbb{Z}^+ \times \Omega \times \mathbb{R}^m \to \Omega \times \mathbb{R}^m,\\ &\Theta^{(t)}(\omega, x) = (\theta^{(t)}\omega, x(t)).\end{aligned}$$

Furthermore, the coupled system can have the form

$$x(t+1) = F(x(t), u(t)),\ t \in \mathbb{Z}^+, \tag{1.10}$$

where u belongs to some function class $\mathcal{U}$ and may be interpreted as an external input or force. Then, defining $[\theta^{(t)}u](\tau) = u(t+\tau)$ as a shift map, the system (1.10) has the form of (1.9). In this paper, we first investigate the general time-varying case of the system (1.8) and also apply our results to systems of the form (1.9).

To study synchronization of the system (1.8), we use its variational equation by linearizing it. Consider the difference $\delta x^i(t) = x^i(t) - f^{(t-t_0)}(s_0)$. This implies that $\delta x^i(t) - \delta x^j(t) = x^i(t) - x^j(t)$ holds for all $i, j = 1, 2, \ldots, m$. We have

$$\delta x^i(t+t_0) = \sum_{j=1}^{m} \frac{\partial f^i_{t+t_0-1}}{\partial x^j}(f^{(t-1)}(s_0))\delta x^j(t+t_0-1),\ i = 1, 2, \ldots, m, \tag{1.11}$$

where for simplicity we use the notation $\frac{\partial f^i_{t+t_0-1}}{\partial x^j}(f^{(t-1)}(s_0))$ to denote $\frac{\partial f^i_{t+t_0-1}}{\partial x^j}$ $(f^{(t-1)}$ $(s_0), \ldots, f^{(t-1)}(s_0))$. Let

$$\delta x(t) = \begin{bmatrix} \delta x^1(t) \\ \vdots \\ \delta x^m(t) \end{bmatrix}, \quad D_t(s) = \left[\frac{\partial f^i_t}{\partial x^j}(s)\right]_{i,j=1}^{m}.$$

The variational equation (1.11) is written in matrix form,

$$\delta x(t+t_0) = D_{t+t_0-1}(f^{(t-1)}(s_0))\delta x(t+t_0-1). \tag{1.12}$$

For the Jacobian matrix, the following lemma is an immediate consequence of hypothesis (H_1).

LEMMA 1.1.

$$\sum_{j=1}^{m} \frac{\partial f^i_t}{\partial x_j}(s, s, \ldots, s) = f'(s), \quad i = 1, 2, \ldots, m \quad \textit{and} \quad t \in \mathbb{Z}^+.$$

Namely, all rows of the Jacobian matrix $[\partial f^i_t/\partial x_j]^m_{i,j=1}$ evaluated on the synchronization manifold $\mathcal{S}$ have the same sum, which is equal to $f'(s)$.

As a special case, if the time variation is driven by some dynamical system $\mathcal{Y} = \{\Omega, \mathcal{F}, P, \theta^{(t)}\}$, then the variational system does not depend on the initial time t_0, but only on (s_0, ω). Thus, the Jacobian matrix can be written in the form $D(f^{(t)}(s_0), \theta^{(t)}\omega) = D_t(f^{(t)}(s))$, by which the variational system can be written as

$$\delta x(t+1) = D(f^{(t)}(s_0), \theta^{(t)}\omega)\delta x(t). \tag{1.13}$$

In this paper, we first extend the concept of the Hajnal diameter to general matrices. A matrix with Hajnal diameter less than one has the property of compressing the convex hull of $\{x^1, \ldots, x^m\}$. Consequently, for an infinite sequence of time-varying

Jacobian matrices, the average compression rate can be used to verify synchronization. Since the Jacobian matrices have identical row sums, the (skew) projection along the diagonal synchronization direction can be used to define the projection joint spectral radius, which equals the Hajnal diameter. Furthermore, we show that the Hajnal diameter is equal to the largest Lyapunov exponent along directions transverse to the synchronization manifold; hence, it can also be used to determine whether the coupled system (1.2) can be synchronized.

Second, we apply these results to discuss the synchronization of the CML with time-varying couplings. As we shall show, the Hajnal diameter of infinite coupling stochastic matrices can be utilized to measure the synchronizability of the coupling process. More precisely, the coupled system (1.7) synchronizes if the sum of the logarithm of the Hajnal diameter and the largest Lyapunov exponent of the uncoupled system is negative. Using the equivalence of the Hajnal diameter, projection joint spectral radius, and transverse Lyapunov exponents, we study some particular examples for which the Hajnal diameter can be computed, including static coupling, a finite coupling set, and a multiplicative ergodic stochastic matrix process. We also present numerical examples to illustrate our theoretical results.

The connection structure of the CML (1.5) naturally gives rise to a graph, where each unit can be regarded as a vertex. Hence, we associate the coupling matrix G with a graph $\Gamma = (V, E)$, with the vertex set $V = \{1, 2, \ldots, m\}$ and the edge set $E = \{e_{ij}\}$, where there exists a directed edge from vertex j to vertex i if and only if $G_{ij} > 0$. The graphs we consider here are assumed to be simple (that is, without loops and multiple edges) but are allowed to be directed and weighted. That is, we do not assume a symmetric coupling scheme.

We extend this idea to an infinite graph sequence $\{\Gamma(t)\}$. That is, we regard a time-varying graph as a graph process $\{\Gamma(t)\}_{t\in\mathbb{Z}^+}$. Define $\Gamma(t) = [V, E(t)]$, where $V = \{1, 2, \ldots, m\}$ denotes the vertex set and $E(t) = \{e_{ij}(t)\}$ denotes the edge set of the graph at time t. The time-varying coupling matrix $G(t)$ might then be regarded as a function of the time-varying graph sequence, i.e., $G(t) = G(\Gamma(t))$. A basic problem that arises is determining which kind of sequence can ensure the synchrony of the coupled system for some chaotic synchronized state $s(t + 1) = f(s(t))$. As we shall show, the property that the union of the $\Gamma(t)$ contains a spanning tree is important for synchronizing chaotic maps. We prove that under certain conditions, the coupling graph process can synchronize some chaotic maps if and only if there exists an integer $T > 0$ such that there exists at least one vertex j from which any other vertex can be accessible within a time interval of length T.

This paper is organized as follows. In section 2, we present some definitions and lemmas on the Hajnal diameter, projection joint spectral radius, projection Lyapunov exponents, and transverse Lyapunov exponents for generalized Jacobian matrix sequences as well as stochastic matrix sequences. In section 3, we study the synchronization of the generalized coupled discrete-time systems with time-varying couplings (1.2). In section 4, we discuss the synchronization of the CML with time-varying couplings (1.7) and study the relation between synchronizability and coupling graph process topologies. In addition, we present some examples where synchronizability is analytically computable. In section 5, we present numerical examples to illustrate the theoretical results, and we conclude the paper in section 6.

2. Preliminaries. In this section we present some definitions and lemmas on matrix sequences. First, we extend the definitions of the Hajnal diameter and the projection joint spectral radius, introduced in [18, 19, 20] for stochastic matrices, to generalized time-varying matrix sequence. Furthermore, we extend Lyapunov exponents and projection Lyapunov exponents to the general time-varying case and discuss their relation. Second, we specialize these definitions to stochastic matrix sequences and introduce the relation between a stochastic matrix sequence and graph topology.

2.1. General definitions. We study the generalized time-varying linear system

$$u(t+t_0+1) = L_{t+t_0}(\varrho^{(t)}(\phi))u(t+t_0), \tag{2.1}$$

where $\varrho^{(t)}$ is defined by a random dynamical system $\{\Phi, \mathcal{B}, P, \varrho^{(t)}\}$, where Φ denotes the state space, $\mathcal{B}$ the σ-algebra on Φ, P the probability measure, and $\varrho^{(t)}$ a semiflow. Studying the linear system (2.1) comes from the variational system of the coupled system (1.2). For the variational system (1.12), $\varrho^{(t)}(\cdot)$ represents the synchronized state flow $f^{(t)}(\cdot)$. And, if $L_t(\cdot)$ is independent of t, then the linear system (2.1) can be rewritten as

$$u(t+1) = L(\varrho^{(t)}(\phi))u(t). \tag{2.2}$$

Thus, it can represent the variational system (1.13) as a special case, where $\varrho^{(t)}$ is the product flow $(f^{(t)}(\cdot), \theta^{(t)}(\cdot))$. Hence, the linear system (2.1) can unify the two cases of variational systems (1.12), (1.13) of the coupled system (1.2), (1.9).

For this purpose, we define a generalized matrix sequence map $\mathcal{L}$ from $\mathbb{Z}^+ \times \Phi$ to $2^{\mathbb{R}^{m\times m}}$,

$$\begin{aligned} \mathcal{L} : \mathbb{Z}^+ \times \Phi &\to 2^{\mathbb{R}^{m\times m}} \\ (t_0, \phi) &\mapsto \{L_{t+t_0}(\varrho^{(t)}\phi)\}_{t\in\mathbb{Z}^+}, \end{aligned} \tag{2.3}$$

where $2^{\mathbb{R}^{m\times m}}$ denotes the set containing all subsets of $\mathbb{R}^{m\times m}$. In [18, 19], the concept of the Hajnal diameter was introduced to describe the compression rate of a stochastic matrix. We extend it to general matrices below.

Definition 2.1. *For a matrix L with row vectors $g_1, \dots, g_m$ and a vector norm $\|\cdot\|$ in $\mathbb{R}^m$, the Hajnal diameter of L is defined by*

$$\mathrm{diam}(L, \|\cdot\|) = \max_{i,j} \|g_i - g_j\|.$$

We also introduce the Hajnal diameter for a matrix sequence map $\mathcal{L}$.

Definition 2.2. *For a generalized matrix sequence map $\mathcal{L}$, the Hajnal diameter of $\mathcal{L}$ at $\phi \in \Phi$ is defined by*

$$\mathrm{diam}(\mathcal{L}, \phi) = \overline{\lim_{t\to\infty}} \sup_{t_0 \geq 0} \left\{ \mathrm{diam} \left(\prod_{k=t_0}^{t_0+t-1} L_k(\varrho^{(k-t_0)}\phi) \right) \right\}^{\frac{1}{t}},$$

where $\prod$ denotes the left matrix product: $\prod_{k=1}^n A_k = A_n \times A_{n-1} \times \cdots \times A_1$.

The Hajnal diameter for the infinite matrix sequence map $\mathcal{L}$ does not depend on the choice of the norm. In fact, all norms in a Euclidean space are equivalent and any additional factor is eliminated by the power $1/t$ and the limit as $t \to \infty$.

Let $\mathcal{H} \subset \mathbb{R}^{m\times m}$ be a class of matrices having the property that all row sums are the same. Thus, all matrices in $\mathcal{H}$ share the common eigenvector $e_0 = [1, 1, \dots, 1]^\top$, where the corresponding eigenvalue is the row sum of the matrix. Then the projection joint spectral radius can be defined for a generalized matrix sequence map $\mathcal{L}$, similar to that introduced in [20] as follows.

DEFINITION 2.3. *Suppose $\mathcal{L}(t_0,\phi) \subset \mathcal{H}$ for $t_0 \in \mathbb{Z}^+$ and $\phi \in \Phi$. Let $\mathcal{E}_0$ be the subspace spanned by the synchronization direction $e_0 = [1,1,\dots,1]^\top$, and let P be any $(m-1)\times m$ matrix with exact kernel $\mathcal{E}_0$. We denote by $\hat{L} \in \mathbb{R}^{(m-1)\times(m-1)}$ the (skew) projection of matrix $L \in \mathcal{H}$ as the unique solution of*

$$PL = \hat{L}P. \tag{2.4}$$

The projection joint spectral radius of the generalized matrix sequence map $\mathcal{L}$ is defined as

$$\hat{\rho}(\mathcal{L},\phi) = \overline{\lim_{t\to\infty}} \sup_{t_0\geq 0} \left\| \prod_{k=t_0}^{t_0+t-1} \hat{L}_k(\varrho^{(k-t_0)}\phi) \right\|^{\frac{1}{t}}.$$

One can see that $\hat{\rho}(\mathcal{L},\phi)$ is independent of the choice of the matrix norm $\|\cdot\|$ induced by the vector norm. The following lemma shows that it is also independent of the choice of the matrix P.

LEMMA 2.4. *Suppose $\mathcal{L}(t_0,\phi) \subset \mathcal{H}$ for all $t_0 \geq 0$ and $\phi \in \Phi$. Then*

$$\hat{\rho}(\mathcal{L},\phi) = \operatorname{diam}(\mathcal{L},\phi).$$

A proof is given in the appendix.

The Lyapunov exponents are often used to study evolution of the dynamics [5, 6]. Here, we extend the definitions of Lyapunov exponents to general time-varying cases.

DEFINITION 2.5. *For the coupled system* (1.2)*, the Lyapunov exponent of the matrix sequence map $\mathcal{L}$ initiated by $\phi \in \Phi$ in the direction $u \in \mathbb{R}^m$ is defined as*

$$\lambda(\mathcal{L},\phi,u) = \overline{\lim_{t\to\infty}} \frac{1}{t} \sup_{t_0\geq 0} \log \left\| \prod_{k=t_0}^{t+t_0-1} L_k(\varrho^{(k-t_0)}\phi)u \right\|. \tag{2.5}$$

The projection along the synchronization direction e_0 can also define a Lyapunov exponent, called the projection Lyapunov exponent:

$$\hat{\lambda}(\mathcal{L},\phi,v) = \overline{\lim_{t\to\infty}} \frac{1}{t} \sup_{t_0\geq 0} \log \left\| \prod_{k=t_0}^{t+t_0-1} \hat{L}_k(\varrho^{(k-t_0)}\phi)v \right\|, \tag{2.6}$$

where $\hat{L}_k(\varrho^k\phi)$ is the projection of matrix $L_k(\varrho^k\phi)$ as defined in Definition 2.3.

It can be seen that the definition of the generalized Lyapunov exponent above satisfies the basic properties of Lyapunov exponents.[1] For more details about generalized Lyapunov exponents, we refer to [22].

LEMMA 2.6. *Suppose $\mathcal{L}(t_0,\phi) \subset \mathcal{H}$ for all $\phi \in \Phi$ and $t_0 \geq 0$. Then*

$$\sup_{v\in\mathbb{R}^{m-1},v\neq 0} \hat{\lambda}(\mathcal{L},\phi,v) = \log\hat{\rho}(\mathcal{L},\phi) = \log\operatorname{diam}(\mathcal{L},\phi).$$

A proof is given in the appendix.

[1]This kind of definition of characteristic exponent is similar to the Bohl exponent used to study uniform stability of time-varying systems in [21].

This lemma implies that the projection joint spectral radius gives the largest Lyapunov exponent in directions transverse to the synchronization direction e_0 of the matrix sequence map $\mathcal{L}$.

When the time dependence arises from being totally driven by some random dynamical system, we can write the generalized matrix sequence map $\mathcal{L}$ as $\mathcal{L}(\phi) = \{L(\varrho^{(t)}\phi)\}_{t\in\mathbb{Z}^+}$ since it is independent of t_0 and is just a map on Φ. As introduced in [23], we have specific definitions for Lyapunov exponents of the time-varying system (2.2) as follows.

For the linear system (2.2), the Lyapunov exponent of the matrix sequence map $\mathcal{L}$ initiated by $\phi \in \Phi$ in the direction $u \in \mathbb{R}^m$ is defined as

$$\lambda(\mathcal{L}, \phi, u) = \overline{\lim_{t\to\infty}} \frac{1}{t} \log \left\| \prod_{k=0}^{t-1} L(\varrho^{(k)}\phi) u \right\|. \tag{2.7}$$

If $\mathcal{L}(\phi) \subset \mathcal{H}$ for all $\phi \in \Phi$, then the Lyapunov exponent in the synchronization direction e_0 is

$$\lambda(\mathcal{L}, \phi, e_0) = \overline{\lim_{t\to\infty}} \frac{1}{t} \log \sum_{k=0}^{t-1} |c(k)|, \tag{2.8}$$

where $c(k)$ denotes the corresponding common row sum at each time k. The projection along the synchronization direction e_0 can also define a Lyapunov exponent, called the projection Lyapunov exponent:

$$\hat{\lambda}(\mathcal{L}, \phi, v) = \overline{\lim_{t\to\infty}} \frac{1}{t} \log \left\| \prod_{k=0}^{t-1} \hat{L}_k(\varrho^{(k)}\phi) v \right\|, \tag{2.9}$$

where $\hat{L}(\varrho^k\omega)$ is the (skew) projection of matrix $L(\varrho^k\omega)$. Also, the Hajnal diameter and projection joint spectral radius become

$$\operatorname{diam}(\mathcal{L}, \phi) = \overline{\lim_{t\to\infty}} \left\{ \operatorname{diam} \left(\prod_{k=0}^{t-1} L(\varrho^{(t)}\phi) \right) \right\}^{\frac{1}{t}}, \quad \hat{\rho}(\mathcal{L}, \phi) = \overline{\lim_{t\to\infty}} \left\| \prod_{k=0}^{t-1} \hat{L}(\varrho^{(k)}\phi) \right\|^{\frac{1}{t}}.$$

According to Lemmas 2.4 and 2.6, $\log \operatorname{diam}(\mathcal{L}, \phi) = \log \hat{\rho}(\mathcal{L}, \phi) = \sup_{v\in\mathbb{R}^{m-1}, v\neq 0} \hat{\lambda}$ $(\mathcal{L}, \phi, v)$. Let λ_0 be the Lyapunov exponent along the synchronization direction e_0 and let $\lambda_1, \lambda_2, \ldots, \lambda_{m-1}$ be the remaining Lyapunov exponents for the initial condition ϕ, counted with multiplicities.

LEMMA 2.7. *Suppose that $\mathcal{L}(\phi) \subset \mathcal{H}$ is time-independent. Let the matrix $D(t) = [D_{ij}(t)]_{i,j=1}^m$ denote the matrix $L(\varrho^{(t)}\phi)$ and let $c(t)$ denote the corresponding common row sum of $D(t)$. If*

1. $\lim_{t\to\infty} 1/t \sum_{k=0}^{t-1} \log |c(k)| = \lambda_0$,
2. $\overline{\lim}_{t\to\infty} 1/t \log^+ |D_{ij}(t)| \leq 0$ *for all* $i, j = 1, 2, \ldots, m$, *where* $\log^+(z) = \max \{\log z, 0\}$,

then

$$\log \operatorname{diam}(\mathcal{L}, \phi) = \log \hat{\rho}(\mathcal{L}, \phi) = \sup_{i\geq 1} \lambda_i.$$

A proof is given in the appendix.

Using the concept of the Hajnal diameter, we can define (uniform) synchronization of the nonautonomous system (1.2) as follows.

DEFINITION 2.8. *The coupled system* (1.2) *is said to be (uniformly locally completely) synchronized if there exists* $\eta > 0$ *such that for any* $\epsilon > 0$ *there exists* $T > 0$ *such that the inequality*

$$\operatorname{diam}\big([x^1(t), x^2(t), \dots, x^m(t)]^\top\big) \leq \epsilon \tag{2.10}$$

holds for all $t > t_0 + T$, $t_0 \geq 0$, *and* $x^i(t_0)$, $i = 1, 2, \dots, m$, *in the* η *neighborhood of* $s(t_0)$ *of a synchronized state* $s(t)$.

2.2. Stochastic matrix sequences. The above definitions can also be used to deal with stochastic matrix sequences.

DEFINITION 2.9. *A matrix* $G \in \mathbb{R}^{m\times m}$ *is said to be a stochastic matrix if its elements are nonnegative and each row sum is* 1.

We consider here the general time-varying case without the assumption of an underlying random dynamical system and write a stochastic matrix sequence as $\mathcal{G} = \{G(t)\}_{t\in\mathbb{Z}^+}$. The case that the time variation is driven by some dynamical system can be regarded as a special one.

DEFINITION 2.10. *The Hajnal diameter of* $\mathcal{G}$ *is defined as*

$$\operatorname{diam}(\mathcal{G}) = \overline{\lim_{t\to\infty}} \sup_{t_0\geq 0} \left(\operatorname{diam} \prod_{k=t_0}^{t_0+t-1} G(k)\right)^{\frac{1}{t}} \tag{2.11}$$

and the projection joint spectral radius for $\mathcal{G}$ *is*

$$\hat{\rho}(\mathcal{G}) = \overline{\lim_{t\to\infty}} \sup_{t_0\geq 0} \left\| \prod_{k=t_0}^{t_0+t-1} \hat{G}(k) \right\|^{\frac{1}{t}}, \tag{2.12}$$

where $\hat{G}(t)$ *is the projection of* $G(t)$, *as in Definition* 2.3.

Then, from Lemma 2.4, we have the following.

LEMMA 2.11. $\operatorname{diam}(\mathcal{G}) = \hat{\rho}(\mathcal{G})$.

To estimate the Hajnal diameter of a product of stochastic matrices, we use the concept of scrambling introduced in [20].

DEFINITION 2.12. *A stochastic matrix* $G = [G_{ij}]_{i,j=1}^m \in \mathbb{R}^{m\times m}$ *is said to be scrambling if for any* i, j *there exists an index* k *such that* $G_{ik} \neq 0$ *and* $G_{jk} \neq 0$.

For $g_i = [g_{i,1}, \dots, g_{i,m}] \in \mathbb{R}^m$ and $g_j = [g_{j,1}, \dots, g_{j,m}] \in \mathbb{R}^m$, define

$$g_i \wedge g_j = \big[\min(g_{i,1}, g_{j,1}), \dots, \min(g_{i,m}, g_{j,m})\big].$$

We use the following quantity introduced in [18, 19] to measure scramblingness:

$$\eta(G) = \min_{i,j} \|g_i \wedge g_j\|_1,$$

where $\|\cdot\|_1$ is the norm given by $\|x\|_1 = \sum_{i=1}^m |x_i|$ for $x = [x_1, \dots, x_m] \in \mathbb{R}^m$. It is clear that $0 \leq \eta(G) \leq 1$ and that $\eta(G) > 0$ if and only if G is scrambling. Thus, the well-known Hajnal inequality has the following generalized form.

LEMMA 2.13 (generalized Hajnal inequality [20, Theorem 6]). *For any vector norm in* $\mathbb{R}^m$ *and any two stochastic matrices* G *and* H,

$$\operatorname{diam}(GH) \leq (1 - \eta(G))\operatorname{diam}(H). \tag{2.13}$$

The concepts of projection joint spectral radius and the Hajnal diameter are linked to the ergodicity of stochastic matrix sequences. We can extend the ergodicity for a matrix set [20, 24] to a matrix sequence as follows.

DEFINITION 2.14 (ergodicity [14, Definition 1]). *A stochastic matrix sequence $\Sigma = \{G(t)\}_{t\in\mathbb{Z}^+}$ is said to be ergodic if for any t_0 and $\epsilon > 0$ there exists $T > 0$ such that for any $t > T$ and some norm $\|\cdot\|$,*

$$\operatorname{diam}\left(\prod_{s=t_0}^{t_0+t-1} G(s)\right) \leq \epsilon. \tag{2.14}$$

Moreover, if for any $\epsilon > 0$ there exists $T > 0$ such that inequality (2.14) *holds for all $t \geq T$ and $t_0 \geq 0$, $\mathcal{G}$ is said to be uniformly ergodic.*

A stochastic matrix $G = [G_{ij}]_{i,j=1}^m$ can be associated with a graph $\Gamma = [V, E]$, where $V = \{1, 2, \ldots, m\}$ denotes the vertex set and $E = \{e_{ij}\}$ the edge set, in the sense that there exists an edge from vertex j to i if and only if $G_{ij} > 0$. Let $\Gamma_1 = [V, E_1]$ and $\Gamma = [V, E_2]$ be two simple graphs with the same vertex set. We also define the union $\Gamma_1 \bigcup \Gamma_2 = [V, E_1 \bigcup E_2]$ (merging multiple edges). It can be seen that for two stochastic matrices G_1 and G_2 with the same dimension and positive diagonal elements, the edge set of $\Gamma_1 \bigcup \Gamma_2$ is contained in that of the corresponding graph of the product matrix $G_1 G_2$. In this way, we can define the union of the graph sequence $\{\Gamma(t)\}_{t\in\mathbb{Z}^+}$ across the time interval $[t_1, t_2]$ by $\bigcup_{k=t_1}^{t_2} \Gamma(k) = [V, \bigcup_{k=t_1}^{t_2} E(k)]$. The following concepts for graphs can be found, e.g., in [25].

DEFINITION 2.15. *A graph Γ is said to have a spanning tree if there exists a vertex, called the root, such that for each other vertex j there exists at least one directed path from the root to vertex j.*

It follows that $\{\Gamma(t)\}_{t\in\mathbb{Z}^+}$ has a spanning tree across the time interval $[t_1, t_2]$ if the union of $\{\Gamma(t)\}_{t\in\mathbb{Z}^+}$ across $[t_1, t_2]$ has a spanning tree. This is equivalent to the existence of a vertex from which all other vertices can be accessible across $[t_1, t_2]$.

DEFINITION 2.16. *A graph Γ is said to be scrambling if for any different vertices i and j there exists a vertex k such that there exist edges from k to i and from k to j.*

It follows that a stochastic matrix G is scrambling if and only if the corresponding graph Γ is scrambling.

LEMMA 2.17 (see [24, Lemma 4]). *Let $G(1), G(2), \ldots, G(m-1)$ be stochastic matrices with positive diagonal elements, where each of the corresponding graphs $\Gamma(1)$, $\Gamma(2), \ldots, \Gamma(m-1)$ have spanning trees. Then $\prod_{k=1}^{m-1} G(k)$ is scrambling.*

Suppose now that the stochastic matrix sequence $\mathcal{G}$ is driven by some metric dynamical system $\mathcal{Y} = \{\Omega, \mathcal{F}, P, \theta^{(t)}\}$. We write $\mathcal{G}$ as $\{G(t) = G(\theta^{(t)}\omega)\}_{t\in\mathbb{Z}^+}$, where $\omega \in \Omega$. Then, as stated in section 2.1, we can define the Lyapunov exponents.

DEFINITION 2.18. *The Lyapunov exponent of the stochastic matrix sequence $\mathcal{G}$ is defined as*

$$\sigma(\mathcal{G}, \omega, u) = \overline{\lim_{t\to\infty}} \frac{1}{t} \log \left\| \prod_{k=0}^{t-1} G(\theta^t \omega) u \right\|.$$

The projection Lyapunov exponents is defined as

$$\hat{\sigma}(\mathcal{G}, \omega, u) = \overline{\lim_{t\to\infty}} \frac{1}{t} \log \left\| \prod_{k=0}^{t-1} \hat{G}(\theta^t \omega) u \right\|,$$

where $\hat{G}(\cdot)$ is the projection of $G(\cdot)$ as defined in Definition 2.3.

For a given $\omega \in \Omega$, one can see that $\mathrm{diam}(\mathcal{G})$ and $\hat{\rho}(\mathcal{G})$ both equal the largest Lyapunov exponent of $\mathcal{G}$ in directions transverse to the synchronization direction under several mild conditions.

In closing this section, we list some notation to be used in the remainder of the paper. The matrix $\hat{L}$ denotes the (skew) projection of the matrix L along the vector e introduced in Definition 2.3, and $\hat{\mathcal{L}}$ is the (skew) projection of the matrix sequence map $\mathcal{L}$ along e. For $x = (x^1, \dots, x^m)^\top \in \mathbb{R}^m$, the average $\frac{1}{m}\sum_{i=1}^m x^i$ of x is denoted by $\bar{x}$. The notation $\|\cdot\|$ denotes some vector norm in the linear space $\mathbb{R}^m$, and also the matrix norm in $\mathbb{R}^{m\times m}$ induced by this vector norm. $f^{(t)}(s_0)$ denotes the t-iteration of the map f with initial condition s_0. We let $x(t, t_0, x_0)$ be the solution of the coupled system (1.2) with initial condition $x(t_0) = x_0$, which we sometimes abbreviate as $x(t)$.

3. Generalized synchronization analysis. For the variational system (1.12), similar to subsection 2.1, we denote by $\mathcal{D}$ the Jacobian sequence map in the generalized sense; i.e., $\mathcal{D}$ is a map from $\mathbb{Z}^+ \times \mathbb{R}$ to $2^{\mathbb{R}^{m\times m}}$: $\mathcal{D}(t_0, s_0) = \{D_{t+t_0}(f^{(t)}(s_0))\}_{t\in\mathbb{Z}^+} \subset \mathcal{H}$ for all $t_0 \in \mathbb{Z}^+$ and $s_0 \in A$. Furthermore, letting

$$B(t, t_0) = \prod_{k=t_0}^{t+t_0-1} D_k(f^{(k-t_0)}(s_0)),$$

we can rewrite the variational system (1.12) as follows:

$$\delta x(t + t_0) = D_{t+t_0-1}(f^{(t-1)}(s_0))\delta x(t + t_0 - 1) = B(t, t_0)\delta x(t_0). \tag{3.1}$$

From Definitions 2.2 and 2.3, we have

$$\mathrm{diam}(\mathcal{D}, s_0) = \overline{\lim_{t\to\infty}} \sup_{t_0\ge 0} \left\{ \mathrm{diam}\left(\prod_{k=t_0}^{t_0+t-1} D_k(f^{(k-t_0)}(s_0)) \right) \right\}^{\frac{1}{t}},$$
$$\hat{\rho}(\mathcal{D}, s_0) = \overline{\lim_{t\to\infty}} \sup_{t_0\ge 0} \left\| \prod_{k=t_0}^{t_0+t-1} \hat{D}_k(f^{(k-t_0)}(s_0)) \right\|^{\frac{1}{t}}.$$

We will also refer to the following hypothesis.

(H_3)

$$\sup_{s_0\in A} \mathrm{diam}(\mathcal{D}, s_0) < 1. \tag{3.2}$$

THEOREM 3.1. *If hypotheses* (H_1)–(H_3) *hold, then the compact set* $A^m \bigcap \mathcal{S}$ *is a uniformly asymptotically stable attractor of the coupled system* (1.2) *in* $\mathbb{R}^m$; *i.e., the coupled system* (1.2) *is uniformly locally completely synchronized.*

Proof. Let

$$\mathrm{diam}(\mathcal{D}, t_0, t, s_0) = \mathrm{diam}\left(\prod_{k=t_0}^{t_0+t-1} D_k\big(f^{(k-t_0)}(s_0)\big) \right),$$
$$\mathrm{diam}(\mathcal{D}, t, s_0) = \sup_{t_0\ge 0} \left\{ \mathrm{diam}\left(\prod_{k=t_0}^{t_0+t-1} D_k(f^{(k-t_0)}(s_0)) \right) \right\}.$$

According to (H_3), letting $1 > d > \sup_{s_0\in A} \mathrm{diam}(\mathcal{D}, s_0)$ and n_0 satisfy $d^{n_0} < \frac{1}{3}$, for any $s_0 \in A$, there exists $n(s_0) \ge n_0$ such that $\mathrm{diam}(\mathcal{D}, t, s_0) < d$ holds for all $t \geq n(s_0)$. By equicontinuity (H_1) and compactness (H_2), there must exist a finite integer set $\mathcal{V} = \{n_1, n_2, \ldots, n_v\}$ satisfying $n_i \geq n_0$ for all $i = 1, 2, \ldots, v$ and a neighborhood U of A such that for any $s_0 \in U$ there exists $n_j \in \mathcal{V}$ such that $\mathrm{diam}\big(\prod_{k=t_0}^{t_0+n_j-1} D_k(f^{(k-t_0)}(s_0))\big) < d^{n_j} < \frac{1}{3}$ holds for all $t_0 \geq 0$.

By hypothesis (H_2), there exists a compact neighborhood W of A such that $U \supset W \supset A$, $f(W) \subset W$, and $\bigcap_{n\geq 0} f^{(n)}(W) = A$ [26]. Let

$$a = \min_{n\in\mathcal{V}} d_H(f^{(n)}(W), W) > 0,$$

where $d_H(\cdot,\cdot)$ denotes the Hausdorff metric in $\mathbb{R}$. Then define a compact set

$$W_\alpha = \left\{ x = (x^1, \ldots, x^m) \in \mathbb{R}^m : \max_{1\leq i\leq m} |x^i - \bar{x}| \leq \alpha \text{ and } \bar{x} \in W \right\}.$$

By the mean value theorem, we have

$$f_k^i(x^1(k), \ldots, x^m(k)) - f(s(k), \ldots, s(k)) = \sum_{j=1}^{m} \frac{\partial f_k^i}{\partial x^j}(\xi_k^{ij}),$$

where ξ_k^{ij} belongs to the closed interval induced by the two ends $x^i(k)$ and $s(k)$. Denote by $D_k(\xi_k)$ the matrix $[\partial f_k^i(\xi_k^{ij})/\partial x^j]_{i,j=1}^m$.

Let $\alpha > 0$ be sufficiently small so that for each $x_0 \in W_\alpha$ with $s(t_0) = \bar{x}_0$ and $x(t_0) = x_0$, there exists $t_1 \in \mathcal{V}$ such that

$$|x^i(t_1, t_0, x_0) - f^{(t_1-t_0)}(\bar{x}_0)| \leq \frac{a}{2},$$
$$\mathrm{diam}\left(\prod_{k=t_0}^{t_0+t_1-1} D_k(\xi_k) \right) < \frac{1}{2}$$

hold for all $t_0 \geq 0$. Then, for any $x_0 \in W_\alpha$, $\bar{x}_0 \in W$, we have

$$\delta x(t_1 + t_0) = \prod_{k=t_0}^{t_1+t_0-1} D_k(\xi(k))\delta x_0 = \tilde{B}(t_1, t_0)\delta x_0,$$

where $\tilde{B}(t_1, t_0) = \prod_{k=t_0}^{t_1+t_0-1} D_k(\xi(k))$. Then

$$\begin{aligned} |\delta x^i(t_1 + t_0) - \delta x^j(t_1 + t_0)| &\leq \sum_{k=1}^{m} |\tilde{B}_{ik}(t_1, t_0) - \tilde{B}_{jk}(t_1, t_0)||\delta x_0^j| \\ &\leq \mathrm{diam}(\tilde{B}(t_1, t_0)) \max_{1\leq i\leq m} |x_0^i - \bar{x}_0|. \end{aligned}$$

Thus, we conclude that

$$\max_{1\leq i,j\leq m} |x^i(t_1 + t_0) - x^j(t_1 + t_0)| \leq \frac{1}{2} \max_{1\leq i,j\leq m} |x_0^i - x_0^j|.$$

By the definition of W_α, we see that $x(t_1 + t_0) \in W_{\alpha/2}$. With initial time $t_0 + t_1$, we can continue this phase and afterwards obtain

$$\lim_{t\to\infty} |x^i(t) - x^j(t)| = 0, \ i, j = 1, 2, \ldots, m,$$

uniformly with respect to $t_0 \in \mathbb{Z}^+$ and $x_0 \in W_\alpha$. Therefore, the coupled system (1.2) is uniformly synchronized. Furthermore, we obtain that $A^m \bigcap \mathcal{S}$ is a uniformly asymptotically stable attractor for the coupled system (1.2), and the convergence rate can be estimated by $O(\{\sup_{s_0\in A}\operatorname{diam}(\mathcal{D}, s_0)\}^t)$ since d is chosen arbitrarily greater than $\sup_{s_0\in A}\operatorname{diam}(\mathcal{D}, s_0)$. The theorem is proved. □

Remark 1. The idea of the above proof comes from that of Theorem 2.12 in [1], with a modification for the time-varying case. In Theorem 2.12 in [1], the authors used normal Lyapunov exponents to prove asymptotic stability of the original autonomous system for the case when it is asymptotically stable in an invariant manifold. In this paper, we directly use the Hajnal diameter of the left product of the infinite Jacobian matrix sequence map to measure the transverse differences of the collections of spatial states. Furthermore, we consider a nonautonomous system here due to time-varying couplings.

Following Lemma 2.4 gives us the following.

COROLLARY 3.2. *If* $\sup_{s_0\in A}\hat{\rho}(\mathcal{D}, s_0) < 1$, *then the coupled system* (1.2) *is uniformly synchronized.*

Consider the special case that the coupled system (1.9) is an RDS on an MDS $\mathcal{Y} = \{\Omega, \mathcal{F}, P, \theta^{(t)}\}$. We can write this coupled system (1.9) as a product dynamical system $\{A\times\Omega, \mathbf{F}, \mathbf{P}, \Theta^{(t)}\}$, where $\mathbf{F}$ is the product σ-algebra on $A\times\Omega$, $\mathbf{P}$ denotes the probability measure, and $\Theta^{(t)}(s_0, \omega) = (\theta^{(t)}\omega, f^{(t)}(s_0))$. Let $D(f^{(t)}(s_0), \theta^{(t)}\omega)$ denote the Jacobian matrix at time t. By Definition 2.5, the Lyapunov exponents for the coupled system (1.9) can be written as follows:

$$\lambda(u, s_0, \omega) = \overline{\lim_{t\to\infty}}\frac{1}{t}\log\left\|\prod_{k=0}^{t-1} D(f^{(k)}(s_0), \theta^{(k)}\omega)u\right\|.$$

It can be seen that the Lyapunov exponent along the diagonal synchronization direction e_0 is

$$\lambda(e_0, s_0, \omega) = \overline{\lim_{t\to\infty}}\frac{1}{t}\sum_{k=0}^{t-1}\log|c(k)|,$$

where $c(k)$ is the common row sum of $D(f^{(k)}(s_0), \theta^{(k)}\omega)$. Let $\lambda_0 = \lambda(e_0, s_0, \omega)$, $\lambda_1, \dots, \lambda_{m-1}$ be the Lyapunov exponents (counting multiplicity) of the dynamical system $\mathcal{L}$ with the initial condition (s_0, ω). From Lemma 2.7, we conclude that $\sup_{i\geq 1}\lambda_i = \log\hat{\rho}(F, s_0, \omega) = \log\operatorname{diam}(F, s_0, \omega)$. If the probability $\mathbf{P}$ is ergodic, then the Lyapunov exponents exist for almost all $s_0 \in A$ and $\omega \in \Omega$, and furthermore they are independent of (s_0, ω).

COROLLARY 3.3. *Suppose that hypotheses* (H_1)–(H_2) *and the assumptions in Lemma* 2.7 *hold. Suppose further that* $A\times\Omega$ *is compact in the weak topology defined in this RDS*, *the semiflow* $\Theta^{(t)}$ *is continuous*, *the Jacobian matrix* $D(\cdot,\cdot)$ *is nonsingular and continuous on* $A\times\Omega$, *and*

$$\sup_{\mathbf{P}\in\mathrm{Erg}_\Theta(\mathrm{A}\times\Omega)}\sup_{i\geq 1}\lambda_i < 0,$$

where $\mathrm{Erg}_\Theta(\mathrm{A}\times\Omega)$ *denotes the ergodic probability measure set supported in* $\{A\times\Omega, \mathbf{F}, \Theta^{(t)}\}$. *Then the coupled system* (1.9) *is uniformly locally completely synchronized.*

Proof. By Theorem 2.8 in [1], we have

$$\sup_{\mathbf{P}\in Erg_\Theta(A\times\Omega)}\lambda_{\max}(\hat{\mathcal{D}}, \mathbf{P}) = \sup_{\|u\|=1,(s_0,\omega)\in A\times\Omega}\overline{\lim_{t\to\infty}}\frac{1}{t}\log\left\|\prod_{k=0}^{t-1}\hat{D}(f^{(k)}(s_0), \theta^{(k)}\omega)u\right\|,$$

where $\hat{\mathcal{D}}$ is the projection of the intrinsic matrix sequence map $\mathcal{D}$ and $\lambda_{\max}(\hat{\mathcal{D}}, \mathbf{P})$ denotes the largest Lyapunov exponent of $\hat{\mathcal{D}}$ according to the ergodic probability $\mathbf{P}$ (the value for almost all (s_0, ω) according to $\mathbf{P}$). From Lemmas 2.4, 2.6, and 2.7, it follows that

$$\begin{aligned}\sup_{\mathbf{P}\in\mathrm{Erg}_\Theta(\mathrm{A}\times\Omega)} \sup_{i\geq 1} \lambda_i &= \sup_{\mathbf{P}\in\mathrm{Erg}_\Theta(\mathrm{A}\times\Omega)} \lambda_{\max}(\hat{\mathcal{D}}, \mathbf{P}) = \sup_{(s_0,\omega)\in A\times\Omega} \lambda_{\max}(\hat{\mathcal{D}}, s_0, \omega)\\ &= \sup_{(s_0,\omega)\in A\times\Omega} \log\hat{\rho}(\mathcal{D}, s_0, \omega) = \sup_{(s_0,\omega)\in A\times\Omega} \log \mathrm{diam}(\mathcal{D}, s_0, \omega).\end{aligned}$$

The corollary is proved as a direct consequence from Theorem 3.1. □

Remark 2. If λ_0 is the largest Lyapunov exponent, then $V = \{u : \ \lambda(u) < \lambda_0\}$ constructs a subspace of $\mathbb{R}^m$ which is transverse to the synchronization direction e_0. Corollary 3.3 implies that if all Lyapunov exponents in the transverse directions are negative, then the coupled system (1.2) is synchronized. Otherwise, if λ_0 is not the largest Lyapunov exponent, then $\sup_{i\geq 1}\lambda_i < 0$ implies that the largest exponent is negative, which means that the synchronized solution $s(t)$ is itself asymptotically stable through the evolution (1.9).

Remark 3. From Lemma 2.7, it can also be seen that when computing $\rho(\mathcal{D})$, it is sufficient to compute the largest Lyapunov exponent of $\hat{\mathcal{D}}$. In [1], the authors proved for an autonomous dynamical system that if all Lyapunov exponents of the normal directions, namely, the Lyapunov exponents for $\hat{\mathcal{D}}$, are negative, then the attractor in the invariant submanifold is an attractor in $\mathbb{R}^m$ (or a more general manifold). In this paper, we extend the proof of Theorem 2.12 in [1] to the general time-varying coupled system (1.2) by discussing the relation between the Hajnal diameter and transverse Lyapunov exponents. In the following sections, we continue the synchronization analysis for nonautonomous dynamical systems.

4. Synchronization analysis of CML with time-varying topologies. Consider the following coupled system with time-varying topologies:

$$x^i(t+1) = \sum_{j=1}^{m} G_{ij}(t) f(x^j(t)), \quad i = 1, 2, \ldots, m, \ t \in \mathbb{Z}^+, \tag{4.1}$$

where $f(\cdot) : \mathbb{R} \to \mathbb{R}$ is C^1 continuous and $G(t) = [G_{ij}(t)]_{i,j=1}^m$ is a stochastic matrix. In matrix form,

$$x(t+1) = G(t)F(x(t)). \tag{4.2}$$

Since the coupling matrix $G(t)$ is a stochastic matrix, the diagonal synchronization manifold is invariant and we have the uncoupled (or synchronized) state as

$$s(t+1) = f(s(t)). \tag{4.3}$$

We suppose that for the synchronized state (4.3), there exists an asymptotically stable attractor A with the (maximum) Lyapunov exponent

$$\mu = \sup_{s_0\in A} \overline{\lim_{t\to\infty}} \frac{1}{t} \sum_{k=0}^{t-1} \log|f'(s(k))|.$$

System (4.1) is a special form of (1.2) satisfying the equicontinuous condition (H_1). Linearizing system (4.1) about the synchronized state yields the variational equation

$$\delta x^i(t+1) = \sum_{j=1}^{m} G_{ij}(t) f'(s(t)) \delta x^i(t), \ i = 1, 2, \ldots, m,$$

and

$$\mathrm{diam}\left(\prod_{k=t_0}^{t_0+t-1} G(k) f'(f^{(k-t_0)}(s_0))\right) = \mathrm{diam}\left(\prod_{k=t_0}^{t_0+t-1} G(k)\right) \left|\prod_{l=0}^{t} f'(f^{(l)}(s_0))\right|.$$

Denote the stochastic matrix sequence $\{G(t)\}_{t\in\mathbb{Z}^+}$ by $\mathcal{G}$. Thus, the Hajnal diameter of the variational system is $\mathrm{diam}(\mathcal{G})e^{\mu}$. Using Theorem 3.1, we have the following result.

THEOREM 4.1. *Suppose that the uncoupled system $s(t+1) = f(s(t))$ satisfies hypothesis* (H_2) *with Lyapunov exponent μ. Let $\mathcal{G} = \{G(t)\}_{t\in\mathbb{Z}^+}$. If*

$$\mathrm{diam}(\mathcal{G})e^{\mu} < 1, \tag{4.4}$$

then the coupled system (4.1) *is synchronized.*

From Theorem 4.1, one can see that the quantity $\mathrm{diam}(\mathcal{G})$ as well as other equivalent quantities such as the projection joint spectral radius and the Lyapunov exponent, can be used to measure the synchronizability of the time-varying coupling, i.e., the coupling stochastic matrix sequence $\mathcal{G}$. A smaller value of $\mathrm{diam}(\mathcal{G})$ implies a better synchronizability of the time-varying coupling topology. If the uncoupled system (4.3) is chaotic, i.e., $\mu > 0$, then the necessary condition for synchronization condition (4.4) is $\mathrm{diam}(\mathcal{G}) < 1$. So, it is important to investigate under what conditions $\mathrm{diam}(\mathcal{G}) < 1$ holds.

Suppose that the stochastic matrix set $\mathcal{M}$ satisfies the following hypotheses.

(H_4) *$\mathcal{M}$ is compact and there exists $r > 0$ such that for any $G = [G_{ij}]_{i,j=1}^m \in \mathcal{M}$, $G_{ij} > 0$ implies $G_{ij} \geq r$ and all diagonal elements $G_{ii} > r$, $i = 1, 2, \ldots, m$.*

We denote the graph sequence corresponding to the stochastic matrix sequence $\mathcal{G}$ by $\mathbf{\Gamma} = \{\Gamma(t)\}_{t\in\mathbb{Z}^+}$. Then we have the following result.

THEOREM 4.2. *Suppose that the stochastic matrix sequence $\mathcal{G} \subset \mathcal{M}$ satisfies hypothesis* (H_4). *Then the following statements are equivalent:*

1. $\mathrm{diam}(\mathcal{G}) < 1$;
2. *there exists $T > 0$ such that for any t_0 the graph $\bigcup_{k=t_0}^{t_0+T} \Gamma(k)$ has a spanning tree*;
3. *the stochastic matrix sequence $\mathcal{G}$ is uniformly ergodic.*

Proof. We first show that (3) $\Rightarrow$ (2) by reduction to absurdity. Let $B(t_0, t) = \prod_{k=t_0}^{t_0+t-1} G(k)$. Since $\mathcal{G}$ is uniformly ergodic, there must exist $T > 0$ such that $\mathrm{diam}(B(t_0, T)) < 1/2$ holds for any $t_0 \geq 0$. So, $v = \prod_{k=t_0}^{t_0+T-1} G(k)u$ satisfies

$$\max_{1\leq i,j\leq m} |v_i - v_j| \leq \mathrm{diam}(B(t_0, T))\|u\|_\infty \leq \frac{1}{2}\|u\|_\infty. \tag{4.5}$$

If the second condition does not hold, then there exists t_T such that the union $\bigcup_{k=t_T}^{t_T+T-1} \Gamma(k)$ does not have a spanning tree. That is, there exist two vertices v_1 and v_2 such that for any vertex z there is either no directed path from z to v_1 or no directed path from z to v_2. Let U_1 (U_2) be the vertex set which can reach v_1 (v_2, respectively) across $[t_T, t_T + T - 1]$. This implies that U_1 and U_2 are disjoint across $[t_T, t_T + T - 1]$ and no edge starts outside of U_1 (U_2) and ends in U_1 (U_2). Furthermore, considering the Frobenius form of $G(t)$, one can see that the elements in the corresponding rows of U_1 (U_2) with columns associated with the complementary set of U_1 (U_2) are all zeros. Let

$$u_i = \begin{cases} 1, & i \in U_1, \\ 0, & i \in U_2, \\ \text{any value in } (0,1), & \text{otherwise.} \end{cases}$$

We have

$$v_i = \begin{cases} 1, & i \in U_1, \\ 0, & i \in U_2, \\ \in [0,1], & \text{otherwise}. \end{cases}$$

This implies that $\max_{1\le i,j\le m} |v_i - v_j| \ge 1 = \|u\|_\infty$, which contradicts (4.5). Therefore, (3) $\Rightarrow$ (2) can be concluded.

We next show (2) $\Rightarrow$ (1). Applying Lemma 2.17, there exists $T > 0$ such that $\prod_{k=t_0}^{t_0+T-1} G(k)$ is scrambling for any t_0. There exists $\delta > 0$ such that $\eta(B(T,t_0)) > \delta > 0$ for all $t_0 \ge 0$ because of the compactness of the set $\mathcal{M}$. So,

$$\begin{aligned} \operatorname{diam}(B(t,t_0)) &= \operatorname{diam}\left\{ B\left(\operatorname{mod}(t,T), t_0 + \left[\frac{t}{T}\right] T\right) \prod_{k=1}^{[\frac{t}{T}]} B(T, t_0 + (k-1)T) \right\} \\ &\le \operatorname{diam}\left\{ \prod_{k=1}^{[\frac{t}{T}]} B(t_0 + kT - 1, t_0 + (k-1)T) \right\} \\ &\le 2(1-\delta)^{[\frac{t}{T}]} \end{aligned} \tag{4.6}$$

holds for any $t_0 \ge 0$. Here, $[t/T]$ denotes the largest integer less than t/T and $\operatorname{mod}(t,T)$ denotes the modulus of the division $t \div T$. Thus,

$$\operatorname{diam}(\mathcal{G}) \le (1-\delta)^{\frac{1}{T}} < 1.$$

This proves (2) $\Rightarrow$ (1). Since (1) $\Rightarrow$ (3) is clear, the theorem is proved. ∎

Remark 4. According to Lemma 2.17, it can be seen that the union of graphs across any time interval of length T has a spanning tree if and only if a union of graphs across any time interval of length $(m-1)T$ is scrambling.

Moreover, from [27], we conclude more results on the ergodicity of stochastic matrix sequences as follows.

PROPOSITION 4.3. *The implication* (1) $\Rightarrow$ (2) $\Rightarrow$ (3) *holds for the following statements:*

1. $\operatorname{diam}(\mathcal{G}) < 1$*;*
2. $\mathcal{G}$ *is ergodic;*
3. *for any* $t_0 \ge 0$*, the union* $\bigcup_{k\ge t_0} \Gamma(k)$ *has a spanning tree.*

Remark 5. It should be pointed out that the implications in Proposition 4.3 cannot be reversed. Counterexamples can be found in [14]. However, in [14], it is also proved under certain conditions that if the stochastic matrices have the property that $G_{ij} > 0$ if and only if $G_{ji} > 0$, then statement 2 is equivalent to statement 3.

Assembling Theorem 4.2, Proposition 4.3, and the results in [14], it can be shown that, for $\mathcal{G} \subset \mathcal{M}$, the implications

$$A_1 \Leftrightarrow A_2 \Leftrightarrow A_3 \Rightarrow A_4 \Rightarrow A_5$$

hold regarding the following statements:

- $\mathcal{A}_1$: $\operatorname{diam}(\mathcal{G}) < 1$.
- $\mathcal{A}_2$: there exists $T > 0$ such that the union across any T-length time interval $[t_0.t_0 + T]$: $\bigcup_{k=t_0}^{t_0+T} \Gamma(k)$ has a spanning tree.
- $\mathcal{A}_3$: $\mathcal{G}$ is uniformly ergodic.
- $\mathcal{A}_4$: $\mathcal{G}$ is ergodic.

- $\mathcal{A}_5$: for any t_0, the union across $[t_0, \infty)$: $\bigcup_{k \geq t_0} \Gamma(k)$ has a spanning tree.

In the following, we present some special classes of examples of CML with time-varying couplings. These classes were widely used to describe discrete-time networks and were studied in some recent papers [5, 6, 20, 27]. The synchronization criterion for these classes can be verified by numerical methods. Thus, the synchronizability diam($\mathcal{G}$) of the time-varying couplings can also be computed numerically.

4.1. Static topology. If $G(t)$ is a static matrix, i.e., $G(t) = G$ for all $t \in \mathbb{Z}^+$, then we can write the coupled system (4.1) as

$$x(t+1) = GF(x(t)). \tag{4.7}$$

PROPOSITION 4.4. *Let* $1 = \sigma_0, \sigma_1, \sigma_2, \dots, \sigma_{m-1}$ *be the eigenvalues of* G *ordered by* $1 \geq |\sigma_1| \geq |\sigma_2| \geq \cdots \geq |\sigma_{m-1}|$. *If* $|\sigma_1| e^{\mu} < 1$, *then the coupled system* (4.7) *is synchronized.*

Proof. Let $v_0 = e_0$ and choose column vectors $v_1, v_2, \dots, v_{m-1}$ in $\mathbb{R}^m$ such that $v_0, v_1, \dots, v_{m-1}$ is an orthonormal basis for $\mathbb{R}^m$. Let $A = [v_0, v_1, \dots, v_{m-1}]$. Then

$$A^{-1}GA = \begin{bmatrix} 1 & \alpha \\ 0 & \hat{G} \end{bmatrix},$$

where the eigenvalues of $\hat{G}$ are $\sigma_1, \dots, \sigma_{m-1}$. By the Householder theorem (see Theorem 4.2.1 in [28]), for any $\epsilon > 0$, there must exist a norm in $\mathbb{R}^m$ such that with its induced matrix norm,

$$|\sigma_1| \leq \|\hat{G}\| \leq |\sigma_1| + \epsilon.$$

Since ϵ is arbitrary, for the static stochastic matrix sequence $\mathcal{G}_0 = \{G, G, \dots, \}$, it can be concluded that $\hat{\rho}(\mathcal{G}_0) = |\sigma_1|$. Using Theorem 4.1, the conclusion follows. Moreover, it can also be obtained that the convergence rate is $O((|\sigma_1| e^{\mu})^t)$. ∎

Remark 6. Similar results have been obtained by several papers concerning synchronization of CML with static connections (see [5, 6, 8, 29]). Here, we have proved this result in a different way as a consequence of our main result.

4.2. Finite topology set. Let $\mathcal{Q}$ be a compact stochastic matrix set satisfying (H_4). Consider the following inclusions:

$$x(t+1) \in \mathcal{Q}F(x(t)), \tag{4.8}$$

i.e.,

$$x(t+1) = G(t)F(x(t)), \tag{4.9}$$

$$G(t) \in \mathcal{Q}. \tag{4.10}$$

Then the synchronization of the coupled system (4.8) can be formulated as follows.

DEFINITION 4.5. *The coupled inclusion system* (4.8) *is said to be synchronized if, for any stochastic matrix sequence* $\mathcal{G} \subset \mathcal{Q}$, *the coupled system* (4.9) *is synchronized.*

In [20], the authors defined the Hajnal diameter and projection joint spectral radius for a compact stochastic matrix set.

DEFINITION 4.6. *For the stochastic matrix set* $\mathcal{Q}$, *the Hajnal diameter is given by*

$$\operatorname{diam}(\mathcal{Q}) = \overline{\lim_{t\to\infty}} \sup_{G(k)\in\mathcal{Q}} \left\{ \operatorname{diam}\left(\prod_{k=0}^{t-1} G(k) \right) \right\}^{\frac{1}{t}},$$

and the projection joint spectral radius is

$$\hat{\rho}(\mathcal{Q}) = \overline{\lim_{t\to\infty}} \left\{ \sup_{G(k)\in\mathcal{Q}} \left\| \prod_{k=0}^{t-1} \hat{G}(k) \right\| \right\}^{\frac{1}{t}}.$$

The following result is from [20].

LEMMA 4.7. *Suppose $\mathcal{Q}$ is a compact set of stochastic matrices. Then*

$$\mathrm{diam}(\mathcal{Q}) = \hat{\rho}(\mathcal{Q}).$$

Using Theorem 4.1, we have the following.

THEOREM 4.8. *If* $\mathrm{diam}(\mathcal{Q})e^{\mu} < 1$, *then the coupled system* (4.8) *is synchronized.*

Moreover, we conclude that the synchronization is uniform with respect to $t_0 \in \mathbb{Z}^+$ and stochastic matrix sequences $\mathcal{G} \subset \mathcal{Q}$. Furthermore, we have the following result on synchronizability of the stochastic matrix set $\mathcal{Q}$.

PROPOSITION 4.9. *Let $\mathcal{Q}$ be a compact set of stochastic matrices satisfying hypothesis* (H_4). *Then the following statements are equivalent:*

- $\mathcal{B}_1$: $\mathrm{diam}(\mathcal{Q}) < 1$.
- $\mathcal{B}_2$: *for any stochastic matrix sequence $\mathcal{G} \subset \mathcal{Q}$, $\mathcal{G}$ is ergodic.*
- $\mathcal{B}_3$: *each corresponding graph of a stochastic matrix $G \in \mathcal{Q}$ has a spanning tree.*

Proof. The implication $\mathcal{B}_1 \Rightarrow \mathcal{B}_2 \Rightarrow \mathcal{B}_3$ is clear by Proposition 4.3. And $\mathcal{B}_3 \Rightarrow \mathcal{B}_1$ can be obtained by the proof of Theorem 4.2 since $\mathcal{Q}$ is a finite set of stochastic matrices satisfying hypothesis (H_4). ∎

Remark 7. By the methods introduced in [30, 31, 32], $\hat{\rho}(\mathcal{Q})$ can be computed to arbitrary precision for a finite set $\mathcal{Q}$ despite a large computational complexity.

4.3. Multiplicative ergodic topology sequence. Consider the stochastic matrix sequence $\mathcal{G} = \{G(t)\}_{t\in\mathbb{Z}^+}$ driven by some dynamical system $\mathcal{Y} = \{\Omega, \mathcal{F}, P, \theta^{(t)}\}$, i.e., $\mathcal{G} = \{G(\theta^{(t)}\omega)\}$ for some continuous map $G(\cdot)$. Recall the Lyapunov exponent for $\mathcal{G}$:

$$\sigma(v,\omega) = \overline{\lim_{t\to\infty}} \frac{1}{t} \log \left\| \prod_{k=0}^{t-1} G(\theta^{(k)}\omega)v \right\|.$$

It is clear that $\sigma(e_0,\omega) = 0$ for all ω and $\sigma(v,\omega) \le 0$ for all ω and $v \in \mathbb{R}^m$. So, the linear subspace

$$L_\omega = \{v, \sigma(v,\omega) < 0\}$$

denotes the directions transverse to the synchronization manifold. If P is an ergodic measure for the MDS $\mathcal{Y}$, then $\sigma(u,\omega)$ and $L(\omega)$ are the same for almost all ω with respect to P [33]. Then we can let σ_1 be the largest Lyapunov exponent of $\mathcal{G}$ transverse to the synchronization direction e_0. By Theorem 4.1 and Corollary 3.3, we have the following.

THEOREM 4.10. *Suppose that $\theta^{(t)}$ is a continuous semiflow*, $G(\cdot)$ *is continuous on all $\omega \in \Omega$ and nonsingular*, *and Ω is compact. If*

$$\sup_{\mathrm{Erf}_\theta(\Omega)} \sigma_1 + \mu < 0,$$

then the coupled system (4.1) *is synchronized.*

Remark 8. There are many papers discussing the computation of multiplicative Lyapunov exponents; for example, see [34, 35]. In particular, [36] discussed the Lyapunov exponents for the product of infinite matrices. By Lemma 2.7, we can compute the largest projection Lyapunov exponent which equals σ_1. We will illustrate this in the following section.

5. Numerical illustrations. In this section, we will numerically illustrate the theoretical results on synchronization of CML with time-varying couplings. In these examples, the coupling matrices are driven by random dynamical systems which can be regarded as stochastic processes. Then the projection Lyapunov exponents are computed numerically by the time series of coupling matrices. In this way, we can verify the synchronization criterion and analyze synchronizability numerically. Consider the following coupled map network with time-varying topology:

$$x^i(t+1) = \frac{1}{\sum\limits_{k=1}^{m} A_{ik}(t)} \sum_{j=1}^{m} A_{ij}(t) f(x^j(t)),\ \ i = 1, 2, \dots, m, \tag{5.1}$$

where $x^i(t) \in \mathbb{R}$ and $f(s) = \alpha s(1-s)$ is the logistic map with $\alpha = 3.9$, which implies that the Lyapunov exponent of f is $\mu \approx 0.5$. The stochastic coupling matrix at time t is

$$G(t) = [G_{ij}(t)]_{i,j=1}^{m} = \left[\frac{A_{ij}(t)}{\sum\limits_{j=1}^{m} A_{ij}(t)}\right]_{i,j=1}^{m}.$$

5.1. Blinking scale-free networks. The blinking scale-free network is a model initiated by a scale-free network and evolves with malfunction and recovery. At time $t = 0$, the initial graph $\Gamma(0)$ is a scale-free network introduced in [37]. At each time $t \geq 1$, every vertex i malfunctions with probability $p \ll 1$. If vertex i malfunctions, all edges linked to it disappear. In addition, a malfunctioned vertex recovers after a time interval T and then causes the reestablishment of all edges linked to it in the initial graph $\Gamma(0)$. The coupling $A_{ij}(t) = A_{ji}(t) = 1$ if vertex j is connected to i at time t; otherwise, $A_{ij}(t) = A_{ji}(t) = 0$ and $A_{ii}(t) = 1$ for all $i, j = 1, 2, \dots, m$.

In Figure 1, we show the convergence of the second Lyapunov exponent σ_1 during the topology evolution with different malfunction probability p. We measure synchronization by the variance $K = 1/(m-1)\langle\sum_{i=1}^{m}(x^i(t) - \bar{x}(t))^2\rangle$, where $\langle\cdot\rangle$ denotes the time average, and we denote $W = \sigma_1 + \mu$. We pick the evolution time length to be 1000 and choose initial conditions randomly from the interval $(0, 1)$. In Figure 2, we show the variation of K and W with respect to the malfunction probability p. It can be seen that the region where W is negative coincides with the region of synchronization, i.e., where K is near zero.

5.2. Blurring directed graph process. A blurring directed graph process is one where each edge weight is a modified Wiener process. In details, the graph process is started with a directed weighted graph $\Gamma(0)$ of which for each vertex pair (i, j), one of two edges $A_{ij}(0)$ and $A_{ji}(0)$ is a random variable uniformly distributed between 1 and 2, and the other is zero with equal probability, for all $i \neq j$; $A_{ii}(0) = 0$ for all $i = 1, 2, \dots, m$. At each time $t \geq 1$, for each $A_{ij}(t-1) \neq 0$, $i \neq j$ we denote the

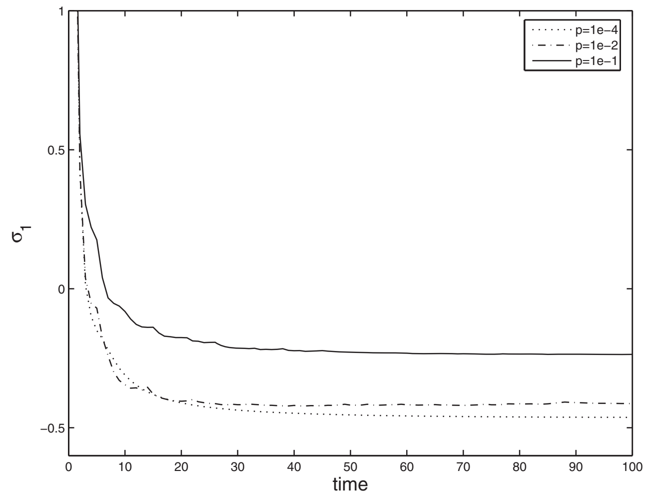


FIG. 1. *Convergence of the second Lyapunov exponent σ_1 for the blinking topology during the topology evolution with the same recovery time $T = 3$ and different malfunction probabilities $p = 10^{-1}$, $p = 10^{-2}$, and $p = 10^{-4}$. The initial scale-free graph is constructed by the method introduced in* [37] *with network size* 500 *and average degree* 12.

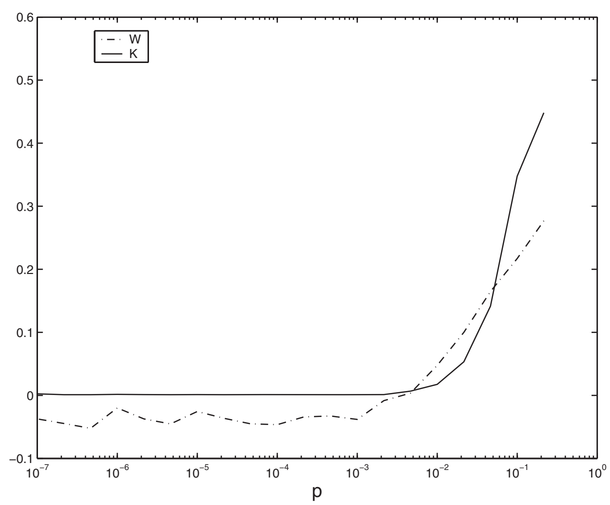


FIG. 2. *Variation of K and W with respect to p for the blinking topology.*

difference $A_{ij}(t) - A_{ij}(t-1)$ by a Gaussian distribution $\mathcal{N}(0, r^2)$ which is statistically independent for all $i \neq j$ and $t \in \mathbb{Z}^+$. If this results in $A_{ij}(t)$ being negative, a weight will be added to the reversal orientation, i.e., $A_{ji}(t) = |A_{ij}(t)|$ and $A_{ij}(t) = 0$. Moreover, if as a result of the process above there exists some index i such that $A_{ij} = 0$ holds for all $j = 1, 2, \ldots, m$, then pick $A_{ii}(t) = 1$.

In Figure 3, we show the convergence of the second Lyapunov exponent σ_1 during the topology evolution for different values of the Gaussian distribution variance r. Picking $r = 0.05$, we show the synchronization of the coupled system (5.1). Let $K(t) = 1/(m-1)\langle\sum_{i=1}^{m}(x^i(t) - \bar{x}(t))^2\rangle_t$, where $\langle\cdot\rangle_t$ denotes the time average from

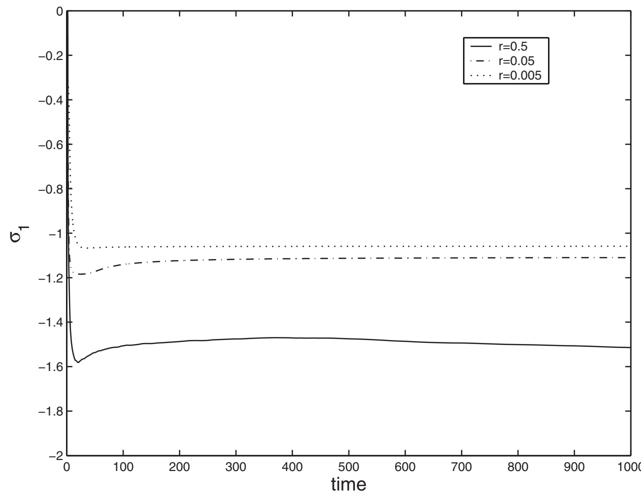


FIG. 3. *Convergence of the second Lyapunov exponent σ_1 for the blurring graph process during the topology evolution with Gaussian variance $r = 0.5$, 0.05, 0.005, and the size of the network $m = 100$.*

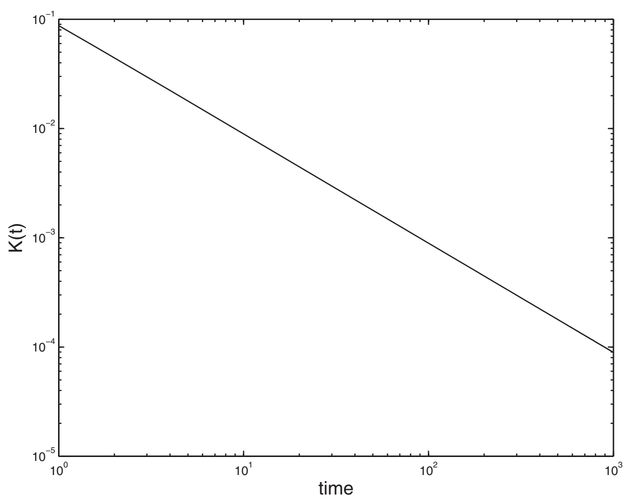


FIG. 4. *Variation of $K(t)$ with respect to time for the blurring graph process.*

0 to t. Since $W = \sigma_1 + \mu$ is about -0.6, i.e., less than zero, the coupled system is synchronized. Figure 4 shows in logarithmic scale the convergence of $K(t)$ to zero.

6. Conclusion. In this paper, we have presented a synchronization analysis for discrete-time dynamical networks with time-varying topologies. We have extended the concept of the Hajnal diameter to generalized matrix sequences to discuss the synchronization of the coupled system. Furthermore, this quantity is equivalent to other widely used quantities such as the projection joint spectral radius and transverse Lyapunov exponents, which we have also extended to the time-varying case. Thus, these results can be used to discuss the synchronization of the CML with time-varying couplings. The Hajnal diameter is utilized to describe synchronizability of the time-varying couplings and obtain a criterion guaranteeing synchronization. Time-varying couplings can be regarded as a stochastic matrix sequence associated with a sequence of graphs. Synchronizability is tightly related to the topology. As we have shown, the statement that $\mathrm{diam}(\mathcal{G}) < 1$, i.e., that chaotic synchronization is possible, is equivalent to saying that there exists an integer T such that the union of the graphs across any time interval of length T has a spanning tree. The methodology will be similarly extended to higher-dimensional maps elsewhere.

Appendix.

Proof of Lemma 2.4. The proof of this lemma comes from [20] with a minor modification. First, we show $\mathrm{diam}(\mathcal{L},\phi) \le \hat{\rho}(\mathcal{L},\phi)$. Let J be any complement of $\mathcal{E}_0$ in $\mathbb{R}^m$ with a basis $u_0,\dots,u_{m-1}$ such that $u_0 = e_0$. Let $A = [u_0, u_1, \dots, u_{m-1}]$, which is nonsingular. Then, for any $t > t_0$ and $t_0 \ge 0$,

$$A^{-1}L_t(\varrho^{(t-t_0)}\phi)A = \begin{bmatrix} c(t) & \alpha_t \\ 0 & \hat{L}_t(\varrho^{(t-t_0)}\phi) \end{bmatrix},$$

where $c(t)$ denotes the row sum of $L_t(\varrho^{(t-t_0)}\phi)$ which is also the eigenvalue corresponding eigenvector e, and $\hat{L}_t(\varrho^{(t-t_0)}\phi)$ can be the solution of linear equation (2.4) with P composed of the rows of A^{-1} except the first row. For any $d > \hat{\rho}(\mathcal{L},\phi)$, there exists $T > 0$ such that the inequality

$$\left\| \prod_{k=t_0}^{t_0+t-1} \hat{L}_k(\varrho^{(k-t_0)}\phi) \right\| \le d^t$$

holds for all $t \ge T$ and $t_0 \ge 0$. Let

$$A^{-1}\prod_{k=t_0}^{t_0+t-1} L_k(\varrho^{(k-t_0)}\phi)A = \begin{bmatrix} \prod_{k=t_0}^{t_0+t-1} c(k) & \alpha_t \\ 0 & \prod_{k=t_0}^{t_0+t-1} \hat{L}_k(\varrho^{(k-t_0)}\phi) \end{bmatrix}.$$

Then

$$\begin{aligned} &\left\| A^{-1}\prod_{k=t_0}^{t_0+t-1} L_k(\varrho^{(k-t_0)}\phi)A - \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \left(\prod_{k=t_0}^{t_0+t-1} c(k), \alpha_t \right) \right\| \\ &= \left\| \begin{bmatrix} 0 & 0 \\ 0 & \prod_{k=t_0}^{t_0+t-1} \hat{L}_k(\varrho^{(k-t_0)}\phi) \end{bmatrix} \right\| \le Cd^t \end{aligned}$$

holds for some constant $C > 0$. Therefore,

$$\left\| \prod_{s=t_0}^{t_0+t-1} L_k(\varrho^{(k-t_0)}\phi) - A \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \left(\prod_{k=t_0}^{t_0+t-1} c(k), \alpha_t \right) A^{-1} \right\| \le C_1 d^t,$$

$$\left\| \prod_{k=t_0}^{t_0+t-1} L_k(\varrho^{(k-t_0)}\phi) - e \cdot q \right\| \le C_1 d^t,$$

where $q = \big[\prod_{k=t_0}^{t_0+t-1} c(k), \alpha_t\big] A^{-1}$ and C_1 is a positive constant. It says that all row vectors of $\prod_{k=t_0}^{t_0+t-1} L_k(\varrho^{(k-t_0)}\phi)$ lie inside the $C_1 d^m$ neighborhood of q. Hence,

$$\mathrm{diam}\bigg(\prod_{k=t_0}^{t_0+t-1} L_k(\varrho^{(k-t_0)}\phi)\bigg) \le C_2 d^t$$

for some constant $C_2 > 0$, all $t \ge T$, and $t_0 \ge 0$. This implies that $\mathrm{diam}(\mathcal{L}, \phi) \le d$. Since d is arbitrary, $\mathrm{diam}(\mathcal{L}, \phi) \le \hat{\rho}(\mathcal{L}, \phi)$ can be concluded.

Second, we show that $\hat{\rho}(\mathcal{L}, \phi) \le \mathrm{diam}(\mathcal{L}, \phi)$. For any $d > \mathrm{diam}(\mathcal{L}, \phi)$, there exists $T > 0$ such that

$$\mathrm{diam}\bigg(\prod_{k=t_0}^{t_0+t-1} L_k(\varrho^{(k-t_0)}\phi)\bigg) \le d^t$$

holds for all $t \ge T$ and $t_0 \ge 0$. Letting q be the first row of $\prod_{k=t_0}^{t_0+t-1} L_k(\varrho^{(k-t_0)}\phi)$, we have

$$\bigg\|\prod_{k=t_0}^{t_0+t-1} L_k(\varrho^{(k-t_0)}\phi) - e \cdot q\bigg\| \le C_3 d^t$$

for some positive constant C_3. Let A be defined as above. Then

$$\bigg\|A^{-1}\prod_{k=t_0}^{t_0+t-1} L_k(\varrho^{(k-t_0)}\phi) A - A^{-1} e \cdot q A\bigg\| \le C_4 d^t,$$

i.e.,

$$\left\|\begin{bmatrix} \prod_{k=t_0}^{t_0+t-1} c(k) & \alpha_t \\ 0 & \prod_{k=t_0}^{t_0+t-1} \hat{L}_k(\varrho^{(k-t_0)}\phi) \end{bmatrix} - \begin{bmatrix} \gamma & \beta \\ 0 & 0 \end{bmatrix}\right\| \le C_4 d^t$$

holds for some γ and β. This implies that

$$\bigg\|\prod_{k=t_0}^{t_0+t-1} \hat{L}_k(\varrho^{(k-t_0)}\phi)\bigg\| \le C_5 d^t$$

holds for all $t \ge T$, $t_0 \ge 0$, and some $C_5 > 0$. Therefore, $\hat{\rho}(\mathcal{L}, \phi) \le d$. The proof is completed since d is chosen arbitrarily. ◻

Proof of Lemma 2.6. Let $\hat{\lambda}_{\max} = \sup_{v \in \mathbb{R}^{m-1}} \hat{\lambda}(\mathcal{L}, \phi, v)$. First, it is easy to see that $\log \hat{\rho}(\mathcal{L}, \phi) \ge \hat{\lambda}_{\max}$. We will show $\log \hat{\rho}(\mathcal{L}, \phi) = \hat{\lambda}_{\max}$. Otherwise, there exists $d \in (\exp(\hat{\lambda}_{\max}), \hat{\rho}(\mathcal{L}, \phi))$. By the properties of Lyapunov exponents, for any normalized orthogonal basis $u_1, u_2, \ldots, u_{m-1} \in \mathbb{R}^{m-1}$ with Lyapunov exponent $\hat{\lambda}(\mathcal{L}, \phi, u_i) = \hat{\lambda}_i$, we have, for any $u \in \mathbb{R}^{m-1}$, $\hat{\lambda}(\mathcal{L}, \phi, u) = \hat{\lambda}_{i_u}$, where $i_u \in \{1, 2, \ldots, m-1\}$. $\hat{\rho}(\mathcal{L}, \phi) > d$ implies that there exist $t_0 \ge 0$ and a sequence t_n with $\lim_{n\to\infty} t_n = +\infty$ such that

$$\bigg\|\prod_{k=t_0}^{t_n+t_0-1} \hat{L}_k(\varrho^{(k-t_0)}\phi)\bigg\| > d^{t_n}$$

for all $n \geq 0$. That is, there also exists a sequence $v_n \in \mathbb{R}^{m-1}$ with $\|v_n\| = 1$ such that

$$\left\| \prod_{k=t_0}^{t_n+t_0-1} \hat{L}_k(\varrho^{(k-t_0)}\phi)v_n \right\| > d^{t_n}.$$

There exists a subsequence of v_n (still denoted by v_n) with $\lim_{n\to\infty} v_n = v^*$. Let $\delta v_n = v_n - v^*$. We have

$$\left\| \prod_{k=t_0}^{t_n+t_0-1} \hat{L}_k(\varrho^{(k-t_0)}\phi)v^* \right\| \geq \left\| \prod_{k=t_0}^{t_n+t_0-1} \hat{L}_k(\varrho^{(k-t_0)}\phi)v_n \right\| - \left\| \prod_{k=t_0}^{t_n+t_0-1} \hat{L}_k(\varrho^{(k-t_0)}\phi)\delta v_n \right\|.$$

Note that we can write $\delta v_n = \sum_{i=1}^{m-1} \delta x_n^i u_i$, where $\delta x_n^i \in \mathbb{R}$ with $\lim_{n\to\infty} \delta x_n^i = 0$. So, there exists an integer N such that $\| \prod_{k=t_0}^{t_n+t_0-1} \hat{L}_k(\varrho^{(k-t_0)}\phi)\delta v_n\| \leq \big(\sum_{i=1}^{m-1} |\delta x_n^i|\big) d^{t_n}$ holds for all $n \geq N$. Then we have

$$\left\| \prod_{k=t_0}^{t_n+t_0-1} \hat{L}_k(\varrho^{(k-t_0)}\phi)v^* \right\| \geq d^{t_n} - d^{t_n}\left(\sum_{i=1}^{m-1} |\delta x_n^i| \right) \geq C d^{t_n}$$

for all $n \geq N$ and some $C > 0$. This implies $\max_{v\in\mathbb{R}^m} \hat{\lambda}(\mathcal{L}, \phi, v) \geq \log d$, which contradicts the assumption $d \in (\exp(\hat{\lambda}_{\max}), \hat{\rho}(\mathcal{L}, \phi))$. Hence, $\hat{\lambda}_{\max} = \log \hat{\rho}(\mathcal{L}, \phi)$. $\square$

Proof of Lemma 2.7. Recall that $\{\Phi, \mathcal{B}, P, \varrho^{(t)}\}$ denotes a random dynamical system, where Φ denotes the state space, $\mathcal{B}$ denotes the σ-algebra, P denotes the probability measure, and $\varrho^{(t)}$ denotes the semiflow. For a given $\phi \in \Phi$ we denote $L(\varrho^{(t)}\phi)$ by $L(t)$. Let $A = [u_1, u_2, \ldots, u_m] \in \mathbb{R}^{m\times m}$, where $u_1, \ldots, u_m$ denotes a basis of $\mathbb{R}^m$ and $u_1 = e$,

$$A^{-1} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix} \in \mathbb{R}^{m\times m}$$

is the inverse of A with

$$\bar{L}(t) = A^{-1}L(t)A = \begin{bmatrix} c(t) & \alpha^\top(t) \\ 0 & \hat{L}(t) \end{bmatrix},\ \hat{L}(t) = A_1^* D(t) A_1,\ \alpha^\top(t) = v_1 L(t) A_1,$$

where $A_1 = [u_2, \ldots, u_m] \in \mathbb{R}^{m\times(m-1)}$, and

$$A_1^* = \begin{bmatrix} v_2 \\ \vdots \\ v_m \end{bmatrix} \in \mathbb{R}^{(m-1)\times m}.$$

One can see that the set of Lyapunov exponents of the dynamical system $\{\bar{L}(t)\}_{t\in\mathbb{Z}^+}$ are the same as those of $\{L(t)\}_{t\in\mathbb{Z}^+}$. For any $z(0) = [x(0), y(0)] \in \mathbb{R}^m$, where $x(0) \in \mathbb{R}$ and $y(0) \in \mathbb{R}^{m-1}$, this evolution $z(t+1) = \bar{L}(t)z(t)$ leads to

$$z(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} c(t-1)x(t-1) + \alpha^\top(t-1)y(t-1) \\ \hat{L}(t-1)y(t-1) \end{bmatrix}.$$

So, we have

$$
\begin{aligned}
y(t) &= \prod_{k=0}^{t-1} \hat{L}(k) y(0), \\
x(t) &= \prod_{k=0}^{t-1} c(k) x(0) + \sum_{k=1}^{t} \prod_{p=t-k+1}^{t-1} c(p) \alpha^{\top}(t-k) \prod_{q=0}^{t-k-1} \hat{L}(q) y(0).
\end{aligned} \tag{A.1}
$$

If the upper bound is less than the lower bound for the left matrix product $\prod$, then the product should be the identity matrix. In the following, we denote by $\hat{\mathcal{L}}$ the projection sequence map of $\mathcal{L}$ and will prove this lemma for two cases.

Case 1. $\lambda_0 \leq \log \hat{\rho}(\mathcal{L}, \phi)$. Since $\hat{\rho}(\mathcal{L}, \phi)$ is just the largest Lyapunov exponent of $\hat{\mathcal{L}}$ defined by $\hat{\lambda}$, from conditions 1 and 2, one can see that for any $\epsilon > 0$, there exists $T > 0$ such that for any $t \geq T$ it holds that $|\alpha(t)| \leq e^{\epsilon t}$, $\left\| \prod_{k=0}^{t-1} \hat{L}(k) \right\| \leq e^{(\hat{\lambda}+\epsilon)t}$, and $e^{(\lambda_0-\epsilon)t} \leq |\prod_{k=0}^{t-1} c(k)| \leq e^{(\lambda_0+\epsilon)t}$. Thus, we can obtain

$$
\begin{aligned}
\prod_{k=t-k+1}^{t-1} |c(p)| &= \prod_{p=0}^{t-1} |c(p)| \times \frac{1}{\prod\limits_{p=0}^{t-k} |c(p)|} \\
&= \begin{cases} e^{(\lambda_0+\epsilon)(t)} e^{-(\lambda_0-\epsilon)(t-k+1)}, & k \leq t-T+1, \\ e^{(\lambda_0+\epsilon)(t-1)} \max\limits_{T \geq q \geq 0} \left(\prod\limits_{p=0}^{q} |c(p)| \right)^{-1}, & t-1 \geq k \geq t-T. \end{cases}
\end{aligned}
$$

Then we have

$$
\begin{aligned}
|x(t)| &\leq \prod_{k=0}^{t-1} |c(k)||x(0)| + \sum_{k=1}^{t-T+1} \prod_{p=t-k+1}^{t-1} |c(p)||\alpha^{\top}(t-k)| \prod_{q=0}^{t-k-1} \|\hat{L}(q)\| \|y(0)\| \\
&\quad + \sum_{k=t-T}^{t-1} \prod_{p=t-k+1}^{t-1} |c(p)| \|\alpha^{\top}(t-k)\| \prod_{q=0}^{t-k-1} \|\hat{L}(q)\| \|y(0)\| \\
&\leq e^{(\lambda_0+\epsilon)t} + \sum_{k=1}^{t-T+1} e^{(\lambda_0+\epsilon)(t-1)} e^{\epsilon t} e^{-(\lambda_0-\epsilon)(t-k)} e^{(\hat{\lambda}+\epsilon)(t-k)} + M_1 e^{(\lambda_0+\epsilon)(t-1)} \\
&\leq e^{(\hat{\lambda}+\epsilon)t} + e^{(\hat{\lambda}+4\epsilon)t} e^{-(\lambda_0+\epsilon)} \sum_{k=1}^{t-T+1} e^{(-\hat{\lambda}+\lambda_0-3\epsilon)k} + M_1 e^{(\lambda_0+\epsilon)t} \\
&\leq M_2 e^{(\hat{\lambda}+4\epsilon)t},
\end{aligned}
$$

where

$$
M_1 = (T+1) \max_{T \geq q \geq 0} \left(\prod_{p=0}^{q} |c(p)| \right)^{-1} e^{\epsilon T} \left(\prod_{p=0}^{q} \|\hat{L}(p)\| \right) \|y(0)\|,
$$

$$
M_2 = 1 + M_1 + e^{-(\lambda_0+\epsilon)} \sum_{k=1}^{\infty} e^{-3\epsilon k}.
$$

So,

$$\overline{\lim_{t\to\infty}}\frac{1}{t}\log\|z(t-1)\|\le\hat{\lambda}+4\epsilon$$

holds for all $z(0)\in\mathbb{R}^m$. Noting that $\hat{\lambda}$ must be less than the largest Lyapunov exponent of $\mathcal{L}$, we conclude that $\hat{\lambda}$ is the largest Lyapunov exponent. This implies the conclusion of the lemma.

Case 2. $\lambda_0>\hat{\lambda}$. Note that for any $\epsilon\in(0,(\lambda_0-\hat{\lambda})/3)$ there exists T such that

$$\prod_{k=0}^{t}|c^{-1}(k)|\|\alpha^\top(t)\|\prod_{l=0}^{t}\|\hat{L}(l)\|\le Ce^{(-\lambda_0+\hat{\lambda}+3\epsilon)t} \tag{A.2}$$

for all $t\ge T$ and some constant $C>0$. Let

$$x=-\sum_{t=0}^{\infty}\prod_{k=0}^{t}c^{-1}(k)\alpha^\top(t)\prod_{l=0}^{t-1}\hat{L}(l)y,$$

which in fact exists and is finite according to inequality (A.2). Then let

$$V_\phi=\left\{z=\left[\begin{array}{c}x\\ y\end{array}\right]:\ x+\sum_{t=0}^{\infty}\prod_{k=0}^{t}c^{-1}(k)\alpha^\top(t)\prod_{l=0}^{t-1}\hat{L}(l)y=0\right\}$$

be the transverse space. For any $\left[\begin{smallmatrix}x(0)\\ y(0)\end{smallmatrix}\right]\in V_\phi$,

$$x(t)=-\sum_{k=t}^{\infty}\prod_{p=t}^{k}c^{-1}(p)\alpha^\top(k)\prod_{q=0}^{k-1}\hat{L}(q)y(0).$$

Noting that there exists $T>0$ such that $\prod_{p=t}^{k}|c^{-1}(p)|\le e^{(-\lambda_0+\epsilon)(k-t)+2\epsilon t}$ for all $t\ge T$, we have

$$\begin{aligned}|x(t)|&\le\sum_{k=t}^{\infty}\prod_{p=t}^{k}|c^{-1}(p)|\|\alpha^\top(k)\|\left\|\prod_{q=0}^{k-1}\hat{L}(q)\right\|\|y(0)\|\\&\le\sum_{k=t}^{\infty}e^{(-\lambda_0+\epsilon)(k-t)}e^{2\epsilon t}e^{\epsilon k}e^{(\hat{\lambda}+\epsilon)k}\\&\le\left\{\sum_{k=t}^{\infty}e^{(-\lambda_0+\hat{\lambda}+3\epsilon)(k-t)}\right\}e^{(\hat{\lambda}+4\epsilon)t}\le M_2e^{(\hat{\lambda}+4\epsilon)t}\end{aligned}$$

for all $t\ge T$and some constants $M_2>0$. So, it can be concluded that

$$\overline{\lim_{t\to\infty}}\frac{1}{t}\log\|z(t-1)\|\le\hat{\lambda}+4\epsilon.$$

Since ϵ is chosen arbitrarily, there exists an $(m-1)$-dimensional subspace $V_\phi=\{z=[x\,y]^\top:\ x=-\sum_{t=0}^{\infty}\prod_{k=0}^{t}c^{-1}(k)\alpha^\top(t)\prod_{l=0}^{t-1}\hat{L}(l)y\}$ of which the largest Lyapunov exponent is less than $\hat{\lambda}$. The largest Lyapunov exponent of V_ϕ is clearly greater than $\hat{\lambda}$. Therefore, we conclude that $\hat{\lambda}$, i.e., $\log(\hat{\rho}(L))$, is the largest Lyapunov exponent of L except for λ_0. The proof is completed. ∎

REFERENCES

[1] P. Ashwin, J. Buescu, and I. Stewart, *From attractor to chaotic saddle: A tale of transverse instability*, Nonlinearity, 9 (1996), pp. 703–737.

[2] J. Milnor, *On the concept of attractors*, Commun. Math. Phys., 99 (1985), pp. 177–195.

[3] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization: A Universal Concept in Nonlinear Sciences*, Cambridge University Press, Cambridge, UK, 2001.

[4] K. Kaneko, *Theory and Applications of Coupled Map Lattices*, Wiley, New York, 1993.

[5] J. Jost and M. P. Joy, *Spectral properties and synchronization in coupled map lattices*, Phys. Rev. E, 65 (2001), article 016201.

[6] Y. H. Chen, G. Rangarajan, and M. Ding, *General stability analysis of synchronized dynamics in coupled systems*, Phys. Rev. E, 67 (2003), article 026209.

[7] F. M. Atay, J. Jost, and A. Wende, *Delays, connection topology, and synchronization of coupled chaotic maps*, Phys. Rev. Lett., 92 (2004), article 144101.

[8] W. Lu and T. Chen, *Synchronization of linearly coupled networks with discrete time systems*, Phys. D, 198 (2004), pp. 148–168.

[9] C. W. Wu, *Synchronization in networks of nonlinear dynamical systems coupled via a directed graph*, Nonlinearity, 18 (2005), pp. 1057–1064.

[10] W. Lu and T. Chen, *Global synchronization of linearly coupled Lipschitz map lattices with a directed graph*, IEEE Trans. Circuits Syst. II Express Briefs, 54 (2007), pp. 136–140.

[11] R. Olfati-Saber and R. M. Murray, *Consensus problems in networks of agents with switching topology and time-delays*, IEEE Trans. Automat. Control, 49 (2004), pp. 1520–1533.

[12] Y. Hatano and M. Mesbahi, *Agreement over random networks*, in Proceedings of the IEEE Conference on Decision and Control, 2004; available online from http://www.aa.washington.edu/faculty/mesbahi/papers/random-cdc.pdf.

[13] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Schochet, *Novel type of phase transitions in a system of self-driven particles*, Phys. Rev. Lett, 75 (1995), pp. 1226–1229.

[14] L. Moreau, *Stability of multiagent systems with time dependent communication links*, IEEE Trans. Automat. Control, 50 (2005), pp. 169–182.

[15] J. H. Lu and G. Chen, *A time-varying complex dynamical network model and its controlled synchronization criterion*, IEEE Trans. Automat. Control, 50 (2005), pp. 841–846.

[16] I. V. Belykh, V. N. Belykh, and M. Hasler, *Connection graph stability method for synchronized coupled chaotic systems*, Phys. D, 195 (2004), pp. 159–187.

[17] D. J. Stilwell, E. M. Bollt, and D. G. Roberson, *Sufficient conditions for fast switching synchronization in time-varying network topologies*, SIAM J. Appl. Dyn. Syst., 5 (2006), pp. 140–156.

[18] J. Hajnal, *Weak ergodicity in nonhomogeneous Markov chains*, Proc. Camb. Phil. Soc., 54 (1958), pp. 233–246.

[19] J. Hajnal, *The ergodic properties of nonhomogeneous finite Markov chains*, Proc. Camb. Phil. Soc., 52 (1956), pp. 67–77.

[20] J. Shen, *A geometric approach to ergodic nonhomogeneous Markov chains*, in Wavelet Analysis and Multiresolution Methods (Urbana-Champaign, IL, 1999), Lecture Notes in Pure and Appl. Math. 212, Dekker, New York, 2000, pp. 341–366.

[21] P. Bohl, *Über Differentialungleichungen*, J. Reine Angew. Math., 144 (1913), pp. 284–313.

[22] L. Barreira and Y. B. Pesin, *Lyapunov Exponents and Smooth Ergodic Theory*, University Lecture Series, AMS, Providence, RI, 2001.

[23] F. Colonius and W. Kliemann, *The Lyapunov spectrum of families of time-varying matrices*, Trans. Amer. Math. Soc., 348 (1996), pp. 4398–4408.

[24] J. Wolfwitz, *Products of indecomposable, aperiodic, stochastic matrices*, Proc. Amer. Math. Soc., 14 (1963), pp. 733–737.

[25] C. Godsil and G. Royle, *Algebraic Graph Theory*, Springer-Verlag, New York, 2001.

[26] J. Milnor, *On the concept of attractor: Correction and remarks*, Commun. Math. Phys., 102 (1985), pp. 517–519.

[27] I. Daubechies and J. C. Lagarias, *Sets of matrices all infinite product of which converge*, Linear Algebra Appl., 161 (1992), pp. 227–263.

[28] D. Serre, *Matrices: Theory and Applications*, Springer-Verlag, New York, 2002.

[29] F. M. Atay, T. Biyikoglu, and J. Jost, *Network synchronization: Spectral versus statistical properties*, Phys. D, 224 (2006), pp. 35–41.

[30] G. Gripenberg, *Computing the joint spectral radius*, Linear Algebra Appl., 234 (1996), pp. 43–60.

[31] Q. Chen and X. Zhou, *Characterization of joint spectral radius via trace*, Linear Algebra Appl., 315 (2000), pp. 175–188.

[32] X. Zhou, *Estimates for the joint spectral radius*, Appl. Math. Comput., 172 (2006), pp. 332–348.
[33] L. Arnold, *Random Dynamical Systems*, Springer-Verlag, Heidelberg, 1998.
[34] V. I. Oseledec, *A multiplicative ergodic theorem. Characteristic Ljapunov, exponents of dynamical systems*, Trans. Moscow Math. Soc., 19 (1968), pp. 197–231.
[35] L. Dieci and E. S. Van Vleck, *Computation of a few Lyapunov exponents for continuous and discrete dynamical systems*, Appl. Numer. Math., 17 (1995), pp. 275–291.
[36] R. Mainieri, *Zeta function for the Lyapunov exponent of a product of random matrices*, Phys. Rev. Lett., 68 (1992), pp. 1965–1968.
[37] A.-L. Barabasi and R. Albert, *Emergence of scaling in random networks*, Science, 286 (1999), pp. 509–512.

SIAM J. MATH. ANAL.
Vol. 39, No. 4, pp. 1260–1294

ON DIFFUSION INDUCED SEGREGATION IN TIME-DEPENDENT DOMAINS*

THOMAS BLESGEN†

Abstract. This article is concerned with the derivation and analysis of a model for diffusion induced segregation phenomena in the physically relevant case that the domain representing the crystal grows in time. A mathematical model is formulated where the phase parameter is a function of bounded variation and the equations are completed with the Gibbs–Thomson law. Based on suitable a priori bounds, methods from geometric measure theory are applied to derive suitable compactness properties which allow us to show the existence of weak solutions in three space dimensions.

Key words. phase transitions, dynamics of phase boundary, Gibbs–Thomson law

AMS subject classifications. 49Q20, 74N20, 80A22

DOI. 10.1137/060652336

1. Introduction.

Diffusion induced segregation (DIS) processes represent a particular class of phase change problems in solids that has been widely studied in mineralogy and crystallography. Typical of DIS is that the segregation starts only after the concentration of one selected component that diffuses into the single crystal from outside exceeds a certain threshold. DIS phenomena are very interesting for geological applications, as the segregation is irreversible and rock samples with DIS can be regarded as a natural geological clock. One prominent application of DIS is the attempt to estimate the time scale for magma ascending from the mantle of the earth by the investigation of a specimen with the so-called chalcopyrite disease within sphalerite. Mineralogical experiments on this particular example were first carried out in [5] and [6] under isothermal conditions. These articles explain conclusively the rearrangement of the lattice as well as the qualitative mechanism responsible for DIS. A collection of experiments revealing DIS phenomena is available in [5] and [6], and the results are compared with geological observations of DIS.

In order to get a better understanding of chalcopyrite disease within sphalerite, a first phase-field model based on partial differential equations with a modified Allen–Cahn equation was developed in [10]. More advanced simulations on the related ternary system of sphalerite, chalcopyrite, and cubanite are done in [11]. A general existence and uniqueness result for the mathematical formulation derived in [10] is contained in [7]. More sophisticated numerical simulations can be found in [8], where ab initio methods are used on a large scale to approximately compute the physical free energies. The results of the ab initio computations are validated with quantum mechanical and molecular dynamics calculations. In particular, the results in [8] provide *quantitative* predictions.

As is explained in [10], the mathematical model developed in earlier articles neglects the attachment of sulphur ions which lead to a growth of the crystal during

*Received by the editors February 17, 2006; accepted for publication (in revised form) May 18, 2007; published electronically December 12, 2007. The author thanks the German Research Community DFG for the financial support within the priority program 1095, Analysis, Modeling and Simulation of Multiscale Problems.
http://www.siam.org/journals/sima/39-4/65233.html
†Max-Planck-Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany (blesgen@mis.mpg.de).

the experiments. Instead, the domain Ω representing the crystal was assumed to be time-independent.

In this article we will close this gap. In the model presented in section 2, boundary conditions on the chemical potential are assumed which are close to the physical reality and which allow the domain to grow. These boundary conditions are connected to a generalized Gibbs–Thomson law. The rest of the paper is devoted to the proof of the existence of weak solutions. We follow the ansatz in [2] which is classical by now. In section 4, a time-discrete scheme is introduced.

We apply methods from geometric measure theory to show suitable a priori bounds and to establish the compactness in space and time of the time-discrete solution. The central argument is Lemma 7.7. In the subset of large discrete-interface velocities its proof is based on the construction of a Besicovitch-type covering that fails for space dimensions $n > 3$. In the set of a small discrete-interface velocity the proof relies on Bernstein's theorem, which is known to hold for space dimensions $n \leq 8$. The other key argument in the proof is Lemma 6.4, which requires $n \geq 3$; see the essential estimate (6.11). Due to these restrictions, the main lemmas are formulated for $n = 3$.

We mention that the employed techniques and results are related to the Stefan problem and the Mullins–Sekerka flow (see in particular the articles [18], [19], [17], [28], and [25]) and also can be transferred to applications in shape optimization problems.

For $h > 0$, time-discrete solutions are constructed in section 4, and it is shown in Theorem 7.4 that for a subsequence χ_h (the characteristic function of a set Ω_h^I) and a function $\chi \in L^1(\Omega_T, \{0, 1\})$,

$$\chi_h \to \chi \qquad \text{in } L^1(\Omega_T)$$

as $h \searrow 0$. Unfortunately, this does *not* imply that

$$|\nabla \chi_h| \rightharpoonup |\nabla \chi| \qquad \text{in } \mathrm{rca}(\Omega_T).$$

(Here $\mathrm{rca}(\Omega_T)$ denotes the space of all regular countable measures on Ω_T.) Therefore we require the following technical assumption from (8.1):

$$\int_{\Omega_T} |\nabla \chi_h| \to \int_{\Omega_T} |\nabla \chi| \qquad \text{as } h \searrow 0,$$

which is needed to prevent a loss of area when passing to the limit. Condition (8.1) is not new. In connection to the Stefan problem with Gibbs–Thomson law, it was stated previously in [19]. Condition (8.1) can also be found in [1].

In general, (8.1) does not hold and is violated because of the following concentration or oscillation effects at the reduced boundary:

1. Several parts of the boundary $\partial^* \Omega_h^I$ meet in the limit.
2. Oscillations of the boundary reduce the area in the limit.

One possible way to avoid condition (8.1) is to construct varifold solutions. For a two-phase flow described by the Navier–Stokes equations, this has been done in [22].

Finally, we remark that for the investigated mathematical system, we cannot expect the uniqueness of solutions. The reason is the same as for the Stefan problem with the Gibbs–Thomson law; see [18] for a proof.

2. Derivation of the model. Let $\Omega \subset \mathbb{R}^n$ be a box chosen large enough such that for times $0 \leq t \leq T$ with given stop time $T > 0$ a time-dependent set $\Omega^I = \Omega^I(t) \subset \mathbb{R}^n$ is contained in Ω. For the proof of existence of weak solutions, we will set

$n = 3$. We assume that Ω is a bounded domain with Lipschitz boundary. We call the set $\Omega^I(t)$ the *inner domain*, as it represents the growing crystal at time t surrounded by a second copper rich mineral. This second phase occupies $\Omega^{II} := \Omega \setminus \overline{\Omega^I}$ which we call the *outer domain*.

We introduce the two space-time cylinders $\Omega_T := \Omega \times (0,T)$, $\Omega_T^I := \Omega^I(t) \times (0,T)$, and by $\theta > 0$ we denote the constant temperature.

We introduce for $1 \le i \le 4$ functions $n_i : \Omega_T \to \mathbb{N}$ which determine the number of lattice positions in Ω_T occupied by species i. The n_i are related to the species by

$$n_1 \ldots \mathrm{Fe}^{3+}, \quad n_2 \ldots \mathrm{Fe}^{2+}, \quad n_3 \ldots \mathrm{Zn}^{2+}, \quad n_4 \ldots \mathrm{Cu}^{+}.$$

Similarly, by n_v we denote the number of vacant lattice positions.

We point out that even though the crystal grows due to the attachment of sulphur ions, the mathematical model does not require a variable for the sulphur concentration. This is because the attachment of S^{2-} is a consequence of the oxidation of Fe, leading to Shottky defects in the crystal. This is explained in detail in [7].

We introduce the vector $m = (m_1, \ldots, m_4)$, where

$$m_i(x,t) := \frac{n_i(x,t)}{\sum_{j=1}^4 n_j(x,t) + n_v(x,t)}, \quad 1 \le i \le 4,$$

is the number density of the ith constituent in Ω_T. We assume that we have a perfect crystal without impurities such that no further constituents need to be considered.

The variable m_5 specifies the electron density. The condition of electric neutrality leads to the formula

$$m_5 = 3m_1 + 2m_2 + 2m_3 + m_4 - 2. \tag{2.1}$$

The coefficients of m_i in (2.1) refer to the positive ionization of the ith constituent, and 2 is subtracted in the formula due to the attachment of S^{2-} ions.

In Ω_T^I the free energy density is given by

$$\begin{aligned} f_I(m) = k_B\theta \left[\sum_{i=1}^4 m_i \ln m_i + \left(1 - \sum_{i=1}^4 m_i\right) \ln \left(1 - \sum_{i=1}^4 m_i\right)\right] \\ + \sum_{i=1}^4 \sum_{j=1}^4 \alpha_{ij} m_i m_j + \sum_{i=1}^4 \beta_i m_i. \end{aligned} \tag{2.2}$$

The matrix $A := (\alpha_{ij})_{1 \le i,j \le 4}$ is symmetric and positive definite with constant coefficients, β_i are positive enthalpic constants, and k_B denotes the Boltzmann constant. The first term $\sum_i m_i \ln m_i$ in (2.2) is the entropic contribution to the free energy as it counts all possible configurations with a given vector m. The second summand with coefficients $\alpha_{ij} = \alpha_{ji}$ measures the elastic energy; i.e., $\alpha_{ij} m_i m_j$ is the contribution due to the interaction of ion i with ion j; see [10] for further details.

For the free energy of the outer phase and for given small $\delta > 0$, we make the ansatz

$$f_{II}(m) = k_B\theta \left[\sum_{i=1}^4 m_i \ln m_i - (\ln \delta + 1) \sum_{i=1}^2 m_i \right]. \tag{2.3}$$

The chemical potential of the ith constituent fulfills for $1 \le i \le 4$ (see (2.13) below)

$$\mu_i = \frac{\partial f_I}{\partial m_i}(m) \quad \text{in } \Omega^I, \qquad \mu_i = \frac{\partial f_{II}}{\partial m_i}(m) \quad \text{in } \Omega^{II}. \tag{2.4}$$

Ansatz (2.3) is chosen such that $\mu_i = k_B\theta \ln(m_i/\delta)$ and μ_i is positive in Ω^{II} for $m_i > \delta$ and $i = 1, 2$.

The oxidation process Fe^{3+}+e^- $\leftrightarrow$ Fe^{2+} in Ω^I_T is formally modeled as a reaction. The reaction vector r^I in Ω^I is given by (see [15] for a general explanation)

$$r^I = (r_1^I, -r_1^I, 0, 0), \qquad r_1^I = r_1^I(m) = k_1 m_2 - k_2 m_5 \tag{2.5}$$

with positive reaction rates k_1, k_2.

The conservation of mass leads to the formulation $\partial_t m_i = -\mathrm{div}(J_i) + r_i(m)$. Onsager's postulate [20], [21] states that the thermodynamic flux J is linearly related to the thermodynamic force. In our case, the thermodynamic forces are the negative chemical potential gradients, and we obtain the phenomenological equations (see [16, p. 137])

$$J_i = -\sum_{j=1}^{4} L_{ij} \nabla \mu_j, \quad 1 \le i \le 4, \tag{2.6}$$

with a mobility tensor $L = (L_{ij})_{1 \le i,j \le M}$ that may depend on μ. The Onsager reciprocity law [20], [21], [16] states that L has to be symmetric, which we assume in the following. To simplify the existence theory, we will further assume that L is positive definite.

The coefficients of L depend on the domain as $L = L^I$ in Ω^I and $L = L^{II}$ in Ω^{II}. The diffusion rates measured for Cu^+ and Zn^{2+} are of the same order and are about 1000 times larger than the diffusivities of Fe^{3+} and Fe^{2+}. Mathematically, we look at an idealized situation, where we set $L^{II}_{ij} := 0$ for $i \ne j$ (this means no cross diffusion) and L^{II}_{11}, $L^{II}_{22} \sim \varepsilon$ and L^{II}_{33}, $L^{II}_{44} \sim \frac{1}{\varepsilon}$. For small $\varepsilon > 0$, this gives rise to the following boundary conditions on $\partial\Omega^I$:

$$-\sum_{j=1}^{4} L^I_{ij}(\mu) \nabla \mu_j \vec{n} = m_i v \qquad \text{for } i = 1, 2, \tag{2.7}$$

$$\mu_i = \varphi_i \qquad \text{for } i = 3, 4. \tag{2.8}$$

The parameter v denotes the speed with which the interface moves outward, $\vec{n}$ is the unit outer normal to Ω^I, and φ_1, φ_2 are two given constants invariant in time and space.

We have the diffusion equations

$$\partial_t m_i = \mathrm{div}\left(\sum_{j=1}^{4} L^I_{ij} \nabla \mu_j\right) + r^I_i(m) \qquad \text{in } \Omega^I, \tag{2.9}$$

$$\partial_t m_i = \mathrm{div}\Big(L^{II}_{ii} \nabla \mu_i\Big) + r^{II}_i(\mu) \qquad \text{in } \Omega^{II}. \tag{2.10}$$

Experimentally it is observed that the outer domain contains only a very small amount of Fe^{3+}, Fe^{2+} except in a small strip near Ω^I. To ensure this condition for the mathematical system if the free boundary $\partial\Omega^I \cap \partial\Omega^{II}$ moves inward, i.e., if $v < 0$, we make the ansatz

$$r^{II}(\mu) = -\frac{1}{\gamma}(\mu_1, \mu_2, 0, 0), \tag{2.11}$$

where $\gamma > 0$ is a small constant related to the thickness of the Fe-containing strip close to Ω^I.

The model is formulated for small positive parameters γ, δ, and ε, but we will show in sections 5 and 7 that the existence theory remains valid in the limit $\gamma, \delta, \varepsilon \searrow 0$; see in particular the assumptions (A1)–(A4) in section 6 on the time-discrete problem.

We postulate that the set $\Omega_0^I := \Omega^I(t = 0)$ has a finite perimeter. For the characteristic function $\chi(\cdot, 0) : \Omega \to \{0, 1\}$ of $\Omega^I(t = 0)$, this means

$$\int_\Omega |\nabla\chi(t = 0)| = \|\chi(t = 0)\|_{BV(\Omega)} < \infty.$$

The condition $\|\chi\|_{BV(\Omega)} < \infty$ means that χ is a function of bounded variation in Ω; see [29], [30]. The symbol $H^{1,2}(\Omega) \subset L^2(\Omega)$ denotes the Sobolev space of functions with first weak derivatives in the Hilbert space $L^2(\Omega)$.

We consider a physical system with surface tension, so the total free energy F is given by

$$F(\chi, \mu) := \int_\Omega |\nabla\chi| + \int_{\partial\Omega^I} v + \int_\Omega (\chi f_I^*(\mu) + (1 - \chi) f_{II}^*(\mu)). \tag{2.12}$$

Here, $f_I^*(\mu)$, $f_{II}^*(\mu)$ are the Legendre–Fenchel transforms of $f_I(\mu)$, $f_{II}(\mu)$ defined by

$$f_I^*(\mu) := \sup_m \{\mu \cdot m - f_I(m)\}, \qquad f_{II}^*(\mu) := \sup_m \{\mu \cdot m - f_{II}(m)\}.$$

In these definitions, $\cdot$ denotes the inner product, i.e., $\mu \cdot m = \sum_{i=1}^4 \mu_i m_i$. If there can be no ambiguity, we drop $\cdot$ and simply write μm in this article.

The use of f_I^* and f_{II}^* exploits duality properties of the problem and allows us to formulate the free energy F as a function of μ and not of m. The Legendre–Fenchel transformation is a frequently used tool in mechanics and originates from convex analysis. General references to the Legendre–Fenchel transformation are [24] and [3]. In the context of diffusion problems, the ansatz (2.12) goes back to [4].

The mathematical description of the system is completed with the condition

$$F(\chi, \mu) \to \min, \tag{2.13}$$

where μ is fixed and fulfills the constraints (2.9), (2.10) and the minimum is sought for $\chi \in BV(\Omega; \{0, 1\})$. When restricting to smooth deformations of $\partial\Omega^I(t)$, the stationarity of F with respect to characteristic functions $\chi \in BV(\Omega; \{0, 1\})$ for fixed μ leads to the Gibbs–Thomson law

$$H + v = f_{II}^*(\mu) - f_I^*(\mu), \tag{2.14}$$

and H is the mean curvature of the interface $\partial\Omega^I(t)$. Below, (2.14) is replaced by the weaker condition (W2) from section 3 that requires less regularity on χ.

The vector m is obtained from μ by the splitting

$$m = \chi m_I + (1 - \chi) m_{II}, \tag{2.15}$$

where the mass vector m_I in the inner domain $\Omega^I(t)$ and the mass vector m_{II} in the outer domain are determined implicitly by

$$\mu = \frac{\partial f_I}{\partial m}(m_I) \quad \text{in } \Omega^I(t), \qquad \mu = \frac{\partial f_{II}}{\partial m}(m_{II}) \quad \text{in } \Omega \setminus \Omega^I(t). \tag{2.16}$$

The free energy densities f_I and f_{II} are strictly convex functions, so (2.16) is meaningful. The decomposition (2.15), (2.16) is essential for the further understanding of this article.

In summary, we are concerned with the free energy minimization (2.13) under the constraints

$$\partial_t(\chi m) = \operatorname{div}(\chi L^I(\mu)\nabla\mu) + \chi r^I(m) \qquad \text{in } \Omega, t > 0, \tag{2.17}$$

$$\partial_t((1-\chi)m) = \operatorname{div}((1-\chi)L^{II}(\mu)\nabla\mu) + (1-\chi)r^{II}(\mu) \qquad \text{in } \Omega, t > 0, \tag{2.18}$$

with r^I, r^{II} given by (2.5), (2.11), where $m(t)$ fulfills (2.15), (2.16), and equipped with the initial conditions

$$\chi(\cdot,0) = \chi_0 \qquad \text{in } \Omega, \tag{2.19}$$

$$\mu(\cdot,0) = \mu_0 \qquad \text{in } \Omega. \tag{2.20}$$

Equation (2.20) also determines $m(t=0)$ with (2.16). Motivated by (2.7) and (2.8), the functions of initial data $\chi_0 \in BV(\Omega)$ and $\mu_0 \in H^{1,2}(\Omega)$ must fulfill the compatibility conditions

$$\mu_{01} = \mu_{02} = 0 \qquad \text{in } \left\{ x \in \Omega \,\middle|\, \operatorname{dist}(x, \overline{\Omega^I(t=0)}) \geq \sqrt{\varepsilon} \right\}, \tag{2.21}$$

$$\mu_{03} = \psi_3, \quad \mu_{04} = \psi_4 \qquad \text{in } \Omega \setminus \overline{\Omega^I(t=0)}, \tag{2.22}$$

where ψ_3, $\psi_4 \in H^{1,2}(\Omega \setminus \overline{\Omega^I(t=0)})$. As the free energy in the outer phase depends on the parameter $\delta > 0$, we demand that, for a constant C independent of δ,

$$\int_{\Omega\setminus\overline{\Omega^I(t=0)}} f_{II}(m(t=0)) \leq C \qquad \text{uniformly in } \delta > 0. \tag{2.23}$$

The system is subject to the boundary conditions (2.7), (2.8), and

$$\mu_i = 0 \quad \text{for } i = 1,2 \qquad \text{on } \partial\Omega \times (0,T), \tag{2.24}$$

$$\mu_i = \varphi_i \quad \text{for } i = 3,4 \qquad \text{on } \partial\Omega \times (0,T). \tag{2.25}$$

For later use we want to introduce some notation. Let

$$\mu^N := (\mu_1, \mu_2), \qquad \mu^D := (\mu_3, \mu_4).$$

The vector m is decomposed in m^N, m^D accordingly.

We define $\hat{f}_{II}^{D,*}(m^N, \mu^D)$ as the partial Legendre–Fenchel transformation of $f_{II}(m)$ with respect to m, that is,

$$\hat{f}_{II}^{D,*}(m^N, \mu^D) := \sup_{m^D} \left\{ \mu^D m^D - f_{II}((m^N, m^D)) \right\}.$$

Due to the form of f_{II}, we may write

$$\hat{f}_{II}^{D,*}(m^N, \mu^D) = f_{II}^{D,*}(\mu^D) - f_{II}^N(m^N),$$

where we introduced the functions

$$f_{II}^N(m^N) := k_B\theta \left(\sum_{i=1}^{2} m_i \ln m_i - \sum_{i=1}^{2} (\ln\delta + 1) m_i \right),$$

$$f_{II}^{D,*}(\mu^D) := \sup_{m^D} \left\{ \mu^D m^D - f_{II}^D(m^D) \right\},$$

$$f_{II}^D(m^D) := k_B\theta \left(\sum_{i=3}^{4} m_i \ln m_i \right).$$

As a prerequisite to the existence theory, we mention the validity of the second law of thermodynamics, which in the isothermal setting reads

$$\frac{d}{dt} F(\chi, \mu) \leq 0.$$

For a direct proof of this inequality see [7]. So it holds that

$$\begin{aligned} \partial_t \int_\Omega \Big[\chi(f_I(m) - m\varphi) + (1-\chi)(f_{II}(m) - m\varphi) \Big] \\ \geq \partial_t \int_\Omega |\nabla\chi| + \partial_t \int_{\partial\Omega^I} v + \int_\Omega \Big(\chi L^I(\mu) + (1-\chi) L^{II}(\mu) \Big) |\nabla\mu|^2 \\ - \int_\Omega (\chi r^I(m) + (1-\chi) r^{II}(\mu))\mu. \end{aligned} \tag{2.26}$$

In (2.26) it holds that

$$- \int_\Omega r\mu := - \int_\Omega (\chi r^I + (1-\chi) r^{II})\mu \geq 0. \tag{2.27}$$

Physically, $-\int_\Omega r\mu$ is the production of entropy due to chemical reactions.

In order to show (2.27), we compute with (2.5)

$$- \sum_{i=1}^{4} r_i^I \mu_i = r_1^I(\mu_2 - \mu_1) = q_I(r_1^I, m_2) + q_I^*(\mu^N, m_2), \tag{2.28}$$

where

$$q_I(r_1^I, m_2) := k_B\theta \left[(k_1 m_2 - r_1^I) \ln \left(1 - \frac{r_1^I}{k_1 m_2} \right) + r_1^I \right],$$

$$q_I^*(\mu^N, m_2) := k_1 m_2 \left[\mu_2 - \mu_1 + k_B\theta \left(\exp\left(\frac{\mu_1 - \mu_2}{k_B\theta} \right) - 1 \right) \right].$$

The function q_I^* denotes, as above, the Legendre–Fenchel transform of q_I. We can check elementarily that the functions q_I and q_I^* are nonnegative, which proves (2.27).

Motivated by (2.11), we define for further use (see (2.26) and (4.3))

$$q_{II}^*(\mu^N) := -\frac{1}{2\gamma}|\mu^N|^2.$$

With (2.28) we have found a general formulation for the reaction terms based on duality. We remark that condition (2.27) can also be derived from an Arrhenius ansatz for the reaction rates; see [7], [9].

Integrating (2.26) with respect to time, we get the a priori estimate

$$\begin{aligned}(2.29)\quad &\sup_{t\in(0,T)}|\nabla\chi(t)| + \int_{\Omega_T}(\chi L^I(\mu) + (1-\chi)L^{II}(\mu))|\nabla\mu|^2 \\ &\quad - \int_{\Omega_T}(\chi r^I(m) + (1-\chi)r^{II}(\mu))\mu + \int_0^T\int_{\partial\Omega^I(t)} v \\ &\leq \int_\Omega \Big(\chi f_I(m) + (1-\chi)f_{II}(m)\Big)(t=T) + \int_\Omega m(0)\varphi - \int_\Omega m(T)\varphi \\ &\quad - \int_\Omega \Big(\chi f_I(m) + (1-\chi)f_{II}(m)\Big)(t=0).\end{aligned}$$

Since $f_{II}(t=0)$ is bounded due to estimate (2.23), we get the boundedness of the right-hand side of (2.29). As L^I, L^{II} are positive definite, we obtain estimates for the functions χ, μ and velocity v. These estimates will be improved later.

3. Weak solutions. We want to briefly specify the class of solutions we are looking for. We call the triple (m, μ, χ) a *weak solution* of the system (2.17)–(2.25) if (2.13) holds, $m = m(\mu)$ is given by (2.15), (2.16), and if (μ, χ) solve the following weak formulations:

(W1) For all $\xi \in C^1([0,T]\times\overline{\Omega};\mathbb{R}^3)$ with $\xi(T) = 0$, it holds that

$$-\int_{\Omega_T}\chi m\partial_t\xi = -\int_{\Omega_T}\chi L^I(\mu)\nabla\mu\nabla\xi + \int_{\Omega_T}\chi r_I(m)\xi + \int_\Omega \chi(t{=}0)m(t{=}0)\xi(t{=}0),$$

$$\begin{aligned}-\int_{\Omega_T}(1-\chi)m\partial_t\xi = &-\int_{\Omega_T}(1-\chi)L^{II}(\mu)\nabla\mu\nabla\xi + \int_{\Omega_T}(1-\chi)r^{II}(\mu)\xi \\ &+ \int_\Omega(1-\chi(t=0))m(t=0)\xi(t=0).\end{aligned}$$

(W2) For all $\zeta \in C^1([0,T]\times\overline{\Omega};\mathbb{R}^3)$ with $\zeta = 0$ on $\partial\Omega\times(0,T)$, it holds that

$$\int_{\Omega_T}\left(\mathrm{div}\zeta - \frac{\nabla\chi}{|\nabla\chi|}\nabla\zeta\frac{\nabla\chi}{|\nabla\chi|}\right)|\nabla\chi| + \int_{\Omega_T} v\zeta\nabla\chi = \int_{\Omega_T}(f_{II}^*(\mu) - f_I^*(\mu))\zeta\nabla\chi.$$

(W3) For all $\xi \in C^1([0,T]\times\overline{\Omega};\mathbb{R})$ with $\xi = 0$ on $\partial\Omega\times(0,T)$ and $\xi(T) = 0$, it holds that

$$\int_{\Omega_T}\chi\partial_t\xi + \int_\Omega \chi_{\Omega_0^I}\xi(0) = -\int_{\Omega_T} v\xi|\nabla\chi|.$$

In [19], further explanations can be found concerning (W2) and (W3).

4. Time-discrete scheme. For fixed $h > 0$, we consider in Ω the time-discrete scheme

$$\chi(t)m(t) - \chi(t-h)m(t-h) = h\,\mathrm{div}[\chi(t-h)L^I(\mu(t-h))\nabla\mu(t)] + \chi(t-h)hr^I(m(t)), \tag{4.1}$$

$$(1-\chi(t))m(t) - (1-\chi(t-h))m(t-h) = h\,\mathrm{div}[(1-\chi(t-h))L^{II}(\mu(t-h))\nabla\mu(t)] + (1-\chi(t-h))hr^{II}(\mu(t)). \tag{4.2}$$

This is an implicit time discretization, except for the coefficients $L^I(\mu)$, $L^{II}(\mu)$, and χ which are treated explicitly. Also we briefly wrote $\mu(t)$ for $\mu(\chi(t))$ and set $\chi(t) := \chi_0$, $m(t) := m(t=0)$ for $-h \le t < 0$.

For given $\chi(t-h)$, $\mu(t-h)$, let the discrete free energy functional be given by

$$\begin{aligned} F_h(\chi(t),\mu(t)) = &\int_\Omega |\nabla\chi(t)| + \int_{\Omega^I(t)\triangle\Omega^I(t-h)} \frac{1}{h}\mathrm{dist}(\cdot,\partial\Omega^I(t-h)) + \int_\Omega m(t-h)\mu(t) \\ &+\int_\Omega \Big[\chi(t)f_I^*(\mu(t)) + (1-\chi(t))f_{II}^*(\mu(t))\Big] \\ &+ h\int_\Omega \Big[\chi(t-h)L^I(\mu(\chi(t-h))) + (1-\chi(t-h))L^{II}(\mu(\chi(t-h)))\Big]|\nabla\mu(t)|^2 \\ &+\int_\Omega \chi(t-h)hq_I^*(\mu^N(t), m_2(t)) - \int_\Omega (1-\chi(t-h))hq_{II}^*(\mu^N(t)). \end{aligned} \tag{4.3}$$

Here we used the notation $A\triangle B := (A\setminus B)\cup(B\setminus A)$ for the symmetrized difference of two sets A and B.

We construct the time-discrete solution $\chi_h \in L^\infty(0,T;\, BV(\Omega;\{0,1\}))$ in the following way. Let $T = hN$. At time $t = 0$, $\Omega^I(t=0)$ and μ_0 are given. For discrete-time values $t = kh$, $k = 1,\ldots,N$, the function $\chi_h(t)$ with $\mu_h(t) = \mu_h(\chi_h(t))$ fulfilling the constraints (4.1), (4.2) iteratively solves the energy minimization problem

$$F_h(\chi_h(t),\mu_h) \to \min \tag{4.4}$$

in the class $BV(\Omega;\{0,1\})$. The discrete mapping m_h is computed from χ_h, μ_h with the help of (2.15), (2.16). We continue χ_h to arbitrary $t \in (0,T)$ by setting

$$\chi_h(t) := \chi_h(kh) \qquad \text{for } t \in ((k-1)h, kh].$$

We introduce the discrete velocity of the interface by

$$v_h(x) := \frac{1}{h}\mathrm{dist}(x,\partial\Omega^I(t-h)). \tag{4.5}$$

The following lemma is a modification of an argument which was used (but not proved) in [19].

LEMMA 4.1 (weak mean curvature equation). *The minimum (χ_h,μ_h) of F_h satisfies the weak mean curvature equation*

$$\begin{aligned} &\int_\Omega \left(\mathrm{div}\zeta - \frac{\nabla\chi_h}{|\nabla\chi_h|}\nabla\zeta\frac{\nabla\chi_h}{|\nabla\chi_h|}\right)|\nabla\chi_h| + \int_\Omega v_h\zeta\nabla\chi_h \\ &\quad= \int_\Omega \Big(f_{II}^*(\mu_h(\chi_h)) - f_I^*(\mu_h(\chi_h))\Big)\zeta\nabla\chi_h \end{aligned}$$

for all $\zeta \in C_0^1(\Omega; \mathbb{R}^3)$.

Proof. We compute the first variation of F_h with respect to deformations of $\Omega^I(t)$; i.e., we compute

$$\frac{d}{ds} F_h(\chi_h \circ \zeta_s, \mu_h(\chi_h \circ \zeta_s))\Big|_{s=0}$$

for all $\zeta \in C_0^1(\Omega; \mathbb{R}^3)$ with $\zeta_0(x) = x$, $\partial_s \zeta_s(x) = \zeta(\zeta_s(x))$. With $\chi_h := \chi_h(t)$ we obtain

$$\begin{aligned}
&\int_\Omega \left(\mathrm{div}\zeta - \frac{\nabla \chi_h}{|\nabla \chi_h|} \nabla \zeta \frac{\nabla \chi_h}{|\nabla \chi_h|} \right) |\nabla \chi_h| + \int_\Omega v_h \zeta \nabla \chi_h \\
&+ \int_\Omega \Big(f_I^*(\mu_h(\chi_h)) - f_{II}^*(\mu_h(\chi_h)) \Big) \zeta \nabla \chi_h \\
&= - \int_\Omega m_h \mu_\zeta + \int_\Omega m_h(t-h) \mu_\zeta \\
&+ h \int_\Omega \mathrm{div} \Big[\Big(\chi_h(t-h) L^I(\mu_h(\chi_h(t-h))) + (1-\chi_h(t-h)) L^{II}(\mu_h(\chi_h(t-h))) \Big) \nabla \mu_h(\chi_h) \Big] \mu_\zeta \\
&+ h \int_\Omega \Big(\chi_h(t-h) r^I(m_h) + (1-\chi_h(t-h)) r^{II}(\mu_h(\chi_h)) \Big) \mu_\zeta,
\end{aligned}$$

where $\mu_\zeta := \mu_h(\chi_h(\zeta(\zeta_s)))$. As the right-hand side is zero, the lemma is proved. ∎

5. A priori estimates. In this section we show a priori estimates for the time-discrete solution. Together with the compactness results, they are the main ingredients for the existence proof.

Generically, we denote by C various constants that may change from estimate to estimate.

LEMMA 5.1 (a priori estimates for the time-discrete solution).

(i) *The following a priori estimate holds*:

(5.1)
$$\begin{aligned}
&\sup_{t \in (0,T)} \int_\Omega |\nabla \chi_h(t)| + \frac{1}{h^2} \int_0^T \int_{\Omega^I(t) \triangle \Omega^I(t-h)} \mathrm{dist}(\cdot, \partial\Omega^I(t-h)) \\
&+ \frac{1}{2\gamma} \int_{\Omega_T} (1 - \chi_h(t-h)) |\mu_h^N|^2 \\
&+ \int_{\Omega_T} \Big(\chi_h(t-h) L^I(\mu_h(\chi_h(t-h))) + (1-\chi_h(t-h)) L^{II}(\mu_h(\chi_h(t-h))) \Big) |\nabla \mu_h(\chi_h)|^2 \\
&\le \int_\Omega \Big(\chi_h f_I(m_h) + (1-\chi_h) f_{II}(m_h) \Big)(t = T) \\
&- \int_\Omega \Big(\chi_h f_I(m_h) + (1-\chi_h) f_{II}(m_h) \Big)(t = 0).
\end{aligned}$$

(ii) *For the chemical potential and the time derivative of* m, *it holds that*

$$\int_{\Omega_T} \chi_h(t-h) |\nabla \mu_h(t)|^2 \le C, \tag{5.2}$$

$$\int_{\Omega_T}(1-\chi_h(t-h))|\nabla\mu_h^N(t)|^2 \le C\varepsilon^{-1}, \tag{5.3}$$

$$\int_{\Omega_T}(1-\chi_h(t-h))|\nabla\mu_h^D(t)|^2 \le C\varepsilon, \tag{5.4}$$

$$\int_{\Omega_T}(1-\chi_h(t-h))|\mu_h^N(t)|^2 \le C\gamma, \tag{5.5}$$

$$\int_{\Omega_T}\|\chi_h\partial_t^h m_h\|^2_{H^{-1}(\Omega)} \le C. \tag{5.6}$$

Proof. (i) Due to the minimality of (χ_h,μ_h) with respect to F_h, we have

$$F_h(\chi_h(t),\mu_h(t)) \le F_h(\tilde\chi,\mu_h(\tilde\chi)) \qquad \text{for all } \tilde\chi\in BV(\Omega;\{0,1\}). \tag{5.7}$$

When choosing $\tilde\chi := \chi_h(t-h)$ in (5.7), we find

$$\int_\Omega|\nabla\chi_h(t)| + \int_{\Omega^I(t)\triangle\Omega^I(t-h)}\frac{1}{h}\mathrm{dist}(\cdot,\partial\Omega(t-h)) + \int_\Omega m_h(t-h)\mu_h(\chi_h(t))$$

$$-\int_\Omega\Big[\chi_h f_I(m_h(t)) + (1-\chi_h(t))f_{II}(m_h(t)) - m_h(t)\mu_h(\chi_h(t))\Big]$$

$$+h\int_\Omega\Big[\chi_h(t-h)L^I(\mu_h(\chi_h(t-h)))+(1-\chi_h(t-h))L^{II}(\mu_h(\chi_h(t-h)))\Big]|\nabla\mu_h(\chi_h(t))|^2$$

$$+h\int_\Omega\chi_h(t-h)q_I^*(\mu_h^N(\chi_h(t)),m_{h,2}(t)) - h\int_\Omega(1-\chi_h(t-h))q_{II}^*(\mu_h^N(\chi_h(t)))$$

$$\le \int_\Omega|\nabla\chi_h(t-h)| - \int_\Omega\Big[\chi_h(t-h)f_I(m_h(t-h)) + (1-\chi_h(t-h))f_{II}(m_h(t-h))\Big]$$

$$+h\int_\Omega\Big[\chi_h(t-h)L^I(\mu_h(\chi_h(t-h)))+(1-\chi_h(t-h))L^{II}(\mu_h(\chi_h(t-h)))\Big]|\nabla\mu_h(\chi_h(t-h))|^2$$

$$+h\int_\Omega\chi_h(t-h)q_I^*(\mu_h^N(\chi_h(t-h)),m_{h,2}(t-h)) - h\int_\Omega(1-\chi_h(t-h))q_{II}^*(\mu_h^N(\chi_h(t-h))).$$

We rewrite this and first observe that q_I, $q_I^*\ge 0$, $q_{II}^*\le 0$. Additionally,

$$-\int_\Omega\chi_h(t)f_I(m_h(t)) - \chi_h(t-h)f_I(m_h(t-h))$$

$$-\int_\Omega(1-\chi_h(t))f_{II}(m_h(t)) - (1-\chi_h(t-h))f_{II}(m_h(t-h))$$

$$= -\int_\Omega\Big[\chi_h(t)\Big(f_I(m_h(t))-f_I(m_h(t-h))\Big) + (1-\chi_h(t))\Big(f_{II}(m_h(t))-f_{II}(m_h(t-h))\Big)\Big]$$

$$-\int_\Omega h\partial_t^h\chi_h(t)\Big(f_I(m_h(t-h)) - f_{II}(m_h(t-h))\Big).$$

The second integral on the right is nonnegative due to (4.4). The first integral on the right, from the definition of the chemical potential μ, can be estimated from above by

$$-\int_\Omega \mu_h(\chi_h(t))m_h(t-h).$$

With this result, $F_h(\chi_h(t),\mu_h(t)) \le F_h(\chi_h(t-h),\mu_h(\chi_h(t-h)))$ finally becomes

(5.8)

$$\begin{aligned}
&\partial_t^h \int_\Omega |\nabla\chi_h| + \frac{1}{h^2}\int_{\Omega^I(t)\triangle\Omega^I(t-h)} \operatorname{dist}(\cdot,\partial\Omega(t-h)) + \frac{1}{2\gamma}\int_\Omega (1-\chi_h(t-h))|\mu_h^N|^2 \\
&\quad + \int_\Omega \Big(\chi_h(t-h)L^I(\mu_h(\chi_h(t-h))) + (1-\chi_h(t-h))L^{II}(\mu_h(\chi_h(t-h)))\Big)|\nabla\mu_h(\chi_h)|^2 \\
&\le \partial_t^h\left(\int_\Omega \chi_h f_I(m_h) + \int_\Omega (1-\chi_h)f_{II}(m_h)\right).
\end{aligned}$$

Integration in time proves (i).

(ii) This follows directly from (i) and (4.1), (4.2). ∎

6. A density lemma. Now we show a density lemma that establishes a strong geometric property of the sets $\Omega^I(t)$.

LEMMA 6.1. *For any $\mu \in H^{1,2}(\Omega;\mathbb{R}^4)$, it holds that*

$$\|f_{II}^*(\mu) - f_I^*(\mu)\|_{L^\infty(\Omega_T)} \le \|f_I\|_{L^\infty((0,T)\times\mathbb{R}^4)} + \|f_{II}\|_{L^\infty((0,T)\times\mathbb{R}^4)}.$$

Proof. Directly from the definitions of f_I^* and f_{II}^*, we see that

$$f_{II}^*(\mu) - f_I^*(\mu) = \sup_{\tilde m}\{\tilde m\cdot\mu - f_{II}(\tilde m)\} - \sup_{\tilde m}\{\tilde m\cdot\mu - f_I(\tilde m)\} \le f_I(m) - f_{II}(m).$$

The analogous estimate of $f_I^*(\mu) - f_{II}^*(\mu)$ gives

$$\|f_{II}^*(\mu) - f_I^*(\mu)\|_{L^\infty} \le \|f_I\|_{L^\infty} + \|f_{II}\|_{L^\infty},$$

which proves the lemma. ∎

We recall the following results from geometric measure theory. For a proof see, for instance, [29], [13].

THEOREM 6.2 (trace operator in BV). *Let $\Omega \subset \mathbb{R}^n$ be an open bounded set and let $\partial\Omega$ be Lipschitz. Then there exists a trace operator* $\operatorname{tr} : BV(\Omega) \to L^1(\partial\Omega, d\mathcal{H}^{n-1})$ *such that for given $f \in BV(\Omega)$,*

$$\int_\Omega f\operatorname{div}\varphi = -\int_\Omega \varphi\cdot\nabla f + \int_{\partial\Omega} \operatorname{tr} f(\varphi\cdot\vec n)d\mathcal{H}^{n-1} \qquad \text{for all } \varphi \in C^1(\mathbb{R}^n,\mathbb{R}^n),$$

where $\vec n$ is the unit outer normal to $\partial\Omega$. Furthermore,

$$\lim_{\varrho\searrow 0}\left(\varrho^{-n}\int_{B_\varrho(x)\cap\Omega} |f(y) - \operatorname{tr} f(x)|dy\right) = 0 \qquad \text{for } d\mathcal{H}^{n-1} - \text{almost all } x \in \partial\Omega.$$

If $E \subset\subset \Omega$ is an open set with a Lipschitz boundary, then $f_{|E} \in BV(E)$ and $f_{|\Omega\setminus\overline{E}} \in BV(\Omega\setminus\overline{E})$ have traces on ∂E. In what follows, we write $f_E^- := \operatorname{tr}(f_{|E})$ and $f_E^+ := \operatorname{tr}(f_{|\Omega\setminus\overline{E}})$. We mention for later use the equality

$$\int_{\partial E} |f_E^- - f_E^+|d\mathcal{H}^{n-1} = \int_{\partial E} |\nabla f|d\mathcal{H}^{n-1}. \tag{6.1}$$

Finally, by χ_E we denote the characteristic function of a set E.

THEOREM 6.3 (isoperimetric inequalities). *Let $\Omega \subset \mathbb{R}^n$ be a bounded set with $\chi_\Omega \in BV(\mathbb{R}^n)$. Then it holds that*

$$\gamma_n \left(\int_{\mathbb{R}^n} \chi_\Omega \right)^{\frac{n-1}{n}} \leq \int_{\mathbb{R}^n} |\nabla \chi_\Omega|, \tag{6.2}$$

$$\min \left\{ \left(\int_{B_\varrho(x)} \chi_\Omega \right)^{\frac{n}{n-1}}, \left(\int_{B_\varrho(x)} \chi_\Omega^c \right)^{\frac{n}{n-1}} \right\} \leq C(n) \int_{B_\varrho(x)} |\nabla \chi_\Omega| \quad \textit{for all } B_\varrho(x) \subset \mathbb{R}^n, \tag{6.3}$$

where $\gamma_n := n\omega_n^{\frac{1}{n}}$ and $\omega_n := \mathcal{L}^n(B_1(0))$ are dimensional constants.

Now we show the following geometric property of the sets $\Omega^I(t)$.

LEMMA 6.4 (density lemma). *Let $\kappa_1 := \frac{1}{4^3}$ and let $\chi_h(t)$ be a minimizer of the functional $F_h(\chi_h(t), \mu_h(\chi_h(t)))$. Then for all $x \in \partial\Omega^I(t)$ and for all $\varrho > 0$ which satisfy*

$$\varrho \leq \frac{\gamma_3}{2\omega_3^{\frac{1}{3}} (\|f_I\|_{L^\infty} + \|f_{II}\|_{L^\infty} + h^{-1} \|\mathrm{dist}(\cdot, \partial\Omega^I(t-h))\|_{L^\infty(\Omega^I(t) \triangle \Omega^I(t-h))})} \tag{6.4}$$

it holds that

$$\kappa_1 \leq \omega_3^{-1} \varrho^{-3} \int_{B_\varrho(x)} \chi_h(t) \leq 1 - \kappa_1. \tag{6.5}$$

Proof. (i) Estimate from below.

Let $\varrho > 0$ satisfy (6.4) and $x \in \partial\Omega^I(t)$ be given. Recall that $\chi_{\Omega^I(t)} = \chi_h(t)$ and recall the notions χ_h^+, χ_h^- introduced after Theorem 6.2. Consider generally $\Omega \subset \mathbb{R}^n$.

First, we show that $\Omega^I(t) \cap B_\varrho(x)$ consists of only one radial component, i.e.,

$$\int_{\partial B_r(x)} \chi_h^- d\mathcal{H}^{n-1} + \int_{\partial B_r(x)} \chi_h^+ d\mathcal{H}^{n-1} > 0 \qquad \text{for all } 0 < r < \varrho. \tag{6.6}$$

Assume (6.6) does not hold for a radius $r \in (0, \varrho)$. We use (5.7) with $\tilde{\chi} := \chi_{\Omega^I(t) \setminus B_r(x)}$ and find

$$\begin{aligned} \int_\Omega |\nabla \chi_h(t)| - \int_\Omega |\nabla \tilde{\chi}| &\leq \int_{B_r(x)} \frac{1}{h} \|\mathrm{dist}(\cdot, \partial\Omega^I(t-h))\|_{L^\infty(\Omega^I \triangle \Omega^I(t-h))} \chi_h(t) \\ &\quad + \int_{B_r(x)} \Big(f_{II}^*(\mu_h(\chi_h(t))) - f_I^*(\mu_h(\chi_h(t))) \Big) \chi_h(t) \\ &\leq \int_{B_r(x)} C_1(h) \chi_h(t), \end{aligned} \tag{6.7}$$

where, led by Lemma 6.1, we introduced the constant

$$C_1(h) := \|f_I\|_{L^\infty} + \|f_{II}\|_{L^\infty} + h^{-1} \|\mathrm{dist}(\cdot, \partial\Omega^I(t-h))\|_{L^\infty(\Omega^I(t) \triangle \Omega^I(t-h))}.$$

Now we use the equalities

$$\int_{\mathbb{R}^n} |\nabla \chi_{\Omega^I(t)\cap B_r(x)}| = \int_{B_r(x)} |\nabla \chi_{\Omega^I(t)}| + \int_{\partial B_r(x)} \chi^-_{\Omega^I(t)} d\mathcal{H}^{n-1},$$

$$\int_{\mathbb{R}^n} |\nabla \chi_{\Omega^I(t)\setminus B_r(x)}| = \int_{\mathbb{R}^n\setminus B_r(x)} |\nabla \chi_{\Omega^I(t)}| + \int_{\partial B_r(x)} \chi^+_{\Omega^I(t)} d\mathcal{H}^{n-1}$$

and the isoperimetric inequality (6.2) to conclude that

$$\begin{aligned}
&\int_{B_r(x)} C_1(h)\chi_{\Omega^I(t)} \\
&\quad \geq \int_{\mathbb{R}^n} |\nabla \chi_{\Omega^I(t)}| - \int_{\mathbb{R}^n\setminus B_r(x)} |\nabla \chi_{\Omega^I(t)}| - \int_{\partial B_r(x)} \chi^+_{\Omega^I(t)} d\mathcal{H}^{n-1} \\
&\quad = \int_{B_r(x)} |\nabla \chi_{\Omega^I(t)}| - \int_{\partial B_r(x)} \chi^+_{\Omega^I(t)} d\mathcal{H}^{n-1} \\
&\quad = \int_{\mathbb{R}^n} |\nabla \chi_{\Omega^I(t)\cap B_r(x)}| - \int_{\partial B_r(x)} \chi^+_{\Omega^I(t)} d\mathcal{H}^{n-1} - \int_{\partial B_r(x)} \chi^-_{\Omega^I(t)} d\mathcal{H}^{n-1} \\
&\quad \geq \gamma_n \left(\int_{B_r(x)} \chi_{\Omega^I(t)}\right)^{\frac{n-1}{n}} - \int_{\partial B_r(x)} \chi^+_{\Omega^I(t)} d\mathcal{H}^{n-1} - \int_{\partial B_r(x)} \chi^-_{\Omega^I(t)} d\mathcal{H}^{n-1}.
\end{aligned}$$

Since we assumed that (6.6) is false, the last two boundary integrals on the right can be estimated from below by zero, and we end up with

$$\gamma_n \leq C_1(h) \left(\int_{B_r(x)} \chi_{\Omega^I(t)}\right)^{\frac{1}{n}} \leq C_1(h)\omega_n^{\frac{1}{n}}\varrho.$$

This is a contradiction to (6.4), so (6.6) is proved.

A similar technique is now used for proving the bound from below in (6.5). This time, we use in (5.7) the comparison function

$$\tilde{\chi} := \chi_{\Omega^I(t)\setminus(B_{\frac{\varrho}{2}+\sigma}(x)\setminus B_{\frac{\varrho}{2}-\sigma}(x))}$$

with $0 < \sigma < \frac{\varrho}{2}$. Let

$$V(\sigma) := \mathcal{L}^n\Big(\Omega^I(t) \cap (B_{\frac{\varrho}{2}+\sigma}(x) \setminus B_{\frac{\varrho}{2}-\sigma}(x))\Big).$$

With this choice of $\tilde{\chi}$, (5.7) yields

$$\int_\Omega |\nabla \chi_{\Omega^I(t)}| - \int_\Omega |\nabla \tilde{\chi}| \leq C_1(h)V(\sigma). \tag{6.8}$$

Similar to the above, we have the identities

$$\int_{\mathbb{R}^n} |\nabla \chi_{\Omega^I(t)\cap(B_{\frac{\varrho}{2}+\sigma}(x)\setminus B_{\frac{\varrho}{2}-\sigma}(x))}|$$
$$= \int_{B_{\frac{\varrho}{2}+\sigma}(x)\setminus B_{\frac{\varrho}{2}-\sigma}(x)} |\nabla \chi_{\Omega^I(t)}| + \int_{\partial B_{\frac{\varrho}{2}+\sigma}(x)} \chi^-_{\Omega^I(t)} d\mathcal{H}^{n-1}$$
$$+ \int_{\partial B_{\frac{\varrho}{2}-\sigma}(x)} \chi^+_{\Omega^I(t)} d\mathcal{H}^{n-1},$$

$$\int_{\mathbb{R}^n} |\nabla \chi_{\Omega^I(t)\setminus(B_{\frac{\varrho}{2}+\sigma}(x)\setminus B_{\frac{\varrho}{2}-\sigma}(x))}|$$
$$= \int_{\mathbb{R}^n\setminus(B_{\frac{\varrho}{2}+\sigma}(x)\setminus B_{\frac{\varrho}{2}-\sigma}(x))} |\nabla \chi_{\Omega^I(t)}| + \int_{\partial B_{\frac{\varrho}{2}+\sigma}(x)} \chi^+_{\Omega^I(t)} d\mathcal{H}^{n-1}$$
$$+ \int_{\partial B_{\frac{\varrho}{2}-\sigma}(x)} \chi^-_{\Omega^I(t)} d\mathcal{H}^{n-1}.$$

With these two equalities and (6.8), we compute with (6.2)

$$C_1(h)V(\sigma) \geq \gamma_n V(\sigma)^{\frac{n-1}{n}} - \int_{\partial B_{\frac{\varrho}{2}+\sigma}(x)} \chi^-_{\Omega^I(t)} d\mathcal{H}^{n-1} - \int_{\partial B_{\frac{\varrho}{2}-\sigma}(x)} \chi^+_{\Omega^I(t)} d\mathcal{H}^{n-1}$$
$$- \int_{\partial B_{\frac{\varrho}{2}+\sigma}(x)} \chi^+_{\Omega^I(t)} d\mathcal{H}^{n-1} - \int_{\partial B_{\frac{\varrho}{2}-\sigma}(x)} \chi^-_{\Omega^I(t)} d\mathcal{H}^{n-1} \tag{6.9}$$
$$= \gamma_n V(\sigma)^{\frac{n-1}{n}} - 2\frac{d}{d\sigma}V(\sigma). \tag{6.10}$$

The last equality holds for almost all $0 < \sigma < \frac{\varrho}{2}$.

From the definition of $V(\sigma)$ and the upper bound (6.4) on ϱ, we directly find

$$V(\sigma) \leq \omega_n \varrho^n \leq \frac{\gamma_n^n}{((n-1)C_1(h))^n},$$

implying at once

$$C_1(h) \leq \frac{\gamma_n}{(n-1)V(\sigma)^{\frac{1}{n}}}.$$

Using the last inequality in (6.10) finally shows the crucial estimate

$$\frac{d}{d\sigma}V(\sigma) \geq \frac{\gamma_n}{2}V(\sigma)^{\frac{n-1}{n}} - \frac{\gamma_n}{2(n-1)}V(\sigma)^{\frac{n-1}{n}} = \frac{\gamma_n}{2}\frac{n-2}{n-1}V(\sigma)^{\frac{n-1}{n}}. \tag{6.11}$$

We integrate (6.11) from $0 < \sigma < \frac{\varrho}{2}$. We see directly

$$\int_0^{\frac{\varrho}{2}} \frac{d}{d\sigma}V(\sigma)\,d\sigma = V(\varrho/2) - V(0) = \int_{B_\varrho(x)} \chi_h(t).$$

For $n = 3$ the integration of (6.11) therefore yields

$$\int_{B_\varrho(x)} \chi_h(t) \geq \frac{\omega_3}{4^3}\varrho^3 = \kappa_1\omega_3\varrho^3,$$

which proves the lower bound in (6.5).

(ii) The upper bound in (6.5) can be derived in the same way as in (i) by replacing $\Omega^I(t)$ with its complement. Obviously,

$$\int_\Omega |\nabla\chi_{\Omega^I(t)}| = \int_\Omega |\nabla\chi_{\mathbb{R}^n\setminus\Omega^I(t)}|.$$

Now we consider minimizers $\tilde{\chi}(t)$ of $\tilde{F}_h(\tilde{\chi}, \tilde{\mu}(\tilde{\chi}))$, where the functional $\tilde{F}_h$ is defined as F_h in (4.3), but with χ replaced by $1-\chi$ and $1-\chi$ replaced by χ.

Then we redo the proof of (i) for the functional $\tilde{F}_h$. We obtain

$$\omega_n^{-1}\varrho^{-n}\int_{B_\varrho(x)} \chi_h(t) \leq 1-\kappa_1.$$

This shows the bound from above in (6.5) and ends the proof. $\square$

COROLLARY 6.5. *Let $\chi_h := \chi_{\Omega^I(t)}$ be a minimizer of the functional $F_h(\chi, \mu(\chi))$ and let $\kappa_2 := \frac{1}{2^3}$. Then for any $\varrho > 0$ and any pair (x, y) with $x \in \partial\Omega^I(t)$, $y \in B_\varrho(x)$, and*

$$\|f_I\|_{L^\infty} + \|f_{II}\|_{L^\infty} \leq \frac{1}{h}\mathrm{dist}(y, \partial\Omega^I(t-h)), \tag{6.12}$$

it holds that

$$\kappa_2 \leq \omega_3^{-1}\varrho^{-3}\int_{B_\varrho(x)} \chi_h(t) \qquad \text{if } x \in \overline{\Omega^I(t)}\setminus\Omega^I(t-h),$$

$$\omega_3^{-1}\varrho^{-3}\int_{B_\varrho(x)} \chi_h(t) \leq 1-\kappa_2 \qquad \text{if } x \in \Omega^I(t-h)\setminus\overline{\Omega^I(t)}.$$

The assertion remains valid if the assumption $x \in \partial\Omega^I(t)$ is replaced by $x \in \Omega$.

Proof. The proof is again similar to the proof of Lemma 6.4. Let $\varrho > 0$ satisfy (6.12). First, we consider the case $x \in \overline{\Omega^I(t)}\setminus\Omega^I(t-h)$ and show that (6.6) holds. First consider generally $\Omega \subset \mathbb{R}^n$. We exploit (5.7) with $\tilde{\chi} := \chi_{\overline{\Omega^I(t)}\setminus B_\varrho(x)}$. Assuming that (6.6) is violated for a radius $r \in (0, \varrho)$, we obtain, analogous to (6.7),

$$\int_\Omega |\nabla\chi_{\Omega^I(t)}| - \int_\Omega |\nabla\tilde{\chi}| \leq \int_{B_r(x)} C_2(h, y)\chi_{\Omega^I(t)}dy \leq 0, \tag{6.13}$$

where we have, according to (6.12),

$$C_2(h, y) := \|f_I\|_{L^\infty} + \|f_{II}\|_{L^\infty} - \frac{1}{h}\mathrm{dist}(y, \partial\Omega^I(t-h)) \leq 0.$$

From (6.13) and (6.2) we conclude that

$$\gamma_n\left(\int_{B_\varrho(x)} \chi_{\Omega^I(t)}\right)^{\frac{n-1}{n}} \leq 0,$$

which is a contradiction.

The analogue to (6.9) is

$$\gamma_n V(\sigma)^{\frac{n-1}{n}} - \int_{\partial B_{\frac{\varrho}{2}+\sigma}(x)} \left(\chi^-_{\Omega^I(t)} + \chi^+_{\Omega^I(t)}\right) d\mathcal{H}^{n-1}$$

$$- \int_{\partial B_{\frac{\varrho}{2}-\sigma}(x)} \left(\chi^-_{\Omega^I(t)} + \chi^+_{\Omega^I(t)}\right) d\mathcal{H}^{n-1} \leq 0,$$

and from that we can proceed as in the proof of Lemma 6.4.

The other case, $x \in \Omega^I(t-h) \setminus \overline{\Omega^I(t)}$, is treated similarly by considering the functional $\tilde{F}$ on the complementary set $(\Omega^I(t))^C$. ∎

Corollary 6.6. *Under the assumptions of Corollary* 6.5 *there exists a constant* C *independent of* h *such that for all* $\varrho > 0$,

$$\int_{B_\varrho(x)} |\nabla \chi_h| \leq C\varrho^2. \tag{6.14}$$

Proof. The proof is analogous to Corollary 6.5. Let $x \in \overline{\Omega^I(t)} \setminus \Omega^I(t-h)$. We use (5.7) with $\tilde{\chi} := \chi_{\overline{\Omega^I(t)} \setminus \Omega(t-h)}$. With condition (6.12) we find

$$\int_{B_\varrho(x)} |\nabla \chi_{\Omega^I(t)}| \leq \int_{\partial B_\varrho(x)} \chi^+_{\Omega^I(t)} d\mathcal{H}^2,$$

and therefore

$$\int_{B_\varrho(x)} |\nabla \chi_{\Omega^I(t)}| \leq \mathcal{H}^2(\partial B_\varrho(x)) \leq C\varrho^2.$$

The other case, $x \in \Omega^I(t-h) \setminus \Omega^I(t)$, can be proved in the same way. ∎

The model (2.17)–(2.25) was formulated for small positive parameters γ, δ, and ε. As will become clear in what follows, the existence theory remains valid if $\gamma = \gamma(h)$, $\delta = \delta(h)$, and $\varepsilon = \varepsilon(h)$ are functions that tend to 0 as $h \searrow 0$. The convergence of the discrete solution to the limit problem can be ensured, provided the following four conditions are met.

The functions $\delta(h) > 0$ and $\varepsilon(h) > 0$ fulfill

(A1) $\lim_{h\searrow 0} h^{\frac{1}{2}} \ln(\delta(h)) = 0$,

(A2) $\lim_{h\searrow 0} (\delta(h)\varepsilon^{-1}(h)) = 0$,

(A3) $\lim_{h\searrow 0} (h^{\frac{1}{4}}\varepsilon^{-\frac{3}{4}}(h)) = 0$,

(A4) $\lim_{h\searrow 0} (h^{\frac{1}{3}}\varepsilon^{-1}) = 0$.

Now we can formulate an estimate on the discrete velocity of the interface.

Lemma 6.7 (L^∞-bound on the discrete velocity). *For any* $h > 0$, *there exists a positive constant* C *independent of* h *such that*

$$\left\| \frac{1}{h} \mathrm{dist}(\cdot, \partial\Omega(t-h)) \right\|_{L^\infty(\Omega^I(t) \triangle \Omega^I(t-h))} < Ch^{-\frac{1}{2}}$$

uniformly in time.

Proof. Without loss of generality, we restrict our attention to the case $x \in \overline{\Omega^I(t)} \setminus \Omega^I(t-h)$. The other case, $x \in \Omega^I(t-h) \setminus \Omega^I(t)$, is treated in the same way.

The proof is done by contradiction. Assume that there exists an $x \in \partial\Omega^I(t)$ such that

$$\frac{1}{h}\operatorname{dist}(x, \partial\Omega^I(t-h)) \geq l h^{-\frac{1}{2}} \qquad \text{for all } l \in \mathbb{N}. \tag{6.15}$$

We use again (5.7) with $\tilde{\chi} := \chi_{\Omega^I(t)\setminus B_{\frac{l}{2}\sqrt{h}}(x)}$ to find

$$\begin{aligned}\int_\Omega |\nabla\tilde{\chi}| - \int_\Omega |\nabla\chi_{\Omega^I(t)}| \geq \int_{\Omega^I(t)\cap B_{\frac{l}{2}\sqrt{h}}(x)} \frac{1}{h}\operatorname{dist}(\cdot, \partial\Omega^I(t-h))\\ - \int_{B_{\frac{l}{2}\sqrt{h}}(x)} \Big(\|f_I\|_{L^\infty} + \|f_{II}\|_{L^\infty}\Big)\chi_{\Omega^I(t)}.\end{aligned} \tag{6.16}$$

The left-hand side of (6.16) can be estimated from above as follows:

$$\begin{aligned}&\int_\Omega |\nabla\chi_{\Omega^I(t)\setminus B_{\frac{l}{2}\sqrt{h}}(x)}| - \int_\Omega |\nabla\chi_{\Omega^I(t)}|\\ &= - \int_{B_{\frac{l}{2}\sqrt{h}}(x)} |\nabla\chi_{\Omega^I(t)}| + \int_{\partial B_{\frac{l}{2}\sqrt{h}}(x)} \chi^+_{\Omega^I(t)} d\mathcal{H}^{n-1}\\ &\leq \mathcal{H}^{n-1}(\partial B_{\frac{l}{2}\sqrt{h}}(x)) = n\left(\frac{l}{2}\sqrt{h}\right)^{n-1}\omega_n.\end{aligned}$$

By Assumption (6.15), the last estimate applied to (6.16) yields

$$\frac{l}{2}h^{-\frac{1}{2}}\int_{B_{\frac{l}{2}\sqrt{h}}(x)} \chi_{\Omega^I(t)} \leq n\left(\frac{l}{2}\sqrt{h}\right)^{n-1}\omega_n + \Big(\|f_I\|_{L^\infty} + \|f_{II}\|_{L^\infty}\Big)\left(\frac{l}{2}\sqrt{h}\right)^n.$$

From assumption (A1), it follows that condition (6.12) is fulfilled. For $n = 3$, we apply Corollary 6.5 and find

$$\frac{l}{2}h^{-\frac{1}{2}}\kappa_2\left(\frac{l}{2}\sqrt{h}\right)^3\omega_3 \leq 3\left(\frac{l}{2}\sqrt{h}\right)^2\omega_3 + \Big(\|f_I\|_{L^\infty} + \|f_{II}\|_{L^\infty}\Big)\left(\frac{l}{2}\sqrt{h}\right)^3.$$

This is a contradiction for sufficiently large l. ∎

LEMMA 6.8 (improved density lemma). *Let $\chi_h(t)$ be a minimizer of the functional $F_h(\chi_h, \mu_h(\chi_h))$. Then there exists a constant C independent of h such that for all $\varrho > 0$ with $\varrho \leq C\sqrt{h}$,*

$$\kappa_1 \leq \omega_3^{-1}\varrho^{-3}\int_{B_\varrho(x)} \chi_h(t) \leq 1 - \kappa_1 \qquad \textit{for all } x \in \partial\Omega^I(t). \tag{6.17}$$

Proof. This follows directly from Lemmas 6.4 and 6.7. ∎

7. Compactness properties of the discrete solution. We proceed with compactness properties of the time-discrete solution χ_h.

LEMMA 7.1 (compactness in space). *For every unit vector $e \in \mathbb{R}^3$, it holds uniformly in h that*

$$\lim_{s\searrow 0}\int_{\Omega_T} |\chi_h(x+se, t) - \chi_h(x,t)| = 0. \tag{7.1}$$

Proof. The claim follows from the a priori estimate (5.1) after observing that

$$\int_\Omega |\chi(\cdot + se) - \chi| \le s \int_\Omega |\nabla \chi|$$

for arbitrary functions $\chi \in BV(\Omega)$. ∎

LEMMA 7.2 (compactness of χ_h in time I). *There exists a constant C such that the discrete solution χ_h satisfies*

$$\int_0^{T-\tau} \int_\Omega |\chi_h(x, t+\tau) - \chi_h(x,t)| < c\tau \qquad \textit{for any } \tau > 0.$$

Proof. First, we consider the case $\tau = kh$ for $k \in \mathbb{N}$. Writing the integrand as a telescopic sum, we see that

$$\begin{aligned}
&\int_0^{T-\tau} \int_\Omega |\chi_h(x,t+\tau) - \chi_h(x,t)| \\
&\quad \le \int_0^{T-\tau} \int_\Omega \sum_{l=1}^k |\chi_h(x, t+lh) - \chi_h(x, t+(l-1)h)| \\
&\quad = \int_0^{T-\tau} \int_\Omega \sum_{l=1}^k h |\partial_t^{-h} \chi_h(x, t+lh)| \le \tau \int_0^T \int_\Omega |\partial_t^{-h} \chi_h(x,t)|.
\end{aligned}$$

Consequently, the proof is finished if we can show that

$$\int_{\Omega_T} |\partial_t^{-h} \chi_h| = \sum_{k=1}^N |\Omega^I(kh) \triangle \Omega^I((k-1)h)| < C \tag{7.2}$$

for a constant C which is independent of h. In order to prove (7.2), we notice that

$$\begin{aligned}
\Omega^I(t) \triangle \Omega^I(t-h) \subset &\Big\{ x \in \Omega^I(t) \triangle \Omega^I(t-h) \,\Big|\, \operatorname{dist}(x, \partial\Omega(t-h)) > c_1 h \Big\} \\
&\cup \Big\{ x \in \Omega^I(t) \triangle \Omega^I(t-h) \,\Big|\, \operatorname{dist}(x, \partial\Omega(t-h)) \le c_1 h \Big\} \\
&=: E_1(t) \cup E_2(t),
\end{aligned}$$

where c_1 is a small positive constant. It remains to shows that the sets $E_1(t)$, $E_2(t)$ for $t = lh$ and $1 \le l \le N$ are bounded. For $E_1(t)$ we have, as a consequence of the free energy inequality (2.26),

$$\begin{aligned}
|E_1(t)| &< \frac{1}{2c_1} \int_{\Omega^I(t) \triangle \Omega^I(t-h)} \operatorname{dist}(\cdot, \Omega^I(t-h)) \\
&\le h \partial_t^h \left[\int_\Omega \Big(\chi_h f_I(m_h) + (1-\chi_h) f_{II}(m_h) \Big) \right](t).
\end{aligned}$$

Discrete integration in time gives

$$\begin{aligned}
\sum_{l=1}^N |E_1(lh)| \le &\int_\Omega \Big(\chi_h f_I(m_h) + (1-\chi_h) f_{II}(m_h) \Big)(t = T) \\
&- \int_\Omega \Big(\chi_h f_I(m_h) + (1-\chi_h) f_{II}(m_h) \Big)(t = 0) \le C.
\end{aligned}$$

In order to show that $E_2(t)$ is bounded, we cover $E_2(t)$ with a family $\mathcal{B}$ of balls with radius $2c_1h$ and center points $x \in \partial\Omega^I(t-h)$. The Besicovitch covering lemma ensures that the covering can be chosen such that any point in $E_2(t)$ is contained in at most M different balls in $\mathcal{B}$, where $M \in \mathbb{N}$ is a fixed number.

With the help of the density lemma, Lemma 6.8, we see that there exists a constant $c_2 > 0$ such that

$$\int_{B_\varrho(x)} \chi_{\Omega^I(t)\triangle\Omega^I(t-h)} \leq \omega_3 \varrho^3 \leq c_2 \varrho \min\left\{ \left(\int_{B_\varrho(x)} \chi_{\Omega^I(t-h)} \right)^{\frac{2}{3}}, \left(\int_{B_\varrho(x)} \chi_{\mathbb{R}^3\setminus\Omega^I(t-h)} \right)^{\frac{2}{3}} \right\}. \tag{7.3}$$

The right-hand side of (7.3) can be bounded with the help of the isoperimetric inequality (6.3) which proves

$$\int_{B_\varrho(x)} \chi_{\Omega^I(t)\triangle\Omega^I(t-h)} \leq c_2 \varrho \int_{B_\varrho(x)} |\nabla \chi_{\Omega^I(t-h)}|.$$

This estimate holds for each ball $B \in \mathcal{B}$, and the union over all elements of $\mathcal{B}$ yields

$$|E_2(lh)| \leq c_1 c_2\, M h \int_\Omega |\nabla \chi_{\Omega^I((l-1)h)}|.$$

After summation, we get

$$\sum_{l=1}^{N} |E_2(lh)| \leq \sum_{l=1}^{N} c_2 h \int_\Omega |\nabla \chi_{\Omega^I((l-1)h)}| \leq C.$$

The generalization to arbitrary $\tau \in (0,T)$ is straightforward. $\square$

LEMMA 7.3 (compactness of χ_h in time II). *The discrete solution χ_h fulfills*

$$\int_\Omega |\chi_h(x,t+\tau) - \chi_h(x,t)| \leq C\tau^{\frac{1}{2}}$$

for all $h \leq \tau \leq T - t$.

Proof. Assume that $\tau = kh$ and $t = mh$ for $k, m \in \mathbb{N}$. As in Lemma 7.2, we conclude that

$$\int_\Omega |\chi_h(x,t+\tau) - \chi_h(x,t)| \leq \int_\Omega \sum_{l=0}^{k-1} |\chi_h(x,(m+l+1)h) - \chi_h(x,(m+l)h)| = \sum_{l=0}^{k-1} |\Omega^I((m+l+1)h) \triangle \Omega^I((m+l)h)|.$$

Therefore it is enough to prove

$$\sum_{l=0}^{k-1} |\Omega^I((m+l+1)h) \triangle \Omega^I((m+l)h)| \leq C\tau^{\frac{1}{2}}.$$

Here we consider the decomposition

$$\Omega^I(t)\triangle\Omega^I(t-h) \subset \left\{ x\in\Omega^I(t)\triangle\Omega^I(t-h) \,\middle|\, \mathrm{dist}(x,\partial\Omega(t-h)) > \frac{c_2}{4}h^{\frac{3}{2}} \right\}$$

$$\cup\left\{ x\in\Omega^I(t)\triangle\Omega^I(t-h) \,\middle|\, \mathrm{dist}(x,\partial\Omega(t-h)) \le \frac{c_2}{4}h^{\frac{3}{2}} \right\},$$

where c_2 is the constant of (7.3). The proof follows as in Lemma 7.2, where we can invoke Lemma 6.8 with $\varrho = \frac{c_2}{2}h^{\frac{1}{2}}$. We obtain

$$\sum_{l=0}^{k-1} |\Omega^I((m+l+1)h)\triangle\Omega^I((m+l)h)| \le c_2 h^{\frac{1}{2}} \le C\tau^{\frac{1}{2}}.$$

The generalization to arbitrary τ and t is straightforward. ◻

THEOREM 7.4 (compactness of χ_h). *There exist a subsequence of χ_h and a function $\chi \in L^1(\Omega_T,\{0,1\})$ such that*

$$\chi_h \to \chi \qquad \textit{in } L^1(\Omega_T).$$

For almost all $t\in(0,T)$, the function $\chi(t)$ is in $BV(\Omega)$ and is a characteristic function of a set $\Omega^I(t)\subset\Omega$. Additionally, we have the convergence of the Radon measures

$$\nabla\chi_h \rightharpoonup \nabla\chi \qquad \textit{in } \ \mathrm{rca}(\Omega_T).$$

Proof. This is a direct consequence of the compactness properties established in Lemmas 7.1 and 7.2. ◻

LEMMA 7.5 (estimate of the discrete velocity I). *There exists a constant C such that for sufficiently large speed s,*

$$\int_{\{|v_h|>s\}\cap\Omega_T} |v_h||\nabla\chi_h| < Cs^{-1}.$$

Proof. We restrict the proof to the case $x\in\overline{\Omega^I(t)}\setminus\Omega^I(t-h)$; case $x\in\Omega^I(t-h)\setminus\Omega^I(t)$ can be proved accordingly.

Let $t = kh$ with $k\in\mathbb{N}$. For given $s>0$, we fix $l\in\mathbb{N}$ and consider those $x\in\partial\Omega^I(t)$ with $x\in\partial\Omega^I(t)\cap\{2^l s<|v_h|\le 2^{l+1}s\}$.

We cover the set

$$\partial\Omega^I(t)\cap\left\{ z\in\Omega \,\middle|\, 2^l s < |v_h(z,t)| \le 2^{l+1}s \right\}$$

with a family of balls $B\in\mathcal{B}_l$, each ball having a center $x\in\partial\Omega^I(t)$ and a radius $\frac{hs}{2}$.

By construction it holds that

$$\begin{aligned}
\int_{(\Omega^I(t)\setminus\Omega^I(t-h))\cap B_{\frac{hs}{2}}(x)} |v_h| \quad &\ge \int_{(\Omega^I(t)\setminus\Omega^I(t-h))\cap B_{\frac{hs}{2}}(x)} \left(2^l s - \frac{1}{2}sh\right) \\
&> \int_{(\Omega^I(t)\setminus\Omega^I(t-h))\cap B_{\frac{hs}{2}}(x)} C2^l s \\
&= \int_{\Omega^I(t)\cap B_{\frac{hs}{2}}(x)} C2^l s = C2^l s \int_{B_{\frac{hs}{2}}(x)} \chi_h(t).
\end{aligned}$$

With Corollary 6.5 this implies the estimate

$$\int_{(\Omega^I(t)\setminus\Omega^I(t-h))\cap B_{\frac{hs}{2}}(x)} |v_h| > C2^l s^4 h^3.$$

With $|v_h| < 2^{l+1}s$ and Corollary 6.6 we find the upper estimate

$$\int_{B_{\frac{hs}{2}}(x)} |v_h||\nabla\chi_h| < \int_{B_{\frac{hs}{2}}(x)} 2^{l+1}s|\nabla\chi_h| \le C2^{l+1}s^3h^2.$$

Consequently,

$$\int_{B_{\frac{hs}{2}}(x)} |v_h||\nabla\chi_h| < Ch^{-1}s^{-1}\int_{(\Omega^I(t)\setminus\Omega^I(t-h))\cap B_{\frac{hs}{2}}(x)} |v_h|$$

for every ball $B \in \mathcal{B}_l$. When taking the union of all balls in $\mathcal{B}_l$, we arrive at

$$\int_{\{2^l s<|v_h|\le 2^{l+1}s\}} |v_h||\nabla\chi_h| < Cs^{-1}h^{-1}\int_{\{(2^l-\frac{1}{2})s<|v_h|<(2^{l+1}+\frac{1}{2})s\}\cap(\Omega^I(t)\triangle\Omega^I(t-h))} |v_h|.$$

Summation over all $l \in \mathbb{N}$ yields

$$\int_{\{|v_h|>s\}\cap\Omega^I(t)} |v_h||\nabla\chi_h| < Ch^{-1}s^{-1}\int_{\Omega^I(t)\triangle\Omega^I(t-h)} |v_h|,$$

and after integration in time,

$$\int_{\{|v_h|>s\}\cap\Omega_T} |v_h||\nabla\chi_h| < Ch^{-1}s^{-1}\int_0^T\int_{\Omega^I(t)\triangle\Omega^I(t-h)} |v_h| \le Cs^{-1}.$$

The last inequality follows from the a priori estimate (5.1) in Lemma 5.1. ∎

LEMMA 7.6 (estimate on the discrete velocity II). *There exists a constant C such that*

$$\int_{\Omega_T} v_h^2|\nabla\chi_h| < C.$$

Proof. We need only prove the assertion if $|v_h| > s$ for some $s > 0$ chosen large. As in Lemma 7.5, let $t = kh$ for $k = 1,\dots,N$ and $x \in \partial\Omega^I(t) \cap \{2^l s < |v_h| \le 2^{l+1}s\}$ for fixed $l \in \mathbb{N}$. Again we restrict the proof to the case $x \in \overline{\Omega^I(t)} \setminus \Omega^I(t-h)$.

We cover the set $\partial\Omega^I(t) \cap \{2^l s < |v_h| \le 2^{l+1}s\}$ by a family of balls $B \in \mathcal{B}_l$ each with radius $2^{l-1}sh$ and center points $x \in \partial\Omega^I(t)$. As in Lemma 7.5, we find

$$\int_{(\Omega^I(t)\setminus\Omega^I(t-h))\cap B_{2^{l-1}sh}(x)} |v_h| > \int_{(\Omega^I(t)\setminus\Omega^I(t-h))\cap B_{2^{l-1}sh}(x)} C2^l s \ge C2^{4l-3}h^3s^4.$$

On the other hand,

$$\int_{B_{2^{l-1}sh}(x)} v_h^2|\nabla\chi_h| \le C(2^{l+1}s)^2(2^{l-1}sh)^2 = C2^{4l-6}h^2s^4.$$

A comparison yields

$$\int_{B_{2^{l-1}sh}(x)} v_h^2|\nabla\chi_h| < Ch^{-1}\int_{(\Omega^I(t)\setminus\Omega^I(t-h))\cap B_{2^{l-1}sh}(x)} |v_h|.$$

Now we can proceed as in Lemma 7.5. ∎

LEMMA 7.7 (error in the discrete velocity). *For all test functions $\xi \in C_0^0(\Omega_T\,;\mathbb{R})$, it holds that*

$$\lim_{h\searrow 0}\int_{\Omega_T}\left(\frac{1}{h}\mathrm{dist}(\cdot,\partial\Omega^I(t-h))|\nabla\chi_h| - \partial_t^{-h}\chi_h\right)\xi = 0.$$

Proof. The principle of the proof is taken from [19]; see also [27] and [26].

We subdivide the proof into two parts. The first part studies the region where the discrete velocity is high. We will show the lemma for $\Omega\subset\mathbb{R}^3$; however, the following proof fails for space dimensions $n\geq 4$.

We cover the set $\Omega^I(t)\triangle\Omega^I(t-h)$ by a family $\mathcal{B}(t)$ of balls, each with radius $h^{\frac12}$ and with center $x\in\Omega^I(t)\triangle\Omega^I(t-h)$. Let

$$\mathcal{B} := \mathcal{B}(h)\cup\mathcal{B}(2h)\cup\cdots\cup\mathcal{B}(Nh).$$

(i) By $\mathcal{B}_1\subset\mathcal{B}$ we denote that subfamily of $\mathcal{B}$ with the property that for every ball B in $\mathcal{B}_1$ there exists a $z\in B\cap(\Omega^I(t)\triangle\Omega^I(t-h))$ with

$$\mathrm{dist}(z,\Omega^I(t-h)) > h^{\frac{9}{16}}. \tag{7.4}$$

Again it is enough to consider the case $x\in\overline{\Omega^I(t)}\setminus\Omega^I(t-h)$. We fix a ball B in $\mathcal{B}_1$. With Corollary 6.5 and (7.4) we obtain

$$\int_{B_{\frac12 h^{9/16}}(x)\cap(\Omega^I(t)\setminus\Omega^I(t-h))}|v_h| > C\frac{h^{-\frac{7}{16}}}{2}\left(\frac{h^{\frac{9}{16}}}{2}\right)^3 = Ch^{\frac54}.$$

This implies

$$\int_{B_{\sqrt{h}}(x)}|\partial_t^{-h}\chi_h| \leq C\int_{B_{\sqrt{h}}(x)}\frac{1}{h}\leq Ch^{\frac12}\leq Ch^{-\frac34}\int_{B_{\frac12 h^{9/16}}(z)\cap(\Omega^I(t)\setminus\Omega^I(t-h))}|v_h|$$
$$\leq ch^{-\frac34}\int_{B_{2\sqrt{h}}(x)\cap(\Omega^I(t)\setminus\Omega^I(t-h))}|v_h|. \tag{7.5}$$

The last estimate, (7.5), holds because

$$B_{\frac12 h^{9/16}}(z)\subset B_{2\sqrt{h}}(x).$$

A further bound can be obtained from Lemma 6.7 and Corollary 6.6:

$$\int_{B_{\sqrt{h}}(x)}|v_h||\nabla\chi_h|\leq Ch^{-\frac12}(h^{\frac12})^2\leq Ch^{\frac12}$$
$$\leq Ch^{-\frac34}\int_{B_{2\sqrt{h}}(x)\cap(\Omega^I(t)\setminus\Omega^I(t-h))}|v_h|. \tag{7.6}$$

When combining estimates (7.5) and (7.6) we obtain, after taking the union of all balls $B\in\mathcal{B}_1$ and after integration in time,

$$\int_0^T\int_{\mathcal{B}_1}\left(|\partial_t^{-h}\chi_h| + |v_h||\nabla\chi_h|\right)\leq Ch^{-\frac34}\int_0^T\int_{\Omega^I(t)\setminus\Omega^I(t-h)}|v_h|\leq Ch^{\frac14}. \tag{7.7}$$

The last inequality in (7.7) is a consequence of the a priori estimate on v_h found in Lemma 5.1. With (7.7) the lemma is shown for the regions with fast discrete velocity v_h and $x \in \overline{\Omega^I(t)} \setminus \Omega^I(t-h)$.

(ii) Now we discuss the regions in Ω with small discrete velocity. Let $\mathcal{B}_2 := \mathcal{B} \setminus \mathcal{B}_1$. By construction,

$$\operatorname{dist}(z, \partial\Omega^I(t-h)) \leq h^{\frac{9}{16}} \tag{7.8}$$

for any ball $B \in \mathcal{B}_2$ and any $z \in B$. Let $\beta := \frac{17}{32}$.

For $t = kh$, $1 \leq k \leq N$, we consider the subcover $\tilde{\mathcal{B}}_2(t)$ of $\mathcal{B}_2(t)$ with balls of radius $h^\beta = h^{\frac{17}{32}}$ and center $x \in \partial\Omega^I(t)$. Also, let $\tilde{\mathcal{B}}_2 := \cup_{k=1}^N \tilde{\mathcal{B}}_2(kh)$.

The following strong assumption is sufficient to prove part (ii).

For every ball $B \in \tilde{\mathcal{B}}_2$ there exists a $\nu \in S^2$ such that

$$\max\left\{ \|\nu_1 - \nu\|_{L^\infty(B)}, \|\nu_2 - \nu\|_{L^\infty(B)} \right\} \leq \omega(h), \tag{7.9}$$

where ν_1, ν_2 denote the unit outer normals to $\Omega^I(t)$ and $\Omega^I(t-h)$, respectively. The function $\omega(h)$ converges uniformly to 0 in every ball $B \in \tilde{\mathcal{B}}_2$.

It is evident that (7.9) implies the assertion of the lemma. Indeed, (7.9) is a very strong condition as it controls the variation of the normals, and thus we obtain directly

$$\left| \int_B \partial_t^{-h} \chi_h - v_h |\nabla \chi_h| \right| \leq \omega(h) \int_B |\partial_t^{-h} \chi_h|$$

for all balls $B \in \tilde{\mathcal{B}}_2$. Taking the union over all balls in $\tilde{\mathcal{B}}_2(t)$ and using discrete integration in time yields, for test functions $\xi \in C^0(\Omega_T; \mathbb{R})$,

$$\left| \int_0^T \int_{\mathcal{B}_2(t)} \Big(\partial_t^{-h} \chi_h - v_h |\nabla \chi_h| \Big) \xi \right| \leq \|\xi\|_{L^\infty(\Omega_T)} \, \omega(h) \int_{\Omega_T} |\partial_t^{-h} \chi_h| \leq C\omega(h).$$

The last estimate follows from the bound on $\|\partial_t^{-h} \chi_h\|_{L^1(\Omega_T)}$ provided in Lemma 7.2.

So it remains to prove (7.9). We will use Bernstein's theorem (see [14], [13], and [12]) for $n = 8$.

Assume that (7.9) does not hold. Then there exists a subsequence of balls $B_{h^\beta}(x_h)$ such that for all ν the norms $\|\nu_{\Omega^I(t)} - \nu\|_{L^\infty(B)}$, $\|\nu_{\Omega^I(t-h)} - \nu\|_{L^\infty(B)}$ do not converge to 0. We blow up the balls $B_{\sqrt{h}}(x_h)$ with a factor $h^{-\beta}$ and shift them such that we obtain a sequence of balls with radius $h^{\frac{1}{2}-\beta}$ and center at 0. These scaled and translated balls will be denoted by $\Omega_h^\beta(t)$. The characteristic functions to $\Omega_h^\beta(t)$ are minimizers of the scaled functional F_h given by (4.3). The compactness of $BV(\Omega)$ with respect to L^1-convergence implies that we can find a subsequence of $\chi_{\Omega_h^\beta(t)}$ and a subsequence of $\chi_{\Omega_h^\beta(t-h)}$ (both denoted as the original sequence) such that

$$\chi_h^\beta(t) := \chi_{\Omega_h^\beta(t)} \to \chi_{\mathcal{T}_1}, \quad \chi_{\Omega_h^\beta(t-h)} \to \chi_{\mathcal{T}_2} \qquad \text{in } L^1_{\mathrm{loc}}(\mathbb{R}^3)$$

for two sets $\mathcal{T}_1$, $\mathcal{T}_2$. Since the velocity term in the rescaled functional is scaled by a factor $h^{2\beta}$, due to Lemma 6.7 we find

$$h^{2\beta} \int_{\Omega_h^\beta(t) \triangle \Omega_h^\beta(t-h)} \frac{1}{h} \operatorname{dist}(\cdot, \partial\Omega^\beta(t-h)) \leq C h^{2\beta} h^{-\frac{1}{2}} h^{-\beta} = C h^{\beta - \frac{1}{2}}.$$

This shows that the velocity term in the rescaled functional converges to 0 for $h \searrow 0$. Due to (5.9), this also holds for the other terms in the scaled functional, except for the area term $\int_{\Omega_h^\beta} |\nabla \chi_h^\beta|$. This implies that the sets $\mathcal{T}_1$, $\mathcal{T}_2$ are area-minimizing in $\mathbb{R}^3$. Bernstein's theorem thus yields that $\mathcal{T}_i$ are half-spaces.

From the bound on the discrete velocity (7.8) we learn that

$$\operatorname{dist}(\cdot, \partial\Omega_h^\beta(t-h)) \le h^{\frac{9}{16}} h^{-\beta} \to 0 \qquad \text{on } B_{h^{\frac{1}{2}-\beta}}(x_h) \cap (\Omega_h^\beta(t) \triangle \Omega_h^\beta(t-h)).$$

This gives $\mathcal{T}_1 = \mathcal{T}_2 =: \mathcal{T}$ and $0 \in \partial\mathcal{T}$. Without loss of generality, we may assume

$$\mathcal{T} = \{z \in \mathbb{R}^3 \mid z_3 < 0\},$$

implying $\nu = (0,0,1)$ for the normal to the minimizing set $\mathcal{T}$.

Let $\varrho > 0$ be fixed. Our construction implies

$$\lim_{h \searrow 0} \varrho^{-3} \int_{B_\varrho(z)} |\nu_1 - \nu|^2 |\nabla \chi_h^\beta| = 0 \qquad \text{uniformly for } z \in B_1(0) \cap \partial\Omega_h^\beta(t).$$

Now we apply the excess-decay-lemma; see, for instance, [29]. It states that

$$\lim_{h \searrow 0} |\nu_1(z) - \nu| = 0 \qquad \text{uniformly for } z \in B_1(0) \cap \partial\Omega_h^\beta(t).$$

The same statement holds for ν_2 and $\Omega_h^\beta(t-h)$. Thus, after rescaling with factor h^β, this gives a contradiction, and (7.9) is proved. ∎

8. Convergence of the discrete solution. We still have to verify that the discrete solutions converge to the solution of the continuous equations. We will prove this in a series of lemmas.

From Theorem 7.4 it does not follow that

$$|\nabla \chi_h| \rightharpoonup |\nabla \chi| \qquad \text{in } \operatorname{rca}(\Omega_T).$$

For the proofs of this section we therefore make the assumption

$$\int_{\Omega_T} |\nabla \chi_h| \to \int_{\Omega_T} |\nabla \chi| \qquad \text{as } h \searrow 0, \tag{8.1}$$

where χ is the function of Theorem 7.4.

First, we show the existence of a velocity v that satisfies (W3).

LEMMA 8.1 (existence of a limit velocity). *Let χ be the characteristic function specified in Theorem* 7.4. *Then there exists a velocity function v which satisfies*

$$v \in L^1((0,T); L^1(\Omega; |\nabla \chi(t)|))$$

such that (W3) *is fulfilled.*

Proof. Discrete integration by parts yields

$$\int_{\Omega_T} \xi \partial_t^{-h} \chi_h + \frac{1}{h} \int_\Omega \int_0^h \chi_{\Omega_0^I} \xi = - \int_{\Omega_T} \chi_h \partial_t^h \xi \tag{8.2}$$

for all $\xi \in C^\infty(\overline{\Omega} \times [0,T]; \mathbb{R})$ with $\xi(T) = 0$.

As was shown in Lemma 7.2, there exists a constant C independent of h such that

$$\int_{\Omega_T} |\partial_t^{-h}\chi_h| < C.$$

Therefore there exists a subsequence of $\partial_t^{-h}\chi_h$ with

$$\partial_t^{-h}\chi_h \rightharpoonup \nu \qquad \text{in } \mathrm{rca}(\Omega_T).$$

With Lemma 7.7 we find at once that

$$v_h|\nabla\chi_h| \rightharpoonup \nu \qquad \text{in } \mathrm{rca}(\Omega_T). \tag{8.3}$$

We prove now that ν is absolutely continuous with respect to the measure $|\nabla\chi|$. Let $E \subset \Omega_T$ with $|\nabla\chi(E)| = 0$. From our assumption (8.1) we infer that there exists a function $g : \mathbb{R} \to \mathbb{R}$ with $g(h) \searrow 0$ as $h \searrow 0$ such that

$$|\nabla\chi_h(E)| \leq g(h).$$

Due to Lemma 7.5 it follows that

$$\begin{aligned}\int_E |v_h||\nabla\chi_h| &= \int_{\{|v_h|\leq s\}\cap E} |v_h||\nabla\chi_h| + \int_{\{|v_h|>s\}\cap E} |v_h||\nabla\chi_h| \\ &\leq sg(h) + Cs^{-1}.\end{aligned}$$

The right-hand side of this estimate can be made arbitrarily small for small h, and thus

$$\lim_{h\searrow 0}\int_E |v_h||\nabla\chi_h| = 0.$$

This shows that $\nu(E) = 0$ and proves the absolute continuity of ν. The existence of $v \in L^1(0,T);\ L^1(\Omega; |\nabla\chi(t)|)$ satisfying

$$\nu = v|\nabla\chi| \tag{8.4}$$

follows now from the Radon–Nikodym theorem.

The measures $v_h|\nabla\chi_h|$ and $v|\nabla\chi|$ are absolutely continuous in time, and furthermore we can approximate ξ in (8.2) by functions with compact support. Passing to the limit $h \searrow 0$ gives (W3). $\square$

LEMMA 8.2 (convergence of the advection term in the weak curvature equation). *Let χ be the characteristic function specified in Theorem* 7.4 *and v be the function specified in Lemma* 8.1. *Then it holds that*

$$\lim_{h\searrow 0}\int_{\Omega_T} v_h\zeta\nabla\chi_h = \int_{\Omega_T} v\zeta\nabla\chi$$

for all $\zeta \in C^1([0,T]\times\overline{\Omega};\mathbb{R}^3)$ with $\zeta = 0$ on $\partial\Omega\times(0,T)$.

Proof. As $|\nabla\chi|$ is a Radon measure, there exists for any $\varepsilon > 0$ a vector-valued mapping $g_\varepsilon \in C^0(\Omega_T;\mathbb{R}^3)$, $|g_\varepsilon| \leq 1$, such that

$$\int_{\Omega_T} |\nabla\chi| - \int_{\Omega_T} g_\varepsilon\nabla\chi < \varepsilon.$$

Consequently,

$$\lim_{h\searrow 0}\left(\int_{\Omega_T}|\nabla\chi_h| - \int_{\Omega_T} g_\varepsilon\nabla\chi_h\right) = \int_{\Omega_T}|\nabla\chi| - \int_{\Omega_T} g_\varepsilon\nabla\chi < \varepsilon. \tag{8.5}$$

With the notation

$$\nu_h := \frac{\nabla\chi_h}{|\nabla\chi_h|}, \qquad \nu := \frac{\nabla\chi}{|\nabla\chi|},$$

we find

$$\begin{aligned}\int_{\Omega_T}(\nu_h - g_\varepsilon)^2|\nabla\chi_h| &= \int_{\Omega_T}(1-2g_\varepsilon\nu_h + g_\varepsilon^2)|\nabla\chi_h| = \int_{\Omega_T}(2-2g_\varepsilon\nu_h+g_\varepsilon^2-1)|\nabla\chi_h|\\ &\le 2\int_{\Omega_T}(1-g_\varepsilon\nu_h)|\nabla\chi_h|.\end{aligned}$$

This estimate in combination with (8.5) yields

$$\lim_{h\searrow 0}\int_{\Omega_T}(\nu_h-g_\varepsilon)^2|\nabla\chi_h| < 2\varepsilon. \tag{8.6}$$

This result, together with Hölder's inequality and the boundedness of $(\int_{\Omega_T}|\nabla\chi_h|)^{\frac{1}{2}}$, implies

$$\lim_{h\searrow 0}\int_{\Omega_T}|\nu_h-g_\varepsilon||\nabla\chi_h| \le C\lim_{h\to 0}\left(\int_{\Omega_T}(\nu_h-g_\varepsilon)^2|\nabla\chi_h|\right)^{\frac{1}{2}} \le C(2\varepsilon)^{\frac{1}{2}}.$$

In the same way, we compute with Lemma 7.6 and Hölder's inequality

$$\begin{aligned}\lim_{h\searrow 0}\int_{\Omega_T}|v_h||g_\varepsilon-\nu_h||\nabla\chi_h| &\le \left(\int_{\Omega_T}|v_h|^2|\nabla\chi_h|\right)^{\frac{1}{2}}\left(\int_{\Omega_T}(\nu_h-g_\varepsilon)^2|\nabla\chi_h|\right)^{\frac{1}{2}}\\ &\le C(2\varepsilon)^{\frac{1}{2}}. \end{aligned}\tag{8.7}$$

Now we are prepared to show the assertion of the lemma. We see that

$$\begin{aligned}\left|\int_{\Omega_T} v_h\zeta\nabla\chi_h - v\zeta\nabla\chi\right| &= \left|\int_{\Omega_T} v_h\zeta\nu_h|\nabla\chi_h| - v\zeta\nu|\nabla\chi|\right|\\ &\le \left|\int_{\Omega_T} v_h\zeta\nu_h|\nabla\chi_h| - v\zeta g_\varepsilon|\nabla\chi|\right| + \left|\int_{\Omega_T} v\zeta g_\varepsilon|\nabla\chi| - v\zeta\nu|\nabla\chi|\right|\\ &=: I_1(\varepsilon,h) + I_2(\varepsilon).\end{aligned}$$

We estimate I_1 and I_2 independently. For the first integral, we have

$$I_1(\varepsilon,h) \le \int_{\Omega_T}|v_h||\zeta||\nu_h-g_\varepsilon||\nabla\chi_h| + \left|\int_{\Omega_T} v_h\zeta g_\varepsilon|\nabla\chi_h| - v\zeta g_\varepsilon|\nabla\chi|\right|. \tag{8.8}$$

Estimate (8.7) confirms the convergence of the first integral in (8.8), the properties (8.3) and (8.4) lead to the convergence of the second integral, and thus

$$\lim_{\varepsilon\searrow 0}\lim_{h\searrow 0} I_1(\varepsilon,h) \le \lim_{\varepsilon\searrow 0} C\varepsilon^{\frac{1}{2}} = 0.$$

For the integral $I_2(\varepsilon)$ we get directly

$$I_2(\varepsilon) \le \omega(\varepsilon)$$

for a function $\omega(\varepsilon)$ that tends to 0 as ε tends to 0. In conclusion we have found that

$$\lim_{h\searrow 0}\left|\int_{\Omega_T} v_h\zeta\nabla\chi_h - v\zeta\nabla\chi\right| = 0,$$

and the proof is finished. $\square$

LEMMA 8.3 (convergence of the weak curvature term). *Let χ be the characteristic function specified in Theorem* 7.4. *Then it holds that*

$$\lim_{h\searrow 0}\int_{\Omega_T} \nu_h\nabla\zeta\nu_h|\nabla\chi_h| = \int_{\Omega_T}\nu\nabla\zeta\nu|\nabla\chi|$$

for all $\zeta \in C^1([0,T]\times\overline{\Omega};\mathbb{R}^3)$ with $\zeta = 0$ on $\partial\Omega\times(0,T)$.

Proof. For $\varepsilon > 0$ and the family g_ε of mappings introduced in Lemma 8.2, one has

$$\begin{aligned}\left|\int_{\Omega_T}\nu_h\nabla\zeta\nu_h|\nabla\chi_h| - \int_{\Omega_T}\nu\nabla\zeta\nu|\nabla\chi|\right| \le{}& \left|\int_{\Omega_T}\nu_h\nabla\zeta\nu_h|\nabla\chi_h| - \int_{\Omega_T} g_\varepsilon\nabla\zeta g_\varepsilon|\nabla\chi_h|\right|\\ &+\left|\int_{\Omega_T} g_\varepsilon\nabla\zeta g_\varepsilon|\nabla\chi_h| - \int_{\Omega_T} g_\varepsilon\nabla\zeta g_\varepsilon|\nabla\chi|\right|\\ &+\left|\int_{\Omega_T} g_\varepsilon\nabla\zeta g_\varepsilon|\nabla\chi| - \int_{\Omega_T}\nu\nabla\zeta\nu|\nabla\chi|\right|.\end{aligned}$$

With the help of (8.6) and (8.1) we see that the right-hand side of this estimate converges to 0 as $h\searrow 0$ and $\varepsilon\searrow 0$. $\square$

It is well known that condition (8.1), together with the convergence of $\chi_h \to \chi$ in $L^1(\Omega_T)$ as stated in Theorem 7.4, implies the convergence of the mean curvature term. This is a consequence of the lemma of Reshetnyak, which we state for the reader's convenience.

LEMMA 8.4 (lemma of Reshetnyak). *Let μ_h be a sequence of vector-valued measures in $\Omega\subset\mathbb{R}^n$ with*

$$\mu_h \to \mu \qquad \text{in } \mathrm{rca}(\Omega;\mathbb{R}^n)$$

and

$$|\mu_h|(\Omega) \to |\mu|(\Omega).$$

Then for all bounded continuous functions in $S^{n-1}\times\Omega$ it holds that

$$f\left(\frac{\mu_h}{|\mu_h|},\cdot\right)|\mu_h| \rightharpoonup f\left(\frac{\mu}{|\mu|},\cdot\right)|\mu| \qquad \text{in } \mathrm{rca}(\Omega;\mathbb{R}^n),$$

where $\frac{\mu}{|\mu|}$ *is defined by the Radon–Nikodym theorem.*

Proof. The proof is similar to the proof of Lemma 8.3; see [23]. ◻

Now we discuss the convergence of the discrete chemical potentials.

LEMMA 8.5 (convergence of μ_h^D). *There exists a function* $\mu^D \in L^2(0,T;\, H^{1,2}(\Omega))$ *and a subsequence* μ_h^D *such that for* $h \searrow 0$,

$$\mu_h^D \rightharpoonup \mu^D \qquad \textit{in } L^2(0,T;\, H^{1,2}(\Omega))$$

and

$$(1-\chi)\mu_D = (1-\chi)\varphi \qquad \textit{for almost every } (x,t) \in \Omega_T.$$

We recall that φ are the boundary conditions on μ^D given by (2.25).

Proof. Due to the a priori estimates, we have for a constant C independent of h that

$$\int_{\Omega_T} |\nabla \mu_h^D|^2 \le C.$$

Together with the boundary conditions for μ_h^D and the Poincaré inequality, this yields

$$\|\mu_h^D\|_{L^2(0,T;\, H^{1,2}(\Omega))} \le C.$$

The second assertion of the lemma follows directly from the a priori estimate, which ensures

$$\int_{\Omega_T} (1-\chi_h(t-h))|\nabla \mu_h^D|^2 \to 0 \qquad \text{for } h \searrow 0.$$

This proves the lemma. ◻

LEMMA 8.6 (convergence of μ_h^N). *There exists a function* $\mu^N \in L^2(0,T; H^{1,\frac{3}{2}}(\Omega))$ *and a subsequence* μ_h^N *such that for* $h \searrow 0$,

$$\mu_h^N \rightharpoonup \mu^N \qquad \textit{in } L^2(0,T;\, L^{\frac{3}{2}}(\Omega)),$$

with

$$(1-\chi)\mu^N = 0 \qquad \textit{for almost every } (x,t) \in \Omega_T.$$

Proof. For any $h > 0$, let $\alpha = \alpha(h) > 0$ be a small real number with $\alpha(h) \searrow 0$ as $h \searrow 0$. We define the mapping $\mu_h^{N,\alpha}$ by

$$\mu_h^{N,\alpha} := \begin{cases} (|\mu_h^N| - \alpha)_+ \frac{\mu_h^N}{|\mu_h^N|} & \text{if } \mu_h^N \neq 0, \\ 0 & \text{else.} \end{cases}$$

For this truncated chemical potential, we compute with the chain rule

$$\begin{aligned}
&\int_{\Omega_T}(1-\chi_h(t-h))|\nabla\mu_h^{N,\alpha}|^{\frac{3}{2}}\\
&\quad=\int_{\Omega_T\cap\{|\mu_h^{N,\alpha}|>\alpha\}}(1-\chi_h(t-h))|\nabla\mu_h^{N,\alpha}|^{\frac{3}{2}}\\
&\quad\le\int_{\Omega_T\cap\{|\mu_h^{N,\alpha}|>\alpha\}}(1-\chi_h(t-h))|\nabla\big((\mu_h^{N,\alpha})^{\frac{5}{4}}\big)|^{\frac{3}{2}}\alpha^{-\frac{3}{8}}\\
&\quad=\frac{5}{4}\alpha^{-\frac{3}{8}}\int_{\Omega_T\cap\{|\mu_h^{N,\alpha}|>\alpha\}}(1-\chi_h(t-h))|\mu_h^{N,\alpha}|^{\frac{3}{8}}|\nabla\mu_h^{N,\alpha}|^{\frac{3}{2}}\\
&\quad\le\frac{5}{4}\alpha^{-\frac{3}{8}}\left(\int_{\Omega_T}(1-\chi_h(t-h))|\nabla\mu_h^{N,\alpha}|^2\right)^{\frac{3}{4}}\left(\int_{\Omega_T}(1-\chi_h(t-h))|\mu_h^{N,\alpha}|^{\frac{3}{2}}\right)^{\frac{1}{4}},
\end{aligned}$$

where Hölder's inequality was used to get the last line.

We apply the a priori estimates (5.3) and (5.5) and obtain

$$\int_{\Omega_T}(1-\chi_h(t-h))|\nabla\mu_h^{N,\alpha}|^{\frac{3}{2}}\le C\alpha^{-\frac{3}{8}}\varepsilon^{-\frac{3}{4}}\gamma^{\frac{3}{16}}\le C.$$

Here we choose $\alpha(h)$ such that $\alpha^{-\frac{3}{8}}\varepsilon^{-\frac{3}{4}}\gamma^{\frac{3}{16}}$ is bounded uniformly in h. So we have found that

$$\int_{\Omega_T}|\nabla\mu_h^{N,\alpha}|^{\frac{3}{2}}\le C.$$

With the boundary conditions for μ_h^N, this ensures the existence of a subsequence, again denoted by $\mu_h^{N,\alpha}$, and the existence of a function $\mu^{N,\alpha}\in L^2(0,T;\,H^{1,\frac{3}{2}}(\Omega))$ with

$$\mu_h^{N,\alpha}\rightharpoonup\mu^{N,\alpha}\qquad\text{in } L^2(0,T;\,H^{1,\frac{3}{2}}(\Omega)).$$

Also we can pick a subsequence $\alpha(h)$ with

$$\mu^{N,\alpha}\rightharpoonup\mu^{N}\qquad\text{in } L^2(0,T;\,H^{1,\frac{3}{2}}(\Omega))$$

for a suitable $\mu^N\in L^2(0,T;\,H^{1,\frac{3}{2}}(\Omega))$.

Finally, by definition of $\mu_h^{N,\alpha}$, we have for every $\alpha\ge 0$ and every $h>0$,

$$\|\mu_h^N-\mu_h^{N,\alpha}\|_{L^2(0,T;\,L^{\frac{3}{2}}(\Omega))}\le C\alpha.$$

By construction it thus holds for $h\searrow 0$ that

$$\mu_h^{N}\rightharpoonup\mu^{N}\qquad\text{in } L^2(0,T;\,L^{\frac{3}{2}}(\Omega)).$$

The second claim of the lemma is again a consequence of the a priori estimate (5.5), which shows as $\gamma(h)\searrow 0$ for $h\searrow 0$,

$$\int_{\Omega_T}(1-\chi_h(t-h))|\mu_h^N|^2\to 0\qquad\text{as } h\searrow 0.$$

This ends the proof. $\square$

Lemma 8.7. *Let χ be the characteristic function specified in Theorem* 7.4. *Then it holds that*

$$\lim_{h\searrow 0}\int_{\Omega_T} f_{II}^*(\mu_h)\zeta\nabla\chi_h = \int_{\Omega_T} K\zeta\nabla\chi$$

for all $\zeta \in C^1([0,T]\times\overline{\Omega};\mathbb{R}^3)$ with $\zeta = 0$ on $\partial\Omega\times(0,T)$, where $K := f_{II}^{D,*}(\varphi)$.

Proof. We reformulate the left-hand side. For a test function $\zeta \in C^1([0,T]\times\overline{\Omega};\mathbb{R}^3)$ with $\zeta = 0$ on $\partial\Omega\times(0,T)$, we compute

$$\begin{aligned}\int_{\Omega_T} f_{II}^*(\mu_h)\zeta\nabla\chi_h &= \int_{\Omega_T}(1-\chi_h)f_{II}^{D,*}(\mu_h^D)\operatorname{div}(\zeta) + \int_{\Omega_T}(1-\chi_h)m_h^D\nabla\mu_h^D\zeta\\ &\quad+\int_{\Omega_T}(1-\chi_h)f_{II}^{N,*}(\mu_h^N)\operatorname{div}(\zeta) + \int_{\Omega_T}(1-\chi_h)m_h^N\nabla\mu_h^N\zeta\\ &=: I_1 + I_2 + I_3 + I_4.\end{aligned}$$

Using the convexity of $f_{II}^{D,*}$ and Lemma 8.5, we have for I_1,

$$\lim_{h\searrow 0}\int_{\Omega_T}(1-\chi_h)f_{II}^{D,*}(\mu_h^D)\operatorname{div}(\zeta) = \int_{\Omega_T}(1-\chi)f_{II}^{D,*}(\varphi)\operatorname{div}(\zeta).$$

So it remains to show that $I_k \to 0$ as $h\searrow 0$ for $k = 2,3,4$. For I_2 this results from

$$|I_2| \le C\int_{\Omega_T}|\chi_h(t-h)-\chi_h||\nabla\mu_h^D||\zeta| + \int_{\Omega_T}(1-\chi_h(t-h))|\nabla\mu_h^D||\zeta|,$$

and the a priori estimates and Lemma 7.2 yield, as desired, $|I_2|\searrow 0$ as $h\searrow 0$.

From $\mu = \frac{\partial f_{II}}{\partial m}$ in $\Omega\setminus\overline{\Omega^I}$ we get

$$m_i^N = \exp(\mu_i^N)\delta \qquad \text{for } i = 1,2$$

and consequently,

$$f_{II}^{N,*}(\mu_h^N) = k_B\theta\delta\sum_{i=1}^{2}\exp\left(\frac{\mu_{h,i}^N}{k_B\theta}\right).$$

Lemma 8.6 therefore implies $|I_3|\searrow 0$ as $\delta(h)\searrow 0$.

Finally,

$$|I_4| = C\left|\int_{\Omega_T}\delta(1-\chi_h)\exp(\mu^N)\nabla\mu_h^N\right| \le C\delta\varepsilon^{-1},$$

where we use again the a priori estimates and Hölder's inequality. With assumption (A2) we obtain $|I_2|\searrow 0$ as $h\searrow 0$, and the lemma is proved. $\square$

Lemma 8.8. *Let χ be the characteristic function specified in Theorem* 7.4. *Then it holds that*

$$\lim_{h\searrow 0}\int_{\Omega_T} f_I^*(\mu_h)\zeta\nabla\chi_h = \int_{\Omega_T} f_I^*(\mu)\zeta\nabla\chi$$

for all $\zeta \in C^1([0,T]\times\overline{\Omega};\mathbb{R}^3)$ *with* $\zeta = 0$ *on* $\partial\Omega\times(0,T)$*, where* $\mu = (\mu^N,\mu^D)$.

Proof. We reformulate the left-hand side. For a test function $\zeta \in C^1([0,T]\times\overline{\Omega};\mathbb{R}^3)$ with $\zeta = 0$ on $\partial\Omega\times(0,T)$, we compute

$$\int_{\Omega_T} f_I^*(\mu_h)\zeta\nabla\chi_h = -\int_{\Omega_T}\chi_h f_I^*(\mu_h)\mathrm{div}(\zeta) - \int_{\Omega_T}\chi_h m_h\nabla\mu_h\zeta. \tag{8.9}$$

From the convexity of f_I^* together with Lemmas 8.5 and 8.6, it follows directly that

$$\lim_{h\searrow 0}\int_{\Omega_T}\chi_h f_I^*(\mu_h)\mathrm{div}(\zeta) = \int_{\Omega_T}\chi_h f_I^*(\mu^N,\mu^D)\mathrm{div}(\zeta).$$

The proof is finished if we can show the convergence of the second integral on the right-hand side in (8.9) for $h\searrow 0$, but this is quite involved.

We use the definition

$$m_h^\beta := \begin{cases} m_h & \text{if } \beta < |m_h| < 1-\beta,\\ 0 & \text{else}\end{cases}$$

to reformulate the second integral on the right in (8.9). Since $\frac{\partial f_I^*}{\partial\mu}(\mu)$ is locally a Lipschitz function, we have for arbitrary vectors μ_1, μ_2,

$$\left|\frac{\partial f_I^*}{\partial\mu}(\mu_1) - \frac{\partial f_I^*}{\partial\mu}(\mu_2)\right| \le C|\mu_1-\mu_2|,$$

where C depends on μ_1, μ_2 and $C\to\infty$ for μ_1, $\mu_2\to\infty$. So we can estimate

$$\begin{aligned}
&\int_{\Omega_T}\chi_h|\nabla m_h^\beta|^{\frac32}\\
&\quad\le \int_{\Omega_T}\chi_h(t-h)|\nabla m_h^\beta|^{\frac32} + \int_{\Omega_T}|\chi_h-\chi_h(t-h)|\,|\nabla m_h^\beta|^{\frac32}\\
&\quad\le C\left(\int_{\Omega_T}\chi_h(t-h)|\nabla m_h^\beta|^2\right)^{\frac34} + C\left(\int_{\Omega_T}|\chi_h-\chi_h(t-h)|\right)^{\frac14}\left(\int_{\Omega_T}|\nabla m_h^\beta|^2\right)^{\frac34}\\
&\quad\le C(\varepsilon^{-\frac34}h^{\frac14}+1).
\end{aligned}$$

Due to assumption (A3) the right-hand side converges to 0 as $h\searrow 0$.

For functions $\varphi,\psi\in C^\infty(B_1(0);\mathbb{R}^+)$ we introduce mollifiers φ_ϱ and ψ_σ by setting $\varphi_\varrho(x) := \varrho^{-3}\varphi(x/\varrho)$ and $\psi_\sigma(t) := \sigma^{-3}\psi(t/\sigma)$. It holds that $\operatorname{supp}\varphi_\varrho\subset B_\varrho(0)$, $\operatorname{supp}\psi_\sigma\subset B_\sigma(0)$. We consider sequences (ϱ,σ) with $\sigma^{\frac12}\varrho^{-1}\to 0$. Then we have the estimate

$$\begin{aligned}
&\int_{\Omega_T}|(\chi_h m_h^\beta)*\varphi_\varrho*\psi_\sigma - \chi_h m_h^\beta|\\
&\quad\le \int_{\Omega_T}|(\chi_h m_h^\beta)*\varphi_\varrho*\psi_\sigma - (\chi_h m_h^\beta)*\varphi_\varrho| + \int_{\Omega_T}|(\chi_h m_h^\beta)*\varphi_\varrho - \chi_h m_h^\beta|\\
&\quad\le C\sigma^{\frac12}\varrho^{-1}\|\partial_t^h(\chi_h m_h^\beta)\|_{L^2(0,T;\,H^{-1,2}(\Omega))} + C\varrho\left(\int_{\Omega_T}|\nabla(\chi_h m_h^\beta)|+1\right).
\end{aligned}$$

Taking (5.6) into account, we arrive at

$$\int_{\Omega_T} |(\chi_h m_h^\beta) * \varphi_\varrho * \psi_\sigma - \chi_h m_h^\beta| \le C(\sigma^{\frac{1}{2}}\varrho^{-1} + \varrho),$$

which holds uniformly in h. It follows that there exists a subsequence of $\chi_h m_h^\beta$ (denoted as the original sequence) and a function $\Gamma \in L^1(\Omega_T)$ such that

$$\chi_h m_h^\beta \to \Gamma \qquad \text{in } L^1(\Omega_T). \tag{8.10}$$

Additionally,

$$\|\chi_h m_h - \Gamma\|_{L^1(\Omega_T)} \le \|\chi_h m_h - \chi_h m_h^\beta\|_{L^1(\Omega_T)} + \|\chi_h m_h^\beta - \Gamma\|_{L^1(\Omega_T)},$$

and the right side of this estimate converges to 0 for $h \searrow 0$. With (8.10) this ensures the strong convergence of a subsequence $\chi_h m_h$ in $L^1(\Omega_T)$. Furthermore we know that there exists a subsequence of m_h such that for any $1 \le p < \infty$,

$$m_h \to m \qquad \text{in } L^p(\Omega_T) \text{ for } h \searrow 0.$$

This yields $\chi_h m_h \rightharpoonup \chi m$ in $L^p(\Omega_T)$ and finally,

$$\chi_h m_h \to \chi m \qquad \text{in } L^p(\Omega_T) \text{ for } h \searrow 0. \tag{8.11}$$

Next we will show that

$$\chi_h \nabla \mu_h \zeta \rightharpoonup \chi \nabla \mu \zeta \qquad \text{in } L^2(0,T;\, L^{\frac{6}{5}}(\Omega)).$$

We use again the mapping $\mu_h^{N,\alpha}$ from the proof of Lemma 8.6. We define the mapping $\mu_{h,\alpha}^N$ by

$$\mu_{h,\alpha}^N := \begin{cases} \min\{\alpha, |\mu_h^{N,\alpha}|\}\frac{\mu_h^N}{|\mu_h^N|} & \text{if } \mu_h^N \ne 0, \\ 0 & \text{else.} \end{cases}$$

From the definition of $\mu_{h,\alpha}^N$ we find

$$\begin{aligned} \left| \int_{\Omega_T} \chi_h(t-h)\nabla\mu_{h,\alpha}^N \zeta \right| &\le C \int_{\Omega_T} |\mu_{h,\alpha}^N||\nabla(\chi_h(t-h)\zeta)| \\ &\le C\alpha \int_{\Omega_T} |\nabla(\chi_h(t-h)\zeta)| \le C\alpha. \end{aligned} \tag{8.12}$$

Since $\alpha(h) \searrow 0$ as $h \searrow 0$, we have for $h \searrow 0$ the convergence

$$\int_{\Omega_T} \chi_h(t-h)\nabla\mu_h^{N,\alpha}\zeta \to \int_{\Omega_T} \chi\nabla\mu^N\zeta,$$

which with (8.12) leads to

$$\int_{\Omega_T} \chi_h(t-h)\nabla\mu_h^N\zeta \to \int_{\Omega_T} \chi\nabla\mu^N\zeta \qquad \text{for } h \searrow 0. \tag{8.13}$$

From the a priori estimates, we can deduce

$$\chi_h(t-h)\nabla\mu_h^N \rightharpoonup \Lambda \qquad \text{in } L^2(0,T;\, L^2(\Omega)),$$

and therefore

$$\int_{\Omega_T} \chi_h(t-h)\nabla\mu_h^N\zeta \to \int_{\Omega_T} \Lambda\zeta \qquad \text{for } h \searrow 0. \tag{8.14}$$

From (8.13) and (8.14) we obtain

$$\int_{\Omega_T} (\chi\nabla\mu^N - \Lambda)\zeta = 0 \qquad \text{for all } \zeta \in C^1([0,T]\times\overline{\Omega};\, \mathbb{R}^3) \text{ with } \zeta = 0 \text{ on } \partial\Omega\times(0,T).$$

Consequently,

$$\chi\nabla\mu^N = \Lambda \qquad \text{for almost every } (x,t) \in \Omega_T.$$

So we have shown that

$$\chi_h(t-h)\nabla\mu_h^N\zeta \rightharpoonup \chi\nabla\mu^N\zeta \qquad \text{in } L^2(0,T;\, L^2(\Omega)) \qquad \text{for } h \searrow 0.$$

Fix an arbitrary $g \in L^2(0,T;\, L^6(\Omega))$. We rewrite the integrand in the form

$$\int_{\Omega_T} \chi_h\nabla\mu_h^N\zeta g = \int_{\Omega_T} (\chi_h - \chi_h(t-h))\nabla\mu_h^N\zeta g + \int_{\Omega_T} \chi_h(t-h)\nabla\mu_h^N\zeta g.$$

With Lemma 7.3 and Hölder's inequality, we can find the estimate

$$\begin{aligned} &\left| \int_{\Omega_T} (\chi_h - \chi_h(t-h))\nabla\mu_h^N\zeta g \right| \\ &\quad \le C\left(\int_{\Omega} |\chi_h - \chi_h(t-h)|\right)^{\frac{2}{3}} \left(\int_{\Omega_T} |\mu_h^N|^2\right)^{\frac{1}{2}} \|g\|_{L^2(0,T;\, L^6(\Omega))} \\ &\quad \le Ch^{\frac{1}{3}}\varepsilon^{-1}\|g\|_{L^2(0,T;\, L^6(\Omega))}. \end{aligned}$$

With assumption (A4) this yields

$$\chi_h\nabla\mu_h^N\zeta \rightharpoonup \chi\nabla\mu^N\zeta \qquad \text{in } L^2(0,T;\, L^{\frac{6}{5}}(\Omega)) \qquad \text{for } h \searrow 0. \tag{8.15}$$

Statements (8.11) and (8.15) combined give

$$\int_{\Omega_T} \chi_h m_h^N \nabla\mu_h^N\zeta \to \int_{\Omega_T} \chi m^N\nabla\mu^N\zeta \qquad \text{for } h \searrow 0.$$

The proof of convergence for the Dirichlet data is analogous. Now we can pass to the limit $h \searrow 0$ in (8.9). □

The following theorem is now a direct consequence of the lemmas shown above.

THEOREM 8.9 (existence of weak solutions). *Let $\Omega \subset \mathbb{R}^3$ be an open bounded set with Lipschitz boundary and let the no-loss-of-area condition* (8.1) *hold. Let $\chi_0 \in BV(\Omega)$, $\mu_0 \in H^{1,2}(\Omega)$. Then there exists (m,μ,χ,v) with*

$$\chi \in L^\infty(0,T;\, BV(\Omega;\, \{0,1\})), \quad \operatorname{supp}\chi(t) \subset\subset \Omega \quad \textit{for all } 0 < t < T,$$

$$v \in L^1(0,T;\, L^1(\Omega;\mathbb{R};|\nabla\chi(t)|)),$$

$$\mu = (\mu^N,\mu^D), \qquad \mu^N \in L^2(0,T;\, H^{1,\frac{3}{2}}(\Omega)), \quad \mu^D \in L^2(0,T;\, H^{1,2}(\Omega)),$$

$$(1-\chi)\mu^N = 0, \quad (1-\chi)\mu^D = (1-\chi)\varphi \qquad \textit{for almost all } (x,t) \in \Omega_T$$

such that (m,μ,χ) is a weak solution in the sense of section 3.

REFERENCES

[1] F. Almgren, J. E. Taylor, and L. Wang, *Curvature-driven flows: A variational approach*, SIAM J. Control Optim., 31 (1993), pp. 387–437.

[2] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[3] V. I. Arnol'd, *Mathematical Methods of Classical Mechanics*, Springer, New York, 1989.

[4] S. Arnrich, T. Blesgen, and S. Luckhaus, *A General Theory for Elastic Phase Transitions in Crystals*, Preprint No. 147, revised version, DFG Priority Program 1095, Analysis, Modeling and Simulation of Multiscale Problems, Germany, 2006.

[5] K. Bente and T. Doering, *Solid-state diffusion in sphalerites—An experimental verification of the chalcopyrite disease*, European J. Mineralogy, 5 (1993), pp. 465–478.

[6] K. Bente and T. Doering, *Experimental studies on the solid-state diffusion of $Cu_{+}In$ in ZnS and on disease, DIS in sphalerite and their geological applications*, Mineralogy and Petrology, 53 (1995), pp. 285–305.

[7] T. Blesgen, *A revised model for diffusion induced segregation processes*, J. Math. Phys., 46 (2005), 022702.

[8] T. Blesgen, *A multiscale approach for quantitative simulations of diffusion induced segregation*, Modelling and Simulation in Materials Science and Engineering, 14 (2006), pp. 389–408.

[9] T. Blesgen and S. Luckhaus, *The dynamics of transition layers in solids with discontinuous chemical potentials*, Math. Methods Appl. Sci., 29 (2006), pp. 525–536.

[10] T. Blesgen, S. Luckhaus, and K. Bente, *Modeling and numerical simulation of diffusion induced segregation*, Crystal Research and Technology, 37 (2002), pp. 570–580.

[11] T. Blesgen, S. Luckhaus, and K. Bente, *Diffusion induced segregation in the case of the ternary system sphalerite, chalcopyrite and cubanite*, Crystal Research and Technology, 39 (2004), pp. 969–979.

[12] E. de Giorgi, *Una estensione del teorema di Bernstein*, Ann. Scuola Norm. Sup. Pisa (3), 19 (1965), pp. 79–85.

[13] H. Federer, *Geometric Measure Theory*, Springer, Berlin, 1996.

[14] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser-Verlag, Basel, 1984.

[15] P. Grindrod, *The Theory and Application of Reaction-Diffusion Equations, Patterns and Waves*, 2nd ed., Clarendon Press, Oxford, UK, 1996.

[16] J. S. Kirkaldy and D. J. Young, *Diffusion in the Condensed State*, The Institute of Metals, London, 1987.

[17] E. Kuwert and R. Schätzle, *Gradient flow for the Willmore functional*, Comm. Anal. Geom., 10 (2002), pp. 307–339.

[18] S. Luckhaus, *The Stefan Problem with the Gibbs-Thomson Law*, Preprint 2.75(591), Universita di Pisa, Pisa, 1991.

[19] S. Luckhaus and T. Sturzenhecker, *Implicit time discretization for the mean curvature flow equation*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 253–271.

[20] L. Onsager, *Reciprocal relations in irreversible processes* I, Phys. Rev., 37 (1931), pp. 405–426.

[21] L. Onsager, *Reciprocal relations in irreversible processes* II, Phys. Rev., 38 (1931), pp. 2265–2279.

[22] P. I. Plotnikov, *Generalized solutions to a free boundary problem of motion of a non-Newtonian fluid*, Siberian Math. J., 34 (1993), pp. 704–716.

[23] Y. G. Reshetnyak, *Weak convergence of completely additive vector functions on a set*, Sibirsk. Mat. Zh., 9 (1968), pp. 1039–1045.

[24] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1997.

[25] M. Röger, *Solutions for the Stefan problem with Gibbs-Thomson law by a local minimisation*, Interfaces Free Bound., 6 (2004), pp. 105–133.

[26] M. Röger, *Existence of weak solutions for the Mullins–Sekerka flow*, SIAM J. Math. Anal., 37 (2005), pp. 291–301.

[27] R. Schätzle, *Hypersurfaces with mean curvature given by an ambient Sobolev function*, J. Differential Geom., 58 (2001), pp. 371–420.

[28] R. Schätzle, *A counterexample for an approximation of the Gibbs-Thomson law*, Adv. Math. Sci. Appl., 7 (1997), pp. 25–36.

[29] L. Simon, *Lectures on Geometric Measure Theory*, Australian National University, Canberra, 1983.

[30] W. Ziemer, *Weakly Differentiable Functions*, Springer, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 39, No. 4, pp. 1295–1327

THE BOUNDARY BEHAVIOR OF BLOW-UP SOLUTIONS RELATED TO A STOCHASTIC CONTROL PROBLEM WITH STATE CONSTRAINT*

TOMMASO LEONORI† AND ALESSIO PORRETTA†

Abstract. We consider solutions of the equation $-\Delta u + \lambda u + |\nabla u|^q = f$, which blow up uniformly at the boundary of a smooth domain, that can be interpreted as the value function of a state constraint control problem for a Brownian motion. We prove a complete asymptotic expansion of the gradient at the boundary, giving the precise behavior of normal and tangent components. The result is achieved by proving Lipschitz regularity for $u - S$, where S is an explicit singular corrector term. As the main motivation and application of our result, we characterize the behavior of the singular optimal control law and of the constrained dynamics near the boundary.

Key words. nonlinear elliptic equations, boundary blow-up solutions, asymptotic expansion, state constraint problem for the Brownian motion

AMS subject classifications. 35K60 (35B37, 49J20)

DOI. 10.1137/070681363

1. Introduction.

In this work we study the solutions of the following elliptic equation:

$$-\Delta u + |\nabla u|^q + \lambda u = f(x) \qquad \text{in} \;\; \Omega \tag{1.1}$$

which satisfy the singular boundary condition

$$u(x) \to +\infty \quad \text{as } x \to \partial\Omega, \tag{1.2}$$

where Ω is a bounded smooth subset in $\mathbb{R}^N$, $N \geq 2$. In (1.1), we assume that $1 < q \leq 2$, $\lambda > 0$, and f satisfies suitable regularity conditions.

Let us stress that the condition $1 < q \leq 2$ is necessary in order that solutions satisfying (1.2) exist. Indeed, if $q > 2$, any solution of (1.1) can be proved to be bounded (universally) and Hölder continuous up to the boundary (see [17]); hence, the maximal trace in that case is bounded: This explains why the study of (1.1) for $q > 2$ needs, in general, a completely different approach.

The study of solutions of (1.1)–(1.2) was suggested in a pioneering paper by Lasry and Lions [15] motivated by a state constraint problem for the Brownian motion, which was roughly presented in an intuitive way as the problem of "constraining a Brownian motion in a given domain by controlling its drift." More precisely, given a Brownian motion B_t (on a standard probability space) and a diffusion process X_t which solves the stochastic differential equation

$$\begin{cases} dX_t = a_t dt + dB_t, \\ X_0 = x \in \Omega, \end{cases} \tag{1.3}$$

they considered the problem of finding optimal feedback controls (i.e., controls $a_t = a(X_t)$, where $a \in C(\Omega; \mathbb{R}^N)$) such that X_t never leaves the domain Ω. Clearly, as

*Received by the editors January 30, 2007; accepted for publication (in revised form) August 7, 2007; published electronically December 12, 2007.
http://www.siam.org/journals/sima/39-4/68136.html
†Dipartimento di Matematica, Università di Roma "Tor Vergata," Via della Ricerca Scientifica 1, 00133 Roma, Italy (leonori@mat.uniroma2.it, porretta@mat.uniroma2.it).

explained in [15], a control has to be singular (at $\partial\Omega$) in order to realize a similar state constraint (for a nondegenerate diffusion). The criterion for optimality was given by the cost functional

$$J(x,a) = E\int_0^\infty \left\{ f(X_t) + C_q\,|a(X_t)|^{q'} \right\} e^{-\lambda\, t}dt, \tag{1.4}$$

where E is the expected value, $q' = \frac{q}{q-1}$, $C_q = (q-1)q^{-\frac{q}{q-1}}$, and $e^{-\lambda t}$ is a discount factor. Then the value function

$$u(x) = \inf_{a\in\mathcal{A}} J(x,a), \quad \mathcal{A} = \{a \in C(\Omega) : X_t \in \Omega\, \forall t > 0, \text{a.s.}\} \tag{1.5}$$

was proved to be the maximal solution of (1.1). We note that here there was no restriction on q, but for $q > 1$. Then, in the case $1 < q \le 2$, they proved that u is the unique solution of (1.1)–(1.2) (in the sense that $u \in W^{2,p}_{\rm loc}(\Omega)$ for any $p < \infty$ and satisfies (1.2) uniformly) and moreover that the optimal feedback state control law is given by

$$a(\cdot) = -q|\nabla u(\cdot)|^{q-2}\nabla u(\cdot); \tag{1.6}$$

i.e., $a(X_t)$ is the unique optimal control. Some asymptotic estimates for u were also proved (in [15] and also in [2]) and in particular that

$$\begin{cases} u(x) \sim C^* d(x)^{-\frac{2-q}{q-1}} \quad \text{as} \quad d(x) \to 0 \qquad \text{if } 1 < q < 2, \\ u(x) \sim -\log(d(x)) \quad \text{as} \quad d(x) \to 0 \qquad \text{if } q = 2, \end{cases} \tag{1.7}$$

where, for $x \in \Omega$, $d(x)$ denotes the distance to the boundary of Ω and C^* is a universal constant, precisely $C^* = \frac{(q-1)^{-\frac{2-q}{q-1}}}{2-q}$.

Let us note that the rate of explosion decreases as q gets close to 2, which is consistent with the fact that, when $q > 2$, solutions of (1.1) are bounded. Further results for the case $q > 2$ can be found in [15], although a characterization of the optimal control law such as (1.6) is missing in that case.

Once more, we stress that we will restrict to the range $1 < q \le 2$, when blow-up solutions exist, and moreover, after the results in [15], the unique optimal control $a(X_t)$ is explicitly given through (1.6). In particular, this gives significant motivation for studying the asymptotic behavior of ∇u, which locally determines, near the boundary, the constrained dynamics (1.3). A first estimate in this sense has been proved in [18], using (1.7) and scaling and blow-up arguments, precisely that

$$\lim_{x\to x_0\in\partial\Omega} \quad d(x)^{\frac{1}{q-1}}\nabla u(x) = (q-1)^{-\frac{1}{q-1}}\,\nu(x_0), \tag{1.8}$$

where $\nu(x)$ is the outward unit vector on $\partial\Omega$. Note that $(q-1)^{-\frac{1}{q-1}} = C^*\frac{2-q}{q-1}$, so that (1.8) is the expected "derivation" of (1.7). Moreover, from (1.7) and (1.8) we observe that the first order asymptotic behavior of u and ∇u depends only on q and is described by the one-dimensional solution of the corresponding ODE.

The aim of the present paper is to give a more precise description of the blow-up of ∇u in order both to point out the influence of the geometry of the domain (by looking at second order effects) and to get a complete picture of the local behavior, near the boundary, of the controlled dynamics. In particular, by studying the asymptotic expansion of ∇u, we are able to detail the roles of normal and tangential directions and the influence of the boundary curvature in the behavior of the optimal control law defined in (1.6). To give a rough idea of the main consequence of our results, we are going to prove that the optimal control is tangentially bounded, it blows up pointing in the inward normal direction, and it achieves its maximum in those points (close to the boundary) where the domain has a maximal mean curvature. Actually this proves (as intuition suggests, when a uniform diffusion is constrained) that the control has to be "stronger" where the domain is more curved.

The above-mentioned properties are contained in the following result on the asymptotic behavior of the optimal (feedback state) control law:

$$a(x) = -q|\nabla u(x)|^{q-2}\nabla u(x). \tag{1.9}$$

Notation. Let us stress that from now on we will denote by $d(x)$ (often simply d) a positive smooth function such that $d(x) \equiv \mathrm{dist}(x, \partial\Omega)$ in a suitable inner neighborhood of $\partial\Omega$. Moreover, we will indicate by $\nu(x) = -\nabla d(x)$ (hence $\nu(x)$ is the outward unit normal vector if $x \in \partial\Omega$) and by $\tau(x)$ any unit vector field at x which satisfies $\tau(x) \cdot \nu(x) = 0$.

Observe that, since Ω will be assumed to be at least of class C^2, any $x \in \Omega$ lying in a sufficiently small neighborhood of the boundary admits a unique projection onto $\partial\Omega$, denoted as $\overline{x} = Proj_{\partial\Omega}(x)$. Moreover, for $\varsigma \in \partial\Omega$, we denote by $H(\varsigma)$ the mean curvature of $\partial\Omega$ computed at ς.

Finally, we write $f(x) = O(g(x))$ in order to say that $|f(x)| \le C\,|g(x)|$ in Ω.

THEOREM 1.1. *Let Ω be a smooth bounded open subset of $\mathbb{R}^N$, and let $H(\overline{x})$ be the mean curvature of $\partial\Omega$ computed at $\overline{x}$, where $\overline{x}$ is the projection of x on $\partial\Omega$.*

Let $f \in W^{1,\infty}(\Omega)$, let u be the unique solution of (1.1)–(1.2) *and $a(x)$ be the optimal control law* (1.9). *Then for any $1 < q < 2$ we have, as $d(x) \to 0$,*

$$a(x) = -\frac{q'}{d(x)}\nu(x) - \frac{q'(N-1)}{2}H(\overline{x})\nu(x) + \eta(x), \tag{1.10}$$

where $\eta(x)$ is such that

$$|\eta(x)| = \begin{cases} O(d^{\frac{2-q}{q-1}}(x)) & \text{if } \ \frac{3}{2} < q < 2, \\ O(d(x)\log d(x)) & \text{if } \ q = \frac{3}{2}, \\ O(d(x)) & \text{if } \ 1 < q < \frac{3}{2}. \end{cases}$$

If $q = 2$, we have, as $d(x) \to 0$,

$$a(x) = -\frac{2}{d(x)}\nu(x) - (N-1)[H(\overline{x}) + o(1)]\nu(x) + \psi(x)\tau(x), \tag{1.11}$$

where $\tau(x) \in \mathbb{R}^N$, $|\tau| = 1$, $\tau \cdot \nu = 0$, and $\psi \in L^\infty(\Omega)$.

To our knowledge, such characterization of the behavior of singular feedback controls related to (1.3)–(1.5) is new in the literature. Actually, similar state constraint problems have been considered in previous works mainly in the case of degenerate diffusions and bounded controls and for a bounded value function (see also, e.g., [7], [6], [13], [14], and references therein), while the case of nondegenerate diffusions and a singular value function seems not to have been developed since the reference work [15].

The results in Theorem 1.1 are the main consequence of our study of second order terms in the asymptotics of the gradient of solutions of (1.1)–(1.2). Note that the estimate (1.8) already implies that $\frac{\partial u(x)}{\partial \tau} = o(\frac{\partial u(x)}{\partial \nu})$ as $d(x) \to 0$, and hence that tangential effects are of lower order. Actually we improve that estimate by proving the following.

THEOREM 1.2. *Let Ω be a smooth bounded open subset of $\mathbb{R}^N$ and $f \in W^{1,\infty}(\Omega)$, and let u be the unique solution of* (1.1)–(1.2). *Then we have, as $d(x) \to 0$, (with the above notations):*

$$\frac{\partial u(x)}{\partial \nu} = \frac{(q-1)^{-\frac{1}{q-1}}}{d(x)^{\frac{1}{q-1}}} \left[1 + \frac{(N-1)H(\overline{x})}{2} d(x) + o(d(x))\right] \qquad \forall\, 1 < q \leq 2, \tag{1.12}$$

and

$$\begin{cases} \frac{\partial u(x)}{\partial \tau} \in L^\infty(\Omega) & \text{if } \frac{3}{2} < q \leq 2, \\ \frac{\partial u(x)}{\partial \tau} = O\left(|\log d|\right) & \text{if } q = \frac{3}{2}, \\ \frac{\partial u(x)}{\partial \tau} = O\left(d^{\frac{2q-3}{q-1}}\right) & \text{if } 1 < q < \frac{3}{2}. \end{cases} \tag{1.13}$$

Note that if $q \leq \frac{3}{2}$ one has that $\frac{\partial u}{\partial \tau}$ blows up, but the optimal control law (1.9), which is the only relevant quantity for the dynamics, always remains tangentially bounded. Actually, the precise behavior of the tangential gradient represents a significant improvement with respect to the estimate (1.8) and, together with the second order terms of normal gradient, allows one to regard the dyamics near the boundary as

$$\begin{cases} dX_t = V(X_t)dt + dB_t, \\ X_0 = x \in \Omega, \end{cases}$$

where $V(x) = -[\frac{q'}{d(x)} + \frac{q'(N-1)}{2} H(\overline{x})]\nu(x)$ (here $q < 2$). Note that this corresponds to a linearization of (1.1), and the behavior of the process can then be described explicitly. In particular, observe how the drift acts differently on points which have the same distance to the boundary but different curvatures.

Let us point out that curvature effects in the boundary blow-up of solutions of elliptic equations were already observed previously in the case of semilinear equations with absorption zeroth order terms. In that context, it was recently proved in [10], [4] (see also references therein) how second order terms in the blow-up of u precisely depend on the mean curvature of the boundary. These results, which also motivated our work, are obtained through a refined construction of sub- and supersolutions. However, in the context of solutions of (1.1)–(1.2), the main local features of the blow-up are observed in the asymptotic behavior of ∇u, which cannot be studied using only comparison functions. Therefore, in some sense our results extend those proved in a different context in [10], [4], although we use a completely different method. Indeed, our estimates on ∇u in (1.12)–(1.13) are not derived by some asymptotic estimate on u, as was the case, for instance, for (1.8), which is proved in [18] using (1.7) and a scaling argument. A similar technique based on second order estimates for u is not possible here unless it is restricted to a smaller range of values of q (see section 3.2 for more details). We develop instead a totally different approach which, through a regularity result for solutions of (1.1)–(1.2), directly leads to a complete asymptotic expansion of ∇u as a vector field.

More precisely, we introduce as a corrector term the formal asymptotic expansion $S = d(x)^{-\frac{2-q}{q-1}} \sum_{k=0}^{m} \sigma_k d(x)^k$ (note that u rescales like $d(x)^{-\frac{2-q}{q-1}}$ from (1.7); this is for $q < 2$), and we prove that $u - S$ is Lipschitz in Ω, up to a suitable (unique and explicit) choice of the coefficients σ_k. As a consequence, we obtain all singular terms in the asymptotic expansion of ∇u. In order to give a proper statement of this result, which is the main content of the article, we denote by α the following number:

$$\alpha = \frac{2-q}{q-1}, \tag{1.14}$$

and we observe that

$$\begin{cases} \alpha = 0 & \iff \quad q = 2\,, \\ 0 < \alpha < 1 & \iff \quad \frac{3}{2} < q < 2, \\ \alpha \geq 1 & \iff \quad 1 < q \leq \frac{3}{2}. \end{cases}$$

THEOREM 1.3. *Let $1 < q \leq 2$ and $u(x)$ be the solution of* (1.1)–(1.2). *Assume that $\partial\Omega \in C^r$, $r = [\alpha] + 5$, and $f(x)$ is a $W^{1,\infty}(\Omega)$ function. Then there exist smooth functions $\sigma_0, \sigma_1, \ldots, \sigma_{[\alpha]+1}$, which can be explicitly determined, such that setting*

$$\begin{cases} S = \sum_{k=0}^{[\alpha]+1} \sigma_k d^{k-\alpha} & \text{if } \alpha \notin \mathbb{N}, \\ S = \sum_{k=0}^{\alpha-1} \sigma_k d^{k-\alpha} + \sigma_\alpha \log d + \sigma_{\alpha+1} d \log d & \text{if } \alpha \in \mathbb{N},\ \alpha \geq 1, \\ S = -\log d & \text{if } \alpha = 0, \end{cases} \tag{1.15}$$

we have

$$u - S \in W^{1,\infty}(\Omega).$$

As a consequence, we have for any $\alpha > 0$:

- *if $\alpha \notin \mathbb{N}$:*

$$\begin{cases} \frac{\partial u(x)}{\partial \nu} - \frac{\alpha C^*}{d^{\alpha+1}} + \sum_{k=1}^{[\alpha]+1} \left[\frac{(k-\alpha)\sigma_k(x)}{d^{\alpha-k+1}(x)} - \frac{\nabla \sigma_{k-1}(x)\cdot\nu}{d^{\alpha-k+1}(x)} \right] \in L^\infty(\Omega), \\ \frac{\partial u(x)}{\partial \tau} - \sum_{k=1}^{[\alpha]} \frac{\nabla\sigma_k(x)\cdot\tau}{d^{\alpha-k}} \in L^\infty(\Omega); \end{cases} \tag{1.16}$$

- *if $\alpha \in \mathbb{N}$:*

$$\begin{cases} \frac{\partial u(x)}{\partial \nu} - \frac{\alpha C^*}{d^{\alpha+1}} + \sum_{k=1}^{\alpha-1} \left[\frac{(k-\alpha)\sigma_k(x)}{d^{\alpha-k+1}(x)} - \frac{\nabla \sigma_{k-1}(x)\cdot\nu}{d^{\alpha-k+1}(x)} \right] + \frac{(\sigma_\alpha - \nabla\sigma_{\alpha-1}\cdot\nu)}{d(x)} \in L^\infty(\Omega), \\ \frac{\partial u(x)}{\partial \tau} - \sum_{k=1}^{\alpha-1} \frac{\nabla\sigma_k(x)\cdot\tau}{d^{\alpha-k}} - (\nabla\sigma_\alpha\cdot\tau)\log d(x) \in L^\infty(\Omega). \end{cases} \tag{1.17}$$

[Let us observe, with respect to the above statement, that, due to (1.14), one has $\alpha \in \mathbb{N} \iff \frac{q}{q-1} \in \mathbb{N}$; i.e., the case $\alpha \in \mathbb{N}$ corresponds to having a cost of integer power type in the functional (1.4).]

The proof of Theorem 1.3 will be given in section 2, consisting of two main ideas. The first one is to look at the equation satisfied by the function $z = u - S$, keeping in mind the first estimate (1.8). Indeed, thanks to (1.8) we have that, near the boundary, $|\nabla z| = o(|\nabla S|)$, which suggests us to write the equation of z as

$$-\Delta z + \lambda z + |\nabla z + \nabla S|^q - |\nabla S|^q = f + F,$$

where $F = \Delta S - \lambda S - |\nabla S|^q$. Here we can prove that $|\nabla z + \nabla S|^q - |\nabla S|^q$ behaves similarly as $q|\nabla S|^q \nabla z \cdot \frac{\nabla S}{|\nabla S|^2} \sim -C\frac{\nabla z \cdot \nabla d}{d}$, so that the left-hand side is a singular but regularizing operator with respect to z. Second, choosing the coefficients of S in a way that F is smooth, we obtain gradient estimates depending only on the regularity of f. These gradient estimates, which are the crucial point in our proof, are obtained by using the classical Bernstein method (see [8]) largely developed by Serrin and Lions (see, e.g., [19], [17], [16], [5]), although we use here a slightly different approach which is more adapted to our situation. Last but not least, we are interested in getting global estimates for the gradients in the whole of Ω and not merely local interior estimates. Following ideas in [17], [16], we make it possible by working with a Neumann-type condition at the boundary. Thus we introduce a regular approximation u_n of the solution u of (1.1)–(1.2), constructed through the following Neumann-type problems:

$$\begin{cases} -\Delta u_n + \lambda u_n + |\nabla u_n|^q = f_n & \text{in } \Omega, \\ \dfrac{\partial u_n}{\partial \nu} = \dfrac{\partial S_n}{\partial \nu} & \text{on } \partial\Omega, \end{cases}$$

where $S_n = \sum_{k=0}^{m} \sigma_k (d(x)+\frac{1}{n})^{k-\alpha}$ (which, for n fixed, is not singular). Here we prove a uniform (in n) version of the preliminary estimates (1.7) and (1.8), and then we prove the Bernstein-type estimates on $z_n = u_n - S_n$, obtaining the Lipschitz regularity claimed in Theorem 1.3.

Note that the results of Theorems 1.1 and 1.2 immediately follow from (1.16)–(1.17) by computing σ_0 and σ_1 (in particular, one has $\sigma_0 = C^*$). Only in the case where $q = 2$ we need more information in order to get (1.12) (and then (1.11)), which will be proved with a simple blow-up argument. In fact, we will show that the normal derivative of $u - S$ tends to zero on the boundary, refining the result in Theorem 1.3.

Finally, let us point out that the regularity condition on f in Theorem 1.3 (and consequently in Theorems 1.1–1.2) can be highly weakened, thanks to our approach to Bernstein's estimates. This will be detailed in section 3, where we prove an extended version of Theorem 1.3 in the case of f singular at the boundary (see Theorem 3.2) as well as an intermediate result (such as Hölder-type regularity for $u - S$) in cases where the singularity of f does not allow $u - S$ to be Lipschitz.

2. Proof of Theorem 1.3.

In order to simplify some expressions, we set $\lambda = 1$ in (1.1); in fact, our results are not really affected by changing the values of λ (see also Remark 2.8).

2.1. The case $1 < q < 2$: Preliminary results.

Let us introduce some notations: For $1 < q < 2$ we set

$$\alpha = \frac{2-q}{q-1} \quad \text{and} \quad C^* = \frac{(q-1)^{-\alpha}}{2-q}, \tag{2.1}$$

and we will often use the following properties of α and C^*:

$$q(\alpha+1) = \alpha + 2 \quad \text{and} \quad (\alpha C^*)^q = \alpha(\alpha+1)C^*. \tag{2.2}$$

Moreover from now on we denote

$$d_n(x) = d(x) + \frac{1}{n},$$

and we will indicate by C various constants (independent on n) whose value may vary from line to line. Finally, we will write $f(x) = O(g(x))$ in order to say that $|f(x)| \leq C\,|g(x)|$ in Ω.

Since we need a suitable smooth approximation of the solution of (1.1)– (1.2), we will consider the sequence $\{u_n\}$ of solutions of the Neumann problem:

$$\begin{cases} -\Delta u_n + u_n + |\nabla u_n|^q = f_n(x) & \text{in}\,\Omega, \\ \dfrac{\partial u_n}{\partial \nu} = \dfrac{\partial S_n}{\partial \nu} & \text{on}\,\partial\Omega, \end{cases} \tag{2.3}$$

where f_n is a suitable regular approximation of f and

$$S_n = C^* d_n(x)^{-\alpha} + \psi(x) d_n(x)^{1-\alpha},$$

with $\psi(x)$ smooth (to be precised later).

Observe that, for fixed n, f_n and S_n being smooth, the existence of a classical solution $u_n \in C^2(\overline{\Omega})$ is a standard result (see, e.g., [16]). The goal of this subsection is to prove the following preliminary result.

Proposition 2.1. *Let u_n be a solution of* (2.3)*, with $1 < q < 2$. Assume that $f_n = O(d_n^{-\alpha-1})$, $|\nabla f_n| \leq C\, d_n^{-\alpha-2}$, and $S_n = C^* d_n^{-\alpha} + \psi(x) d_n^{1-\alpha}$, with $\psi(x)$ smooth. Then we have*

$$|\nabla u_n| \leq C d_n^{-\alpha-1} \qquad \textit{in } \Omega, \tag{2.4}$$

and

$$\lim_{d_n(x)\to 0} |d_n^{\alpha+1}(x) \nabla u_n(x) - \alpha C^* \nu(x)| = 0. \tag{2.5}$$

Note that, since $d_n(x) = d(x) + \frac{1}{n}$, the limit as $d_n(x) \to 0$ means that both n tends to ∞ and $d(x) \to 0$, as a two-variable limit.

The proof of Proposition 2.1 will be achieved through the following lemmas, which contain first order (uniform) estimates for u_n of the same kind as (1.7) and (1.8) proved in [15] and in [18], respectively.

Lemma 2.2. *Let u_n be a solution of* (2.3)*. Then there exists $C > 0$ such that:*

- *for* $0 < \alpha < 1$

$$|u_n - C^* d_n^{-\alpha}| \leq C, \tag{2.6}$$

- *for* $\alpha = 1$

$$\left| u_n - \frac{C^*}{d_n} \right| \leq C \left| \log d_n \right|, \tag{2.7}$$

- *for* $\alpha > 1$,

$$|u_n - C^* d_n^{-\alpha}| \leq C d_n^{1-\alpha}, \tag{2.8}$$

where α and C^* have been defined in (2.1).

Proof. Let us first consider $0 < \alpha < 1$ and

$$\overline{\psi}_n(x) = C^* d_n^{-\alpha} - M_0 d_n^{1-\alpha} + M_1,$$

where $M_0,\ M_1 > 0$. We have

$$
\begin{aligned}
&-\Delta\overline{\psi}_n + |\nabla\overline{\psi}_n|^q + \overline{\psi}_n - f_n(x) \\
&= -\alpha(\alpha+1)C^* d_n^{-\alpha-2} + \Delta d\alpha C^* d_n^{-\alpha-1} - \alpha M_0(1-\alpha)d_n^{-\alpha-1} + M_0(1-\alpha)d_n^{-\alpha} \\
&\quad + (\alpha C^*)^q d_n^{-q(\alpha+1)}\left|1 + \frac{M_0(1-\alpha)}{\alpha C^*}d_n\right|^q + C^* d_n^{-\alpha} - M_0 d_n^{1-\alpha} + M_1 - f_n(x).
\end{aligned}
$$

Thus, using (2.1)–(2.2), after standard computations (here we use that, in a neighborhood of $t = 0$, $(1+t)^q = 1 + qt + O(t^2)$) we find

(2.9)
$$
-\Delta\overline{\psi}_n + |\nabla\overline{\psi}_n|^q + \overline{\psi}_n - f_n(x) \geq \left[2M_0(1-\alpha) + \alpha C^*\Delta d\right] d_n^{-\alpha-1} - f_n(x) + O(d_n^{-\alpha}) + M_1.
$$

Moreover

$$
\frac{\partial\overline{\psi}_n}{\partial\nu} = \alpha C^* n^{\alpha+1} + (1-\alpha)M_0 n^{\alpha} \quad \text{on } \partial\Omega. \tag{2.10}
$$

Hence, by choosing M_0 and M_1 large enough we get that the right-hand side of (2.9) is positive and, by (2.10), $\partial_\nu\overline{\psi}_n \geq \partial_\nu S_n$ on $\partial\Omega$, so that $\overline{\psi_n}(x)$ turns out to be a supersolution for (2.3).

Similarly, arguing as above, there exist $M_2,\ M_3 > 0$ such that

$$
\underline{\psi}_n(x) = C^* d_n^{-\alpha} + M_2 d_n^{1-\alpha} - M_3
$$

is a subsolution for (2.3). Hence by the maximum principle we deduce that

$$
\underline{\psi}_n(x) \leq u_n(x) \leq \overline{\psi}_n(x),
$$

so that (2.6) holds true.

In the same way we prove estimates (2.7) and (2.8) by choosing, respectively,

$$
\overline{\psi}_n(x) = \frac{4}{d_n} - M_0\log d_n + M_1, \qquad \underline{\psi}_n(x) = \frac{4}{d_n} + M_2\log d_n - M_3, \quad \text{if } \alpha = 1,
$$

and

$$
\overline{\psi}_n(x) = C^* d_n^{-\alpha} + M_0 d_n^{1-\alpha} + M_1, \qquad \underline{\psi}_n(x) = C^* d_n^{-\alpha} - M_2 d_n^{1-\alpha} - M_3, \quad \text{if } \alpha > 1,
$$

where $M_i > 0$ are large enough, for $i = 0, 1, 2, 3$. □

Remark 2.3. Let us note that (2.6), (2.7), and (2.8) imply

$$
\lim_{d_n(x)\to 0} |d_n^{\alpha}(x)u_n(x) - C^*| = 0. \tag{2.11}
$$

We will often use this weaker result later. We also remark that in order to obtain (2.11) it is enough to ask that $f_n = O(d_n^{-\gamma})$, with $\gamma < \alpha + 2$. The proof is similar but for a different choice of the sub and supersolutions; for instance, it is enough to choose $\overline{\psi_n}(x) = C^* d_n(x)^{-\alpha} + \rho_n(x)$, where $\rho_n(x) = O(d_n(x)^{2-\gamma})$.

The next lemma will be very useful in getting global Bernstein-type estimates, as, for instance, in [17], [16]. Indeed it allows us to get information about the normal derivative of a weighted power of the gradient of a solution of (2.3). Observe that the homogeneous Neumann condition is crucial in order to prove such a result.

LEMMA 2.4. *Let η be a smooth (say, $C^2(\overline{\Omega})$) function such that*

$$\frac{\partial \eta}{\partial \nu} = 0 \quad on \ \ \partial\Omega,$$

and let $\Phi(s) \in C^1(0,a)$, $a > 0$, be a positive increasing function such that,

$$\forall 0 < s < a, \qquad \Phi'(s) - C_0\Phi(s) > 0, \tag{2.12}$$

where $C_0 = 2\|D^2 d\|_\infty$. Then

$$\frac{\partial}{\partial \nu}\Big[|\nabla \eta|^2 \Phi(d)\Big] \le 0 \quad on\ \partial\Omega. \tag{2.13}$$

Remark 2.5. Let us observe that both $\Phi(s) = s^\beta$ and $\Phi(s) = e^{\theta(s)}$ satisfy condition (2.12) for any $\beta > 0$ and for any increasing $\theta(s)$ such that $\theta(0) = 0$ and $\theta'(0) > C_0$, respectively. Moreover, we remark that if $\Phi(s)$ satisfies (2.12), the same is true for $\Phi(s + \frac{1}{n})$ if n is large enough.

Proof. Let us compute

$$\frac{\partial |\nabla \eta|^2 \Phi(d)}{\partial \nu} = -\Phi'(d)|\nabla \eta|^2 + 2\Phi(d)\,(D^2\eta\,\nabla\eta\,\nu)\,;$$

since η satisfies the homogeneous Neumann boundary condition, there exists $\mu(x)$ such that on $\partial\Omega$ we have

$$\mu\nu = D(\nabla\eta\cdot\nu) = D^2\eta\ \nu + D\nu\nabla\eta.$$

Thus using again the boundary condition $\frac{\partial \eta}{\partial \nu} = 0$ and since $D\nu = -D^2 d$, we get

$$\frac{\partial |\nabla \eta|^2 \Phi(d)}{\partial \nu} = -\Phi'(d)|\nabla \eta|^2 + 2\Phi(d) D^2 d\,\nabla\eta\,\nabla\eta \le |\nabla\eta|^2\Big[2\Phi(d)\|D^2 d\|_\infty - \Phi'(d)\Big],$$

and by (2.12) we deduce (2.13). ∎

Here we prove a first estimate on the gradient of the solution of the problem (a suitable translation of (2.3)):

$$\begin{cases} -\Delta z_n + z_n + |\nabla z_n + G_n|^q = F_n & \text{in } \Omega, \\ \dfrac{\partial z_n}{\partial \nu} = 0 & \text{on } \partial\Omega. \end{cases} \tag{2.14}$$

LEMMA 2.6. *Let $z_n \in C^2(\overline{\Omega})$ be a solution of* (2.14), *where G_n and F_n are a vector field and a function, respectively, such that*

$$|G_n| \le C d_n^{-\alpha-1}, \quad |DG_n| \le C d_n^{-\alpha-2}, \quad |F_n| \le C d_n^{-\alpha-2}, \quad |\nabla F_n| \le C d_n^{-\alpha-3}. \tag{2.15}$$

Then we have

$$|\nabla z_n| \le C d_n^{-\alpha-1}. \tag{2.16}$$

Proof. Observe first that the maximum principle and the assumptions on F_n and G_n imply $|z_n| \le C\, d_n^{-\alpha-2}$. Let us define $w_n = |\nabla z_n|^2 e^{\theta d_n}$, where θ is a positive number. By Lemma 2.4 we have, for θ and n large enough,

$$\frac{\partial w_n}{\partial \nu} \le 0 \quad \text{on}\,\partial\Omega. \tag{2.17}$$

On the other hand,

$$\Delta w_n = \theta^2 w_n + \theta \Delta d w_n + 2\theta \nabla w_n \cdot \nabla d - 2\theta^2 w_n + 2e^{\theta d_n}\left[\nabla \Delta z_n \cdot \nabla z_n + |D^2 z_n|^2\right],$$

and so, since by the Cauchy–Schwarz inequality

$$|D^2 z_n|^2 \geq \frac{1}{N}(\Delta z_n)^2,$$

we have, recalling that $\Delta z_n = z_n + |\nabla z_n + G_n|^q - F_n$,
(2.18)

$$\begin{aligned}\Delta w_n \geq (-\theta^2 + \theta\Delta d) w_n + 2\theta \nabla w_n \cdot \nabla d + 2e^{\theta d_n}\Bigg[\frac{1}{N}\left(|\nabla z_n + G_n|^q + z_n - F_n\right)^2 \\ + q\nabla z_n \cdot (D^2 z_n + DG_n)\frac{\nabla z_n + G_n}{|\nabla z_n + G_n|^{2-q}} - \nabla F_n \cdot \nabla z_n + |\nabla z_n|^2\Bigg].\end{aligned}$$

Since $|z_n| \leq C\, d_n^{-\alpha-2}$, we have

$$\frac{1}{N}\left(|\nabla z_n + G_n|^q + z_n - F_n\right)^2 \geq c_N |\nabla z_n + G_n|^{2q} - C\, d_n^{-2(\alpha+2)} - |F_n|^2.$$

Moreover, the Young inequality with exponents $(\frac{2q}{q-1}, \frac{2q}{q+1})$ implies

$$\left|\frac{\nabla z_n + G_n}{|\nabla z_n + G_n|^{2-q}} \cdot \Big(D^2 z_n + DG_n\Big)\nabla z_n\right|$$

$$\leq \frac{c_N}{2q}|\nabla z_n + G_n|^{2q} + C\left|\frac{1}{2}\nabla|\nabla z_n|^2 + DG_n \nabla z_n\right|^{\frac{2q}{q+1}},$$

and similarly, for all $\delta > 0$, we have

$$|\nabla F_n \cdot \nabla z_n| \leq \delta |\nabla z_n|^{2q} + \frac{C}{\delta^{\frac{1}{2q-1}}}|\nabla F_n|^{\frac{2q}{2q-1}}.$$

Thus we get from (2.18)

$$\begin{aligned}-\Delta w_n + c_N e^{\theta d_n}|\nabla z_n + G_n|^{2q} \leq (\theta^2 - 2 - \theta\Delta d) w_n \\ - 2\theta \nabla w_n \cdot \nabla d + 2e^{\theta d_n}\Bigg[\delta|\nabla z_n|^{2q} + \frac{C}{\delta^{\frac{1}{2q-1}}}|\nabla F_n|^{\frac{2q}{2q-1}} \\ + C\left|\frac{1}{2}\nabla|\nabla z_n|^2 + DG_n\nabla z_n\right|^{\frac{2q}{q+1}} + C\, d_n^{-2(\alpha+2)} + |F_n|^2\Bigg].\end{aligned} \tag{2.19}$$

Let us note that, since $\frac{2q}{q+1} > 1$ ($q > 1$) and $\nabla|\nabla z_n|^2 = e^{-\theta d_n}(\nabla w_n - \theta w_n \nabla d)$, we have

(2.20)
$$\left|\frac{1}{2}\nabla|\nabla z_n|^2 + DG_n\nabla z_n\right|^{\frac{2q}{q+1}} \le \tilde{C}\left[|\nabla w_n|^{\frac{2q}{q+1}} + |w_n|^{\frac{2q}{q+1}} + \Big(|DG_n||\nabla z_n|\Big)^{\frac{2q}{q+1}}\right]$$
$$\le \tilde{C}\left[|\nabla w_n|^{\frac{2q}{q+1}} + |w_n|^{\frac{2q}{q+1}}\right] + \eta|\nabla z_n|^{2q} + C\eta^{-\frac{1}{q}}|DG_n|^2,$$

where η is any positive constant. Using that, by convexity,

$$|\nabla z_n + G_n|^{2q} \ge \frac{1}{2^{2q-1}}|\nabla z_n|^{2q} - |G_n|^{2q},$$

we deduce from (2.19) and (2.20), choosing δ and η small enough:

$$-\Delta w_n + \tilde{C}_N|\nabla z_n|^{2q} \le w_n(\theta^2 - 2 - \theta\Delta d) - 2\theta\nabla w_n\cdot\nabla d + Ce^{\theta d_n}\Big[|\nabla F_n|^{\frac{2q}{2q-1}}$$
$$+ |\nabla w_n|^{\frac{2q}{q+1}} + w_n^{\frac{2q}{q+1}} + |DG_n|^2 + |G_n|^{2q} + C\, d_n^{-2(\alpha+2)} + |F_n|^2\Big].$$

Finally, since $\frac{2q}{q+1} < q$, using the Young inequality and that $1 \le e^{\theta d_n} \le C_\Omega$, we conclude that there exists $\tau > 0$ and constants C_0, $C_1 > 0$ such that

$$-\Delta w_n + \tau w_n^q - C_0|\nabla w_n|^{\frac{2q}{q+1}} \le C_1\left(|\nabla F_n|^{\frac{2q}{2q-1}} + |DG_n|^2 + |G_n|^{2q} + |F_n|^2 + d_n^{-2(\alpha+2)}\right).$$

The above inequality and (2.17) say that w_n is a subsolution for the problem

$$\begin{cases} -\Delta\psi + \tau\psi^q - C_0|\nabla\psi|^{\frac{2q}{q+1}} = \tilde{F}_n & \text{in } \Omega,\\ \dfrac{\partial\psi}{\partial\nu} = 0 & \text{on } \partial\Omega,\end{cases}$$

where

$$\tilde{F}_n = C_1\left(|\nabla F_n|^{\frac{2q}{2q-1}} + |DG_n|^2 + |G_n|^{2q} + |F_n|^2 + d_n^{-2(\alpha+2)}\right).$$

Let us note that from (2.15), using that

$$\frac{2q}{2q-1}(\alpha+3) = 2q(\alpha+1) = 2(\alpha+2), \tag{2.21}$$

we obtain that $|\tilde{F}_n| \le C_F d_n^{-2(\alpha+2)}$. In order to prove (2.16), let us prove that $\varphi = t d_n^{-2(\alpha+1)}$ is a supersolution for large $t \in \mathbb{R}^+$. First let us observe that $\frac{\partial\varphi}{\partial\nu} \ge 0$; moreover

$$-\Delta\varphi + \tau\varphi^q - C_0|\nabla\varphi|^{\frac{2q}{q+1}}$$
$$= -t(2\alpha+2)(2\alpha+3)d_n^{-2(\alpha+2)} + t(2\alpha+2)d_n^{-2\alpha-3}\Delta d + \tau t^q d_n^{-2q(\alpha+1)} - C_0 t^{\frac{2q}{q+1}} d_n^{-\frac{2q}{q+1}(2\alpha+3)},$$

and using (2.21), recalling once again that $q > \frac{2q}{q+1}$, we have, for t large,

$$-\Delta\varphi + \tau\varphi^q - C_0|\nabla\varphi|^{\frac{2q}{q+1}} - C_F d_n^{-2(\alpha+2)} \ge \frac{1}{2}t^q d_n^{-2(\alpha+2)} - C_F d_n^{-2(\alpha+2)} - C.$$

Thus there exists $t^* > 0$ such that the right-hand side of the previous inequality is positive and the maximum principle lets us conclude that $w_n \leq \varphi$ and finally (2.16). ◻

Our next step consists in a refinement of (2.16) which sounds like the asymptotic estimate (1.8) proved in [18], though here we need a uniform estimate with respect to n as well. We will use for this purpose a scaling and blow-up argument for which we need a suitable localization near the boundary. Let us denote $\Sigma = \partial\Omega$ and

$$\Omega_\rho = \{x \in \Omega \ : \ d(x) < \rho\} \ .$$

Given a generic point on $\partial\Omega$, let us denote by $(\sigma_1, \ldots, \sigma_{N-1})$ its coordinates in some local chart and by $\overline{x}(\sigma)$ its representation in Cartesian coordinates. It is well known (see, e.g., [12]) that if Ω is of class C^k, $k > 1$, then there exists a positive number ρ such that any point x in Ω_ρ can be represented as

$$x = \delta\hat{n} + \overline{x},$$

where $\overline{x}$ is the projection of x onto $\partial\Omega$, $\hat{n}$ is the inward normal computed at $\overline{x}$, and $\delta = d(x)$ =dist$(x, \partial\Omega)$. Since $\overline{x}$ and $\hat{n}$ are functions of σ, we have $x = \delta\hat{n}(\sigma) + \overline{x}(\sigma)$, and the mapping $x \to (\delta, \sigma)$ is a C^{k-1} diffeomorphism of Ω_ρ onto $(0, \rho) \times U$, where $U \subset \mathbb{R}^{N-1}$ is an open set. If we denote by $T(x)$ this mapping, in the new coordinates (δ, σ) the Laplace operator takes the form

$$\Delta_x = \frac{1}{|\det(DT^{-1})|} \ \mathrm{div}_{\ \delta,\sigma} \left(A(\delta, \sigma)\nabla_{\delta,\sigma}\right), \tag{2.22}$$

where $A(\delta, \sigma) = (a_{ij})$ is defined by $a_{ij} = |\det(DT^{-1})| \, \nabla T_i \cdot \nabla T_j$ (see, e.g., [9]). However, since $\nabla d(x) \cdot \nabla\sigma_j(x) = 0$ for any $j = 1, \ldots, N-1$ (the projection of x is invariant along the normal), then we have $a_{1i} = a_{i1} = 0$ for $i = 2, \ldots, N$, while $a_{11} = |\nabla d(x)|^2$ and $a_{ij} = \nabla\sigma_{i-1} \cdot \nabla\sigma_{j-1}$ for $i, j = 2, \ldots, N$. Let us note that A has regular coefficients since Ω is smooth, and moreover A is uniformly elliptic. In order to fix the ideas, we may consider that the curve Σ is locally the graph of a smooth function f and hence that $\overline{x}(\sigma) = (\sigma, f(\sigma))$, $f : U \subset \mathbb{R}^{N-1} \mapsto \mathbb{R}$. In that case one explicitly computes $|\det(DT^{-1})| = \sqrt{1 + |\nabla f|^2} + O(\delta)$, which shows that the map T and T^{-1} are nondegenerate and yields the ellipticity of A (in a neighborhood of the boundary).

In the next lemma, we are going to use such a local change of coordinates $x \mapsto (\delta, \sigma)$ in order to straighten the boundary: The main advantage is that the matrix $A(\delta, \sigma)$ has zeros in the first row and column but for the term a_{11}. The reader familiar with differential geometry will also recognize in (2.22) the form of the Laplacian in curvilinear coordinates, since it can be rewritten as

$$\Delta_x = \frac{1}{\sqrt{g}} \frac{\partial}{\partial\delta} \left(\sqrt{g} \frac{\partial}{\partial\delta}\right) + \frac{1}{\sqrt{g}} \sum_{i=1}^{N-1} \frac{\partial}{\partial\sigma_i} \left(\sqrt{g}\, g^{ij} \frac{\partial}{\partial\sigma_j}\right), \tag{2.23}$$

where the last part, when $\delta = 0$, is nothing but the Laplace–Beltrami operator. Let us recall (see also [12]) that (2.23) may be used to compute the mean curvature at a given point $x_0 \in \partial\Omega$ in terms of the distance function. Indeed, by choosing σ_i as the principal directions at x_0 (hence the matrix A is diagonal at x_0), one can verify that

$$\Delta d\,|_{\partial\Omega} = -(N-1)H(\sigma), \tag{2.24}$$

where we indicate by $H(\sigma)$ the mean curvature of Σ at σ.

LEMMA 2.7. *Let* $z_n \in C^2(\overline{\Omega})$ *be a solution of* (2.14) *such that*

$$\lim_{d_n(x)\to 0} z_n(x)d_n^{\alpha}(x) = 0. \tag{2.25}$$

Assume moreover that (2.15) *holds true. Then*

$$\lim_{d_n(x)\to 0} d_n^{\alpha+1}(x)|\nabla z_n(x)| = 0. \tag{2.26}$$

Proof. We argue by contradiction: Let us suppose that there exists $\overline{\varepsilon}$ and a sequence $x_j \in \Omega$ such that $d(x_j) + \frac{1}{n_j} < \frac{1}{j}$ and

$$d_{n_j}^{\alpha+1}(x_j)|\nabla z_{n_j}(x_j)| > \overline{\varepsilon}. \tag{2.27}$$

Let us denote $\delta_j = d(x_j)$; we can extract a subsequence j_k such that $x_{j_k} \to x_0 \in \partial\Omega$ and

$$\frac{1}{1+\delta_{j_k} n_{j_k}} \to l \in [0,1]. \tag{2.28}$$

We localize then around x_0 using the system of local coordinates introduced above: In a neighborhood J of x_0 (with $\{x_j\} \subset J\cap\Omega$) we define a map $T : J\cap\Omega \to (0,\rho_0)\times U$ for some $\rho_0 > 0$, $U \subset \mathbb{R}^{N-1}$ such that, if $y = (y_1, y') = T(x)$, we have $y_1 = d(x) \in (0,\rho_0)$ and $y' \in U$ denotes the local coordinates in $\mathbb{R}^{N-1}$ of the projection of x on the boundary. With respect to the above presentation, we changed only the notations, setting $y = (\delta,\sigma)$, $\delta = y_1$, $\sigma = y'$. Moreover we set $T(x_0) = (0, y'_0)$, and without loss of generality, we may assume that $U = B(y'_0, R)$ for some $R > 0$.

Define now $v_n(y) = z_n(x) = z_n(T^{-1}(y))$, so that $\nabla z_n(x) = DT(x)\nabla v_n(y)$. Then the map T changes (2.14) into

$$\begin{cases} -\dfrac{1}{|\det(T^{-1})|} \operatorname{div}(A\nabla v_n) + |DT\,\nabla v_n + \tilde{G}_n|^q + v_n = \tilde{F}_n & \text{in } (0,\rho_0)\times B(y'_0,R), \\ A\nabla v_n \cdot e_1 = 0 & \text{on } \{y_1 = 0\}\times B(y'_0,R), \end{cases} \tag{2.29}$$

where $A(y)$ is the matrix defined as in (2.22), $\tilde{F}_n(y) = F_n(x)$, and $\tilde{G}_n(y) = G_n(x)$. Clearly from (2.25) we know that

$$\lim_{d_n(y)\to 0} v_n(y)d_n^{\alpha}(y) = 0, \qquad \text{where } d_n(y) = y_1 + \frac{1}{n}. \tag{2.30}$$

Moreover we are in the hypotheses of Lemma 2.6 so that $|\nabla z_n| \le C d_n^{-\alpha-1}$, which in turn yields

$$|DT\,\nabla v_n(y)| \le C\left(y_1 + \frac{1}{n}\right)^{-\alpha-1}. \tag{2.31}$$

We will contradict (2.27) by proving that

$$\left(\delta_{j_k} + \frac{1}{n_{j_k}}\right)^{\alpha+1} \nabla v_{n_{j_k}}(y_{j_k}) \to 0, \tag{2.32}$$

where $y_{j_k} = T(x_{j_k}) = (\delta_{j_k}, y'_{j_k})$, which converges toward $(0, y'_0)$.

For simplicity, we denote henceforth $\delta_k = \delta_{j_k}$, $n_k = n_{j_k}$, and $y'_k = y'_{j_k}$. Let us introduce the rescaled variable $\xi = (\xi_1, \xi')$ defined by

$$y = \begin{cases} y_1 = \left(\delta_k + \frac{1}{n_k}\right)\xi_1, \\ y' = \left(\delta_k + \frac{1}{n_k}\right)\xi' + y'_k. \end{cases}$$

Then ξ belongs to a domain D_k such that $D_k(\xi) \subset \mathbb{R}^N_+$ and

$$\left\{\xi \in \mathbb{R}^N_+ \ : \ \xi = (\xi_1, \xi'),\, 0 < \xi_1 < \frac{\rho_0}{\delta_k + \frac{1}{n_k}},\, \xi' \in B\left(0, \frac{R}{2(\delta_k + \frac{1}{n_k})}\right)\right\} \subset D_k(\xi);$$

hence $D_k(\xi)$ tends to $\mathbb{R}^N_+$ as $k \to \infty$. Now we set

$$\begin{aligned} v_k(\xi) &= \left(\delta_k + \frac{1}{n_k}\right)^{\alpha} v_{n_k}(y) \\ &= \left(\delta_k + \frac{1}{n_k}\right)^{\alpha} v_{n_k}\left(\left(\delta_k + \frac{1}{n_k}\right)\xi_1, \left(\delta_k + \frac{1}{n_k}\right)\xi' + y'_k\right). \end{aligned} \tag{2.33}$$

Since, using assumption (2.15), (2.31), and that $q(\alpha+1) = \alpha+2$, we have

$$|DT\,\nabla v_n + \tilde{G}_n|^q + |v_n| + |\tilde{F}_n| \le C\left(y_1 + \frac{1}{n}\right)^{-\alpha-2},$$

by rescaling (2.29) we get that v_k solves an equation of the type

$$-\operatorname{div}\Big(A_k(\xi)\nabla v_k(\xi)\Big) = H_k(\xi) \quad \text{in} \quad \mathcal{D}_k(\xi),$$

where $A_k(\xi) = A(y)$ and H_k is a sequence satisfying, in D_k,

$$|H_k| \le C\left(\delta_k + \frac{1}{n_k}\right)^{\alpha+2}\left(y_1 + \frac{1}{n_k}\right)^{-\alpha-2} = C\left[\frac{1}{\xi_1 + \frac{1}{n_k\delta_k+1}}\right]^{\alpha+2}. \tag{2.34}$$

Note that, since

$$d_{n_k}(y) = y_1 + \frac{1}{n_k} = \left(\delta_k + \frac{1}{n_k}\right)\left[\xi_1 + \frac{1}{n_k\delta_k + 1}\right], \tag{2.35}$$

(2.28) implies

$$\frac{\left(\delta_k + \frac{1}{n_k}\right)}{d_{n_k}(y)} = \frac{1}{\xi_1 + \frac{1}{n_k\delta_k+1}} \to \frac{1}{\xi_1 + l}. \tag{2.36}$$

In particular, by (2.30) and (2.33), $v_k(\xi)$ converges to 0 in $C^0_{\mathrm{loc}}(\overline{\mathbb{R}^N_+})$ if $l > 0$ or in $C^0_{\mathrm{loc}}(\mathbb{R}^N_+)$ if $l = 0$.

Observe that the points $y_k = (\delta_k, y'_k)$ correspond to $\xi_k = (\frac{n_k\delta_k}{\delta_k n_k+1}, 0)$, so that $\xi_k \to (1-l, 0)$ and ξ_k reaches the boundary of $\mathbb{R}^N_+$ only if $l = 1$. We distinguish henceforth two cases: If $l \ne 1$, the sequence ξ_k is contained in a fixed compact $K \subset \mathbb{R}^N_+$, so we only need to reason away from $\xi_1 = 0$. Since from (2.34) we have that

H_k is locally uniformly bounded in $\mathbb{R}^N_+$, Theorem 9.11 in [12] allows us to obtain a local $W^{2,p}_{\mathrm{loc}}(\mathbb{R}^N_+)$ bound for any $p>1$ and consequently the $C^1_{\mathrm{loc}}(\mathbb{R}^N_+)$ compactness for $v_k(\xi)$. Then we deduce that $v_k(\xi)\to 0$ in $C^1_{\mathrm{loc}}(\mathbb{R}^N_+)$, and computing ∇v_k on ξ_k we get

$$\nabla v_k(\xi_k) = \left(\delta_k + \frac{1}{n_k}\right)^{\alpha+1} \nabla v_{n_k}(\delta_k, y'_k) \longrightarrow 0, \tag{2.37}$$

which is (2.32).

If $l=1$, the sequence $\xi_k = (\frac{n_k\delta_k}{\delta_k n_k+1}, 0) \to (0,0)$, and hence we need a C^1 compactness up to the boundary. This can be done, for instance, in the same spirit of [11], by extending through reflection the problem on the whole of $\mathbb{R}^N$. Thus set $v_k^*(\xi)$ as

$$v_k^*(\xi) = \begin{cases} v_k(\xi_1,\xi') & \text{if } \xi\in D_k, \\ v_k(-\xi_1,\xi') & \text{if } \xi\in D_k^*, \end{cases}$$

where $D_k^* = \{(\xi_1,\xi') : (-\xi_1,\xi')\in D_k\}$, and similarly the matrix $A_k^*(\xi)$ and the datum $H_k^*(\xi)$ as

$$A_k^*(\xi) = \begin{cases} A_k(\xi_1,\xi') & \text{if } \xi\in D_k, \\ A_k(-\xi_1,\xi') & \text{if } \xi\in D_k^*, \end{cases} \quad \text{and} \quad H_k^*(\xi) = \begin{cases} H_k(\xi_1,\xi') & \text{if } \xi\in D_k, \\ H_k(-\xi_1,\xi') & \text{if } \xi\in D_k^*. \end{cases}$$

Note that in the reflected problem the mixed terms $a_{1k}(\xi_1,\xi')$, with $k\neq 1$, change into $-a_{1k}(-\xi_1,\xi')$: However, our change of coordinates gives $a_{1k}=0$ for $k\neq 1$ and hence we really have $A_k^*(\xi_1,\xi') = A_k(-\xi_1,\xi')$ for $\xi_1<0$. This allows us to deduce that A_k^* is continuous across the hyperplane $\xi_1=0$, and actually A_k^* is Lipschitz continuous on $\mathbb{R}^N$ with the same Lipschitz constant as A_k.

Using the boundary condition

$$A\nabla v_n\cdot e_1 = 0 \quad \text{on } \{\xi_1=0\},$$

then $v_k^*(\xi)$ solves

$$-\operatorname{div}\,(A_k^*(\xi)\nabla v_k^*(\xi)) = H_k^*(\xi) \quad \text{in} \quad \overline{D_k\cup D_k^*}^{\,o}.$$

Clearly we have $\overline{D_k\cup D_k^*}\to\mathbb{R}^N$ as $k\to+\infty$. Moreover, since $l=1$, we deduce from (2.28) and (2.34) that H_k^* is locally uniformly bounded in $\mathbb{R}^N$, in particular in any compact neighborhood of $\xi=0$. Similarly (2.30) and (2.33) imply, using (2.36) with $l=1$, that $v_k\to 0$ in $C^0_{\mathrm{loc}}(\overline{\mathbb{R}^N_+})$, and hence v_k^* converges to zero uniformly on compact sets. Finally, by regularity of A_k and by definition of A_k^* we deduce that A_k^* is equi-Lipschitz continuous, so that Calderon–Zygmund regularity applies. Applying again Theorem 9.11 in [12] to v_k^*, we obtain a local $W^{2,p}_{\mathrm{loc}}(\mathbb{R}^N)$ bound for any $p>1$ and then the $C^1_{\mathrm{loc}}(\mathbb{R}^N)$ compactness for $v_k^*(\xi)$. Consequently $v_k(\xi)$ converges to 0 in $C^1_{\mathrm{loc}}(\overline{\mathbb{R}^N_+})$, and therefore $\nabla v_k(\xi_k)\to 0$, where $\xi_k = (\frac{n_k\delta_k}{\delta_k n_k+1}, 0)$. Thus we obtain (2.37), and hence (2.32) holds in any case. This contradicts (2.27). ◻

Collecting the results of Lemmas 2.2, 2.6, and 2.7, we have all of the ingredients for the proof of Proposition 2.1.

Proof of Proposition 2.1 Let us observe that since $S_n = C^* d_n^{-\alpha} + \psi(x) d_n^{1-\alpha}$, with $\psi(x)$ smooth,

$$|S_n|\le C d_n^{-\alpha}, \quad |\nabla S_n|\le C d_n^{-\alpha-1}, \quad |D^2 S_n|\le C d_n^{-\alpha-2}, \quad \text{and} \quad |D^3 S_n|\le C d_n^{-\alpha-3},$$

and moreover $z_n = u_n - S_n$ solves

$$\begin{cases} -\Delta z_n + z_n + |\nabla z_n + \nabla S_n|^q = f_n(x) + \Delta S_n - S_n & \text{in } \Omega, \\ \dfrac{\partial z_n}{\partial \nu} = 0 & \text{on } \partial\Omega. \end{cases}$$

We can first apply Lemma 2.2 to u_n, which implies (see (2.11) and the definition of S_n) that $z_n d_n^\alpha \to 0$ as $d_n \to 0$. Subsequently Lemmas 2.6 and 2.7 applied to z_n give that $|\nabla z_n| \le C d_n^{-\alpha-1}$ and $d_n^{\alpha+1} \nabla z_n \to 0$ as $d_n \to 0$. By the definition of S_n, we deduce (2.4) and (2.5). ∎

2.2. Bernstein-type estimates. Now we come to the proof of Theorem 1.3; to this purpose, we consider the solutions u_n of (2.3), and we look for Lipschitz estimates on the translated function

$$z_n = u_n - S_n.$$

The choice of the "corrector term" S_n will be explained below; unfortunately, this choice (uniquely determined) will be slightly different according to whether $\alpha = \frac{2-q}{q-1}$ belongs to $\mathbb{N}$ or not, so that we have to treat separately some details of the two situations. However, in any case the function S_n will satisfy the assumptions of Proposition 2.1, in particular $S_n = C^* d_n^{-\alpha} + O(d_n^{1-\alpha})$ and, what is more important, $\nabla S_n = -\alpha\, C^* d_n^{-\alpha-1} \nabla d + O(d_n^{-\alpha})$. Therefore, estimate (2.5) implies that $d_n^{\alpha+1} \nabla u_n - d_n^{\alpha+1} S_n \to 0$ as $d_n \to 0$, which means that $|\nabla z_n| = o(d_n^{-\alpha-1}) = o(|\nabla S_n|)$. This fact plays a crucial role in the estimates below, and this is why we needed the preliminary Proposition 2.1 (except for $q = 2$, i.e., $\alpha = 0$).

Proof of Theorem 1.3. *Case* 1: $\alpha \notin \mathbb{N}$. Let us set

$$S_n = \sum_{k=0}^{m} \sigma_k d_n^{k-\alpha},$$

where $m = [\alpha] + 1$, $\sigma_0 = C^*$ (defined in (2.2)), and σ_k, $k = 1, \dots, m$, are smooth functions we will fix later. Let also f_n be a smooth approximation of f. Then let u_n be the solution of (2.3), and define $z_n = u_n - S_n$ so that z_n satisfies

$$\frac{\partial z_n}{\partial \nu} = 0 \quad \text{on} \quad \partial\Omega \tag{2.38}$$

and the following equation:

$$-\Delta z_n + z_n + |\nabla z_n + \nabla S_n|^q - |\nabla S_n|^q = f_n - F_n \quad \text{in} \quad \Omega, \tag{2.39}$$

where

$$F_n = -\Delta S_n + S_n + |\nabla S_n|^q. \tag{2.40}$$

Applying Proposition 2.1, and rephrasing (2.4) and (2.5) in terms of z_n, we get

$$|\nabla z_n| \le C d_n^{-\alpha-1} \qquad \text{in } \Omega, \tag{2.41}$$

$$d_n^{\alpha+1} \nabla z_n \to 0 \quad \text{as } d_n \to 0. \tag{2.42}$$

Step 1. We apply Bernstein's method on $|\nabla z_n|^2$ weighted with a suitable power of d_n, so let us define

$$w_n = |\nabla z_n|^2 d_n^\beta,$$

where β is a positive number. First, by Lemma 2.4 and (2.38), we deduce the condition on the normal derivative of w_n on the boundary, namely,

$$\frac{\partial w_n}{\partial \nu} \leq 0 \quad \text{on} \quad \partial\Omega.$$

Moreover, computing the Laplacian of w_n we observe that it satisfies

$$\Delta w_n = -\beta(\beta+1) d_n^\beta \frac{|\nabla z_n|^2}{d_n^2} + \beta d_n^{\beta-1} \Delta d |\nabla z_n|^2 + 2\beta \frac{\nabla w_n \cdot \nabla d}{d_n} + d_n^\beta [\Delta |\nabla z_n|^2], \tag{2.43}$$

and

$$\Delta |\nabla z_n|^2 = 2\left(\nabla z_n \cdot \nabla \Delta z_n + |D^2 z_n|^2\right). \tag{2.44}$$

Using (2.39) we have

$$\nabla z_n \cdot \nabla \Delta z_n = |\nabla z_n|^2 + H \cdot \nabla z_n + \nabla F_n \cdot \nabla z_n - \nabla f_n \cdot \nabla z_n, \tag{2.45}$$

where, recalling that $q(\alpha+1) = \alpha+2$,

$$H = \nabla(|\nabla z_n + \nabla S_n|^q - |\nabla S_n|^q) = \nabla\left(\frac{|d_n^{\alpha+1}\nabla z_n + d_n^{\alpha+1}\nabla S_n|^q - |d_n^{\alpha+1}\nabla S_n|^q}{d_n^{\alpha+2}}\right).$$

Now let us compute

$$\nabla S_n = \sum_{k=0}^m (k-\alpha)\sigma_k d_n^{k-\alpha-1}\nabla d + d_n^{k-\alpha}\nabla\sigma_k,$$

and we set

$$T = d_n^{\alpha+1}\nabla S_n = \sum_{k=0}^m (k-\alpha)\sigma_k d_n^k \nabla d + d_n^{k+1}\nabla\sigma_k; \qquad \xi = d_n^{\alpha+1}\nabla z_n + T.$$

Note that, since $\sigma_0 = C^*$, T can be written more explicitly as

$$T = -\alpha C^* \nabla d + \sum_{k=1}^m (k-\alpha)\sigma_k d_n^k \nabla d + d_n^{k+1}\nabla\sigma_k = -\alpha C^*\nabla d + O(d_n). \tag{2.46}$$

In particular both $|T|$ and $|\xi|$ (due to (2.42)) do not vanish in a suitable neighborhood of $\partial\Omega$, and we have

$$|T|^{q-2} = (\alpha C^*)^{q-2} + o(1) \quad \text{and} \quad |\xi|^{q-2} = (\alpha C^*)^{q-2} + o(1). \tag{2.47}$$

We compute then

$$H \cdot \nabla z_n = \nabla\left(\frac{|\xi|^q - |T|^q}{d_n^{\alpha+2}}\right) \cdot \nabla z_n =$$

$$-(\nabla d \cdot \nabla z_n)\frac{(\alpha+2)}{d_n^{\alpha+3}}[|\xi|^q - |T|^q] + \frac{q}{d_n^{\alpha+2}}|\xi|^{q-2} D\xi\,\xi\,\nabla z_n - \frac{q}{d_n^{\alpha+2}}|T|^{q-2} DT\,T\,\nabla z_n\,;$$

since by the definition of ξ (we use now the notation $(a\otimes b)cd=(a\cdot c)(b\cdot d)$ for vector fields)

$$|\xi|^{q-2}D\xi\xi\nabla z_n = |\xi|^{q-2}\Big[(\alpha+1)d_n^{\alpha}\nabla z_n\otimes\nabla d + d_n^{\alpha+1}D^2 z_n\Big]\xi\nabla z_n \\ + |\xi|^{q-2}DT(d_n^{\alpha+1}\nabla z_n + T)\nabla z_n,$$

we obtain

$$H\cdot\nabla z_n = -(\nabla d\cdot\nabla z_n)\frac{(\alpha+2)}{d_n^{\alpha+3}}[|\xi|^q - |T|^q] \\ + \frac{q|\xi|^{q-2}}{d_n^{\alpha+2}}\left[(\alpha+1)(\nabla d\cdot\nabla z_n)(\nabla z_n\cdot\xi)d_n^{\alpha} + \frac{d_n^{\alpha+1}}{2}\nabla|\nabla z_n|^2\cdot\xi\right] \\ + \frac{q|\xi|^{q-2}DT\nabla z_n\nabla z_n}{d_n} + \frac{q}{d_n^{\alpha+2}}\left[|\xi|^{q-2} - |T|^{q-2}\right]DT\,T\,\nabla z_n.$$

Let us observe that

$$\Big||\xi|^q - |T|^q - q d_n^{\alpha+1}|T|^{q-2}\nabla z_n\cdot T\Big| \le C|\nabla z_n|^2 d_n^{2\alpha+2}$$

and

$$\Big||\xi|^{q-2} - |T|^{q-2}\Big| \le C|\nabla z_n| d_n^{\alpha+1}, \tag{2.48}$$

so that

$$H\cdot\nabla z_n \ge -q\frac{(\alpha+2)}{d_n^2}(\nabla d\cdot\nabla z_n)(\nabla z_n\cdot T)|T|^{q-2} - C|\nabla d\cdot\nabla z_n|\frac{|\nabla z_n|^2 d_n^{2\alpha+2}}{d_n^{\alpha+3}} \\ + \frac{q(\alpha+1)}{d_n^2}(\nabla d\cdot\nabla z_n)(\nabla z_n\cdot\xi)|T|^{q-2} - \frac{C}{d_n^2}|\nabla d\cdot\nabla z_n|\,|\nabla z_n\cdot\xi||\nabla z_n|d_n^{\alpha+1} \\ + \frac{q|\xi|^{q-2}}{2d_n}\nabla|\nabla z_n|^2\cdot\xi + \frac{q|\xi|^{q-2}DT\nabla z_n\nabla z_n}{d_n} - C\frac{|\nabla z_n|^2}{d_n}.$$

In the third term on the right, we write that $\xi = T + d_n^{\alpha+1}\nabla z_n$. Then, using (2.42) and (2.47), we can simplify the previous inequality by writing

$$H\cdot\nabla z_n \ge -\frac{q}{d_n^2}(\nabla d\cdot\nabla z_n)(\nabla z_n\cdot T)|T|^{q-2} \\ + \frac{q|\xi|^{q-2}}{2d_n}\nabla|\nabla z_n|^2\cdot\xi - \frac{|\nabla z_n|^2}{d_n^2}\omega_n,$$

where (here and later) ω_n is some function satisfying

$$|\omega_n| \le C(d_n + |\nabla z_n| d_n^{\alpha+1}). \tag{2.49}$$

Note that (2.42) implies $\omega_n\to 0$ as $d_n\to 0$. Now we use (2.46), (2.47), and the algebraic relations in (2.2) (i.e., $(C^*\alpha)^{q-1}=\alpha+1$, $q(\alpha+1)=\alpha+2$), and we deduce

$$H\cdot\nabla z_n \ge (\alpha+2)\left(\frac{\nabla d\cdot\nabla z_n}{d_n}\right)^2 + \frac{q|\xi|^{q-2}}{2d_n}\nabla|\nabla z_n|^2\cdot\xi - \frac{|\nabla z_n|^2}{d_n^2}\omega_n.$$

Substituting the previous inequality into (2.45) we finally get

$$\begin{aligned}\nabla z_n \cdot \nabla \Delta z_n &\geq |\nabla z_n|^2 + \nabla F_n \cdot \nabla z_n - \nabla f_n \cdot \nabla z_n \\ &+ (\alpha+2)\left(\frac{\nabla d \cdot \nabla z_n}{d_n}\right)^2 + \frac{q|\xi|^{q-2}}{2d_n}\nabla|\nabla z_n|^2 \cdot \xi - \frac{|\nabla z_n|^2}{d_n^2}\omega_n.\end{aligned} \tag{2.50}$$

Moreover since

$$\nabla|\nabla z_n|^2 = \frac{\nabla w_n}{d_n^\beta} - \beta\frac{|\nabla z_n|^2}{d_n}\nabla d \tag{2.51}$$

we have, recalling that $\xi = d_n^{\alpha+1}\nabla z_n + T$,

$$\begin{aligned}\frac{q|\xi|^{q-2}}{2d_n}\nabla|\nabla z_n|^2 \cdot \xi &= -\beta\frac{q|\xi|^{q-2}}{2d_n^2}|\nabla z_n|^2\nabla d \cdot T \\ -\beta\frac{q|\xi|^{q-2}}{2}\frac{|\nabla z_n|^2}{d_n^2}\nabla d \cdot \nabla z_n\, d_n^{\alpha+1} &+ \frac{q|\xi|^{q-2}}{2}d_n^{-\beta-1}\nabla w_n \cdot \xi.\end{aligned}$$

Then using (2.42) we get from (2.50)

$$\begin{aligned}\nabla\Delta z_n \cdot \nabla z_n &\geq |\nabla z_n|^2 + \nabla F_n \cdot \nabla z_n - \nabla f_n \cdot \nabla z_n + (\alpha+2)\left(\frac{\nabla d \cdot \nabla z_n}{d_n}\right)^2 \\ &-\beta\frac{q|\xi|^{q-2}}{2d_n^2}|\nabla z_n|^2\nabla d \cdot T + \frac{q|\xi|^{q-2}}{2}d_n^{-\beta-1}\nabla w_n \cdot \xi - \frac{|\nabla z_n|^2}{d_n^2}\omega_n.\end{aligned} \tag{2.52}$$

Here we use again (2.46), (2.47), and (2.2), so that

$$-\beta\frac{q|\xi|^{q-2}}{2d_n^2}|\nabla z_n|^2\nabla d \cdot T = \beta\frac{(\alpha+2)}{2d_n^2}|\nabla z_n|^2[1+o(1)].$$

We conclude then from (2.52), since $\omega_n = o(1)$ as $d_n \to 0$,

$$\begin{aligned}\nabla\Delta z_n \cdot \nabla z_n &\geq |\nabla z_n|^2 + \nabla F_n \cdot \nabla z_n - \nabla f_n \cdot \nabla z_n \\ +(\alpha+2)\left(\frac{\nabla d \cdot \nabla z_n}{d_n}\right)^2 &+ \beta\frac{(\alpha+2)}{2d_n^2}|\nabla z_n|^2[1+o(1)] + \frac{q|\xi|^{q-2}}{2}d_n^{-\beta-1}\nabla w_n \cdot \xi.\end{aligned} \tag{2.53}$$

Using the above inequality in (2.43) and (2.44), we finally obtain

$$\begin{aligned}-\Delta w_n + 2w_n + \beta(\alpha+1-\beta)d_n^\beta\frac{|\nabla z_n|^2}{d_n^2}(1+o(1)) + 2(\alpha+2)\left(\frac{\nabla d \cdot \nabla z_n}{d_n}\right)^2 d_n^\beta \\ + 2\beta\frac{\nabla w_n \cdot \nabla d}{d_n} + \frac{q|\xi|^{q-2}}{d_n}\nabla w_n \cdot \xi \\ + 2d_n^\beta\nabla F_n \cdot \nabla z_n - 2d_n^\beta\nabla f_n \cdot \nabla z_n + 2d_n^\beta|D^2 z_n|^2 \leq 0.\end{aligned} \tag{2.54}$$

Now let us write explicitly the term F_n defined in (2.40), i.e.,

$$\begin{aligned}F_n &= -\Delta S_n + S_n + |\nabla S_n|^q \\ = -\sum_{k=0}^m (k-\alpha)&\left[(k-\alpha-1)\sigma_k d_n^{k-\alpha-2} + (\Delta d + 2\nabla\sigma_k \cdot \nabla d)d_n^{k-\alpha-1}\right] + \Delta\sigma_k d_n^{k-\alpha} \\ &+ \sum_{k=0}^m \sigma_k d_n^{k-\alpha} + \frac{1}{d_n^{\alpha+2}}\left|\sum_{k=0}^m (k-\alpha)\sigma_k d_n^k\nabla d + d_n^{k+1}\nabla\sigma_k\right|^q.\end{aligned} \tag{2.55}$$

Recall that $\sigma_0 = C^*$ is a constant: Then we can write last term as

$$\frac{1}{d_n^{\alpha+2}}\left|\sum_{k=0}^{m}(k-\alpha)\sigma_k d_n^k \nabla d + d_n^{k+1}\nabla\sigma_k\right|^q = \sum_{k=0}^{m} l_k d_n^{k-\alpha-2} + O(d_n^{m-\alpha-1}),$$

where $l_0 = (\alpha\sigma_0)^q$ and l_k, $k = 1, \dots, m$ are coefficients that depend on $\sigma_0, \dots, \sigma_k$, precisely

$$l_k = -q(k-\alpha)(\alpha\sigma_0)^{q-1}\sigma_k + \tilde{l}_k(\sigma_0, \dots, \sigma_{k-1}), \quad 1 \le k \le [\alpha]+1.$$

Hence, since $q(\alpha\sigma_0)^{q-1} = \alpha + 2$ (see (2.2)) and $m = [\alpha]+1$, we have from (2.55)

$$F_n = \sum_{k=1}^{m}\left[(\alpha-k)(k+1)\sigma_k + \psi_k(\sigma_0, \dots, \sigma_{k-1})\right] d_n^{k-\alpha-2} + O(d_n^{[\alpha]-\alpha})$$

for some functions ψ_k. Then by induction we can choose $\sigma_k = -\frac{\psi_k(\sigma_0,\dots,\sigma_{k-1})}{(\alpha-k)(k+1)}$ in order to have

$$|F_n| = \left| -\Delta S_n + S_n + |\nabla S_n|^q\right| = O\left(d_n^{[\alpha]-\alpha}\right)$$

and

$$|\nabla F_n \cdot \nabla z_n| \le C d_n^{[\alpha]-\alpha-1}\left(|\nabla d \cdot \nabla z_n| + d_n\,|\nabla z_n|\right).$$

Consequently, by Young's inequality we have

$$|\nabla F_n \cdot \nabla z_n| \le \frac{(\alpha+2)}{2}\frac{|\nabla d \cdot \nabla z_n|^2}{d_n^2} + C\left(d_n^{2([\alpha]-\alpha)} + |\nabla z_n|^2\right). \tag{2.56}$$

Thus in a suitable neighborhood Ω_{δ_0} of $\partial\Omega$, we have that w_n solves

$$\begin{aligned} &-\Delta w_n + 2w_n + \beta(\alpha+1-\beta)\,\frac{w_n}{d_n^2}\,(1+o(1)) + (\alpha+2)\left(\frac{\nabla d\cdot\nabla z_n}{d_n}\right)^2 d_n^\beta \\ &\quad + \frac{q|\xi|^{q-2}}{d_n}\nabla w_n\cdot\xi + 2\beta\frac{\nabla w_n\cdot\nabla d}{d_n} - 2d_n^\beta\nabla f_n\cdot\nabla z_n \le C\,d_n^{\beta+2([\alpha]-\alpha)}. \end{aligned} \tag{2.57}$$

By Young's inequality, we have for all $\varepsilon > 0$

$$2d_n^\beta|\nabla f_n\cdot\nabla z_n| \le \varepsilon\frac{w_n}{d_n^2} + C_\varepsilon d_n^{2+\beta}\,|\nabla f_n|^2 \le \varepsilon\frac{w_n}{d_n^2} + C_\varepsilon. \tag{2.58}$$

Here we set $\beta < \alpha+1$ and δ_0, n_0 such that in Ω_{δ_0} and for $n > n_0$ we have $\beta(\alpha+1-\beta)\,(1+o(1)) \ge c > 0$: Hence, fixing ε small enough and dropping positive terms, we get for some constants c_0, $C_0 > 0$

(2.59)
$$-\Delta w_n + 2w_n + c_0\frac{w_n}{d_n^2} + 2\beta\frac{\nabla w_n\cdot\nabla d}{d_n} + \frac{q|\xi|^{q-2}}{d_n}\nabla w_n\cdot\xi \le C_0\left(1 + \frac{d_n^\beta}{d_n^{2(\alpha-[\alpha])}}\right) \quad \text{in} \quad \Omega_{\delta_0},$$

with the condition

$$\frac{\partial w_n}{\partial\nu} \le 0 \quad \text{on} \quad \partial\Omega.$$

Since

$$C_0\left(1+\frac{d_n^\beta}{d_n^{2(\alpha-[\alpha])}}\right)\le\frac{C_1}{d_n^2},$$

the maximum principle implies that

$$\max_{\Omega_{\delta_0}} w_n\le\frac{C_1}{c_0}+\max_{\partial\Omega_{\delta_0}\setminus\partial\Omega} w_n,$$

and last term is uniformly bounded thanks to (2.41): Then,

$$\forall\, 0<\beta<\alpha+1,\quad |\nabla z_n|^2\le C d_n^{-\beta}\qquad\text{in}\quad\Omega. \tag{2.60}$$

Step 2. In order to get a uniform $L^\infty(\Omega)$ estimate on ∇z_n, now we deal with

$$w_n=e^{\theta(d_n)}|\nabla z_n|^2, \tag{2.61}$$

where $\theta(s)$ is a positive increasing function such that $\theta(0)=0$. Then w_n satisfies in Ω

$$\begin{aligned}&-\Delta w_n+\left[\theta''(d_n)+\theta'(d_n)^2+\theta'(d_n)\Delta d\right]w_n\\&+2e^{\theta(d_n)}\theta'(d_n)\nabla|\nabla z_n|^2\cdot\nabla d+2e^{\theta(d_n)}(|D^2z_n|^2+\nabla z_n\cdot\nabla(\Delta z_n))=0.\end{aligned} \tag{2.62}$$

Following previous computations we obtain (2.50); then since

$$\nabla|\nabla z_n|^2=\nabla w_n e^{-\theta(d_n)}-\theta'(d_n)|\nabla z_n|^2\nabla d \tag{2.63}$$

we get

$$\begin{aligned}\nabla z_n\cdot\nabla\Delta z_n&\ge|\nabla z_n|^2+\nabla F_n\cdot\nabla z_n-\nabla f_n\cdot\nabla z_n+(\alpha+2)\left(\frac{\nabla d\cdot\nabla z_n}{d_n}\right)^2\\&+\frac{q|\xi|^{q-2}}{2d_n}e^{-\theta(d_n)}\nabla w_n\cdot\xi-\theta'(d_n)\frac{q|\xi|^{q-2}}{2d_n}|\nabla z_n|^2\nabla d\cdot\xi-\frac{|\nabla z_n|^2}{d_n^2}\omega_n.\end{aligned}$$

Here recall that $\xi=d_n^{\alpha+1}\nabla z_n+T$; we use (2.46), (2.47), and the relations between q, α, C^* in (2.2), and we obtain, similarly as we did for (2.53), that

$$\begin{aligned}\nabla\Delta z_n\cdot\nabla z_n&\ge|\nabla z_n|^2+\nabla F_n\cdot\nabla z_n-\nabla f_n\cdot\nabla z_n+\frac{\alpha+2}{d_n^2}(\nabla d\cdot\nabla z_n)^2\\&+\frac{q|\xi|^{q-2}}{2d_n}e^{-\theta(d_n)}\nabla w_n\cdot\xi+\frac{(\alpha+2)\theta'(d_n)}{2d_n}|\nabla z_n|^2[1+o(1)]-\omega_n\frac{|\nabla z_n|^2}{d_n^2},\end{aligned}$$

where ω_n satisfies (2.49). Choosing $\sigma_1,\dots,\sigma_m$ as before we have

$$\nabla F_n\cdot\nabla z_n\ge-\frac{\alpha+2}{2}\frac{(\nabla z_n\cdot\nabla d)^2}{d_n^2}-C\left[d_n^{2([\alpha]-\alpha)}+|\nabla z_n|^2\right], \tag{2.64}$$

so we conclude that

$$\begin{aligned}\nabla\Delta z_n\cdot\nabla z_n&\ge|\nabla z_n|^2-\nabla f_n\cdot\nabla z_n+\frac{\alpha+2}{2d_n^2}(\nabla d\cdot\nabla z_n)^2\\&+\frac{q|\xi|^{q-2}}{2d_n}e^{-\theta(d_n)}\nabla w_n\cdot\xi+\frac{(\alpha+2)\theta'(d_n)}{2d_n}|\nabla z_n|^2[1+o(1)]-\omega_n\frac{|\nabla z_n|^2}{d_n^2}-Cd_n^{2([\alpha]-\alpha)}.\end{aligned}$$

Substituting the above inequality into (2.62) and using (2.63), we find

$$\begin{aligned}
&-\Delta w_n + 2w_n + \left[\frac{(\alpha+2)\theta'(d_n)}{d_n} + \theta''(d_n) - \theta'(d_n)^2 + \theta'(d_n)\Delta d - \frac{2\omega_n}{d_n^2}\right] w_n \\
&\quad + \frac{\alpha+2}{d_n^2}(\nabla d\cdot\nabla z_n)^2 e^{\theta(d_n)} + \frac{q|\xi|^{q-2}}{d_n}\nabla w_n\cdot\xi + 2\theta'(d_n)\nabla w_n\cdot\nabla d \\
&\quad + 2e^{\theta(d_n)}|D^2 z_n|^2 \le 2e^{\theta(d_n)}[\nabla f_n\cdot\nabla z_n + C d_n^{2([\alpha]-\alpha)}].
\end{aligned} \tag{2.65}$$

Several choices of functions $\theta(s)$ are now possible (see also the results in section 3): For instance, if we take $\theta(s) = s^\eta$, $\eta > 0$, we get

$$\frac{(\alpha+2)\theta'(d_n)}{d_n} + \theta''(d_n) - \theta'(d_n)^2 + \theta'(d_n)\Delta d - \frac{2\omega_n}{d_n^2} \ge \frac{\eta(\alpha+1+\eta)}{d_n^{2-\eta}}(1+o(1)) - \frac{2\omega_n}{d_n^2}.$$

However, by the definition of ω_n in (2.49), using now (2.60) with $\beta \le \alpha$ implies that

$$\omega_n \le C(d_n + |\nabla z_n| d_n^{\alpha+1}) \le C\, d_n, \tag{2.66}$$

and hence, for any $\eta < 1$,

$$\left[\frac{(\alpha+2)\theta'(d_n)}{d_n} + \theta''(d_n) - \theta'(d_n)^2 + \theta'(d_n)\Delta d - \frac{2\omega_n}{d_n^2}\right] \ge \frac{\eta(\alpha+1+\eta)}{d_n^{2-\eta}}(1+o(1)).$$

Using Young's inequality to take care of the term with f_n, we deduce from (2.65)

$$\begin{aligned}
&-\Delta w_n + 2w_n + \frac{\eta(\alpha+1+\eta)}{d_n^{2-\eta}}(1+o(1))w_n + \frac{\alpha+2}{d_n^2}(\nabla d\cdot\nabla z_n)^2 e^{\theta(d_n)} \\
&\quad + \frac{q|\xi|^{q-2}}{d_n}\nabla w_n\cdot\xi + 2\theta'(d_n)\nabla w_n\cdot\nabla d \le C[1 + d_n^{2([\alpha]-\alpha)}].
\end{aligned} \tag{2.67}$$

Observe that for any $\eta > 0$ the function e^{s^η} satisfies Lemma 2.4, and hence

$$\frac{\partial w_n}{\partial \nu} \le 0 \quad \text{on} \quad \partial\Omega.$$

Therefore, by applying the maximum principle we easily conclude from (2.67) as soon as $\eta < 1 + [\alpha] - \alpha$:

$$|\nabla z_n| \le C\left[1 + \max_{\partial\Omega_{\delta_0}\setminus\partial\Omega} |\nabla z_n|\right] \quad \text{in} \quad \Omega_{\delta_0},$$

and the last term is uniformly bounded by (2.41). This proves a uniform Lipschitz estimate for z_n (recall that z_n is locally uniformly bounded). Then standard results imply that z_n converges to a function z such that $z+S$ is a solution of (1.1)–(1.2), and hence the uniqueness result implies that $z = u-S$. We conclude that $u-S$ is Lipschitz, and this regularity, decomposed on normal and tangentials directions, yields (1.16).

Case 2: $\alpha \in \mathbb{N}$, $\alpha \ne 0$. The only difference between the two cases lies in the formal expansion of u_n. Indeed we have to modify S_n as follows:

$$S_n = \sum_{k=0}^{\alpha-1} \sigma_k(x) d_n^{k-\alpha} + \sigma_\alpha(x)\log d_n + \sigma_{\alpha+1}(x) d_n \log d_n + \tilde\sigma_{\alpha+1}(x) d_n, \tag{2.68}$$

where again $\sigma_0 = C^*$ and $\tilde\sigma_{\alpha+1}$, σ_i, $i = 0, \ldots, \alpha+1$, are smooth functions to be fixed. Hence now we have (after standard computations)

$$T = d_n^{\alpha+1}\nabla S_n = -\alpha C^*\nabla d + \sum_{k=1}^{\alpha-1} d_n^k[(k-\alpha)\sigma_k\nabla d + \nabla\sigma_{k-1}] + d_n^\alpha(\sigma_\alpha\nabla d + \nabla\sigma_{\alpha-1}) \tag{2.69}$$

$$+\, d_n^{\alpha+1}\left[\log d_n(\nabla\sigma_\alpha + \sigma_{\alpha+1}\nabla d) + (\sigma_{\alpha+1} + \tilde\sigma_{\alpha+1})\nabla d + d_n \log d_n \nabla\sigma_{\alpha+1} + d_n\nabla\tilde\sigma_{\alpha+1}\right].$$

Then, as in (2.46), we have that $T = -\alpha\, C^*\nabla d + O(d_n)$ and (2.47) still holds, so we can argue as above until we obtain (2.54). The choice of S_n is involved in the regularity of F_n defined in (2.40): In fact we find

$$\begin{aligned} &-\Delta S_n + S_n + |\nabla S_n|^q \\ &= -\sum_{k=0}^{\alpha-1}(k-\alpha)\left[(k-\alpha-1)\sigma_k d_n^{k-\alpha-2} + (\Delta d + 2\nabla\sigma_k\cdot\nabla d)d_n^{k-\alpha-1}\right] + \Delta\sigma_k d_n^{k-\alpha} \\ &\quad + \sigma_\alpha\, d_n^{-2} - [\sigma_\alpha\Delta d + \nabla\sigma_\alpha\nabla d + \sigma_{\alpha+1}]\, d_n^{-1} + \tau(x)\log d_n \\ &\quad + \sum_{k=0}^{\alpha-1}\sigma_k d_n^{k-\alpha} + \frac{1}{d_n^{\alpha+2}}\left|d_n^{\alpha+1}\nabla S_n\right|^q, \end{aligned}$$

where τ is smooth (say, $W^{1,\infty}(\Omega)$). From the expression in (2.69), we notice that by choosing $\sigma_{\alpha+1} = -\nabla\sigma_\alpha\cdot\nabla d$ we can write the last term as

$$\frac{1}{d_n^{\alpha+2}}\left|d_n^{\alpha+1}\nabla S_n\right|^q = \sum_{k=0}^{\alpha+1} l_k d_n^{k-\alpha-2} + \tau_1(x)\log d_n,$$

where τ_1 is smooth, $l_0 = (\alpha C^*)^q$ and the l_k, $k = 1, \ldots, \alpha+1$, are coefficients that depend on $\sigma_0, \ldots, \sigma_k$ as follows:

$$\begin{aligned} l_k &= -q(k-\alpha)(\alpha C^*)^{q-1}\sigma_k + \tilde l(\sigma_0, \ldots, \sigma_{k-1}), \quad 1 \le k \le \alpha-1, \\ l_\alpha &= q(\alpha C^*)^{q-1}\sigma_\alpha + \tilde l(\sigma_0, \ldots, \sigma_{\alpha-1}), \\ l_{\alpha+1} &= q(\alpha C^*)^{q-1}(\sigma_{\alpha+1} + \tilde\sigma_{\alpha+1}) + \tilde l(\sigma_0, \ldots, \sigma_\alpha). \end{aligned}$$

Hence, using $q(\alpha C^*)^{q-1} = \alpha + 2$, as in Case 1 we can choose by induction σ_k, $k = 1, \ldots, \alpha$ (which are the same as in the previous case), and $\tilde\sigma_{\alpha+1}$ in a way that

$$|F_n| = O(|\log d_n|)$$

and

$$|\nabla F_n\cdot\nabla z_n| \le C d_n^{-1}\left(|\nabla z_n\cdot\nabla d| + d_n|\log d_n||\nabla z_n|\right).$$

Then, by Young's inequality,

$$|\nabla F_n\cdot\nabla z_n| \le \frac{(\alpha+2)}{2}\left(\frac{|\nabla z_n\cdot\nabla d|}{d_n}\right)^2 + C(|\log d_n|^2 + |\nabla z_n|^2), \tag{2.70}$$

which replaces (2.56). We conclude as in the previous case by getting estimate (2.60), and then, using again (2.70), as in Step 2 we obtain the Lipschitz estimate and the claimed regularity for $u - S$. We deduce the asymptotic development (1.17) (using again that $\sigma_{\alpha+1} = -\nabla\sigma_\alpha \cdot \nabla d$).

Case 3: $\alpha = 0$. We recall that $\alpha = 0$ means $q = 2$: This case in fact is much simpler than the previous ones, since (2.39) reads as

$$-\Delta z_n + z_n + |\nabla z_n|^2 + 2\nabla z_n \cdot \nabla S_n = f_n - F_n$$

and there is no need to use Taylor's expansion of the nonlinearity (this is why Lemma 2.7 is not needed). In particular, the computation of $\nabla z_n \nabla \Delta z_n$ is straightforward, and one can easily follow the previous steps. As far as the choice of S_n is concerned, this is exactly as in (2.68), which here assumes the simple form $S_n = -\log(d_n) + \frac{\Delta d}{2}\, d_n$. ∎

We observe that the computation of the functions σ_k in the previous theorem can be done explicitly. In particular, it turns out that

$$\sigma_1(x) = \frac{(q-1)^{-\frac{1}{q-1}}}{1-\alpha}\,\frac{\Delta d(x)}{2} \quad \text{if } \alpha \neq 1, \ \ \sigma_1(x) = \frac{(q-1)^{-\frac{1}{q-1}}\Delta d(x)}{2} \quad \text{if } \alpha = 1,$$

and hence, recalling (2.24),

$$\begin{aligned} \sigma_1(x) &= \frac{(q-1)^{-\frac{1}{q-1}}}{1-\alpha}\,\frac{(N-1)}{2} H(\overline{x}) \quad \text{if } \alpha \neq 1, \\ \sigma_1(x) &= \frac{(q-1)^{-\frac{1}{q-1}}(N-1)}{2} H(\overline{x}) \quad \text{if } \alpha = 1. \end{aligned} \tag{2.71}$$

Remark 2.8. Nothing really changes in the proof of Theorem 1.3 according to the value of λ appearing in (1.1), but for the possibly different values of the coefficients σ_k defining the corrector term S. However, the first two terms $\sigma_0 = C^*$ and σ_1 defined in (2.71) would remain the same for solutions of (1.1), independently of the value of λ, and so the conclusions of Theorems 1.2 and 1.1. In particular, let us stress that the gradient estimates remain valid for the solution of the ergodic problem, which is well defined from the results in [15].

2.3. The case $q = 2$ and Theorem 1.3 refined. Let us complete here the proof of Theorems 1.2 and 1.1 by getting an improved asymptotic estimate for the case $q = 2$.

PROPOSITION 2.9. *Let $\Omega \in C^5$ and f be in $W^{1,\infty}(\Omega)$. Let $u(x)$ be the unique solution of*

$$\begin{cases} -\Delta u + u + |\nabla u|^2 = f(x) & \textit{in} \quad \Omega, \\ u(x) \to +\infty & \textit{as} \quad d(x) \to 0. \end{cases} \tag{2.72}$$

Then

$$\lim_{d(x)\to 0} \frac{\partial u(x)}{\partial \nu} - \left(\frac{1}{d(x)} - \frac{\Delta d(x)}{2}\right) = 0. \tag{2.73}$$

Proof. Let us set $S = -\log d(x) + \frac{\Delta d}{2} d(x)$ and $z = u - S$. Applying Theorem 1.3 we have that z is Lipschitz continuous, which implies

$$|\nabla(u(x) + \log d(x))| \leq C \quad \text{in} \quad \Omega. \tag{2.74}$$

From (2.74), $u+\log d$ can be extended continuously to $\partial\Omega$ and for all $x_0\in\partial\Omega$,

$$|u(x)+\log d(x)-\rho(x_0)|\le C|x-x_0|, \tag{2.75}$$

where $\rho(x_0)=(u+\log d)|_{x_0}$. Let us consider now a point $P_0\in\Omega$ close to the boundary such that its projection on $\partial\Omega$ coincides with x_0, and let us introduce in $B(P_0,d_0)$, $d_0=|x_0-P_0|$, a new system of coordinates $(\eta_1,\dots,\eta_N)=(\eta_1,\eta')$ such that the origin O coincides with x_0 and the η_1-axis coincides with $\nabla d(x_0)=-\nu(x_0)$. Since the equation is invariant with respect to rotations and translations, $u(\eta)$ still solves (2.72), and estimate (2.75) holds true. We introduce the rescaled variable $\xi=\frac{\eta}{\delta}$, i.e.,

$$\xi_1=\frac{\eta_1}{\delta},\qquad \xi'=\frac{\eta'}{\delta}\quad \forall\delta>0,$$

and we define

$$v_\delta(\xi)=\frac{u(\delta\xi)+\log d(\delta\xi)-\frac{\Delta d(x_0)d(\delta\xi)}{2}-\rho(x_0)}{\delta}. \tag{2.76}$$

Hence $v_\delta(\xi)$ solves the following equation:

$$\begin{gathered}-\Delta v_\delta+\delta^2 v_\delta-2\delta\frac{\nabla v_\delta\cdot\nabla d(\delta\xi)}{d(\delta\xi)}+\delta|\nabla v_\delta|^2+\delta\Delta d(x_0)\nabla v_\delta\cdot\nabla d(\delta\xi)\\ +\frac{\delta}{d(\delta\xi)}(\Delta d(\delta\xi)-\Delta d(x_0))+\delta\left(\frac{\Delta d(x_0)}{2}\right)^2-\delta\frac{\Delta d(x_0)\Delta d(\delta\xi)}{2}-\delta\log d(\delta\xi)\\ +\delta\frac{\Delta d(x_0)}{2}d(\delta\xi)+\delta\rho(x_0)-\delta f(\delta\xi)=0,\end{gathered} \tag{2.77}$$

in the domain D_δ, where

$$D_\delta=\left\{(\xi_1,\xi')\,:\,\left(\xi_1-\frac{d_0}{\delta}\right)^2+|\xi'|^2\le\left(\frac{d_0}{\delta}\right)^2\right\}.$$

Note that $D_\delta\to\mathbb{R}^N_+$ as $\delta\to 0$. Moreover, from (2.74) and (2.75), we have $|v_\delta|\le C|\xi|$ and $|\nabla v_\delta|\le C$, and hence v_δ is bounded in $W^{1,\infty}_{loc}(\mathbb{R}^N_+)$. Now, by regularity of the function $d(x)$, we have $d(\eta)=\eta_1+O(|\eta|^2)$, and hence

$$d(\delta\xi)=\delta\xi_1+\delta^2O(|\xi|^2),$$

so that both $\frac{d(\delta\xi)}{\delta}$ and $\frac{\delta}{d(\delta\xi)}$ are locally bounded in D_δ. Then we can apply the Calderon–Zygmund theorem to (2.77) which implies that v_δ is bounded in $W^{2,p}_{\mathrm{loc}}(\mathbb{R}^N_+)$, for all $p>1$, and consequently, by Sobolev embedding, it is locally relatively compact in $C^1(\mathbb{R}^N_+)$. Letting $\delta\to 0$, since $\nabla d(\delta\xi)$ converges toward $e_1=(1,0,\dots,0)$, we find that, up to subsequences, v_δ converges toward a solution v of the equation

$$-\Delta v-2\frac{\nabla v\cdot e_1}{\xi_1}=0\quad\text{in}\quad \mathbb{R}^N_+, \tag{2.78}$$

with the additional information that $|\nabla v|\le C$ and $|v(\xi)|\le c|\xi|$. We claim now that $\partial_{\xi_1}v=0$ for any solution v of (2.78) having these properties. Indeed, let us derive (2.78) with respect to the ξ_1 direction so that $z=\partial_{\xi_1}v$ solves

$$-\Delta z-2\frac{\nabla z\cdot e_1}{\xi_1}+2\frac{z}{\xi_1^2}=0\quad\text{in}\quad\mathbb{R}^N_+,\quad z\in L^\infty(\mathbb{R}^N_+), \tag{2.79}$$

and, say, $|z| \le M$. Standard arguments imply that (2.79) has a maximal and a minimal solution in the class $|z| \le M$. Indeed, if we take a sequence D_ε of smooth bounded domains with $d(D_\varepsilon, \partial\mathbb{R}^N_+) = \varepsilon$, $D_{\varepsilon_1} \subset D_{\varepsilon_2}$, for all $\varepsilon_2 < \varepsilon_1$ such that $D_\varepsilon \to \mathbb{R}^N_+$ as $\varepsilon \to 0$, then we can construct a maximal solution $Z(\xi)$ as the limit of solutions in D_ε of the problem

$$\begin{cases} -\Delta z_\varepsilon - 2\dfrac{\nabla z_\varepsilon \cdot e_1}{\xi_1} + 2\dfrac{z_\varepsilon}{\xi_1^2} = 0 & \text{in } D_\varepsilon, \\ z_\varepsilon = M & \text{on } \partial D_\varepsilon. \end{cases}$$

Note that by the strong maximum principle, $0 \le z_\varepsilon \le M$ and $z \le z_\varepsilon$ for any other solution of (2.79) with $|z| \le M$, and hence the limit Z is maximal in this class. Since the equation is invariant with respect to translations along $\xi_2, \dots, \xi_N$, the maximality of Z implies that $Z(\xi) = Z(\xi_1)$. In the same way, since the equation is linear, we construct a minimal solution that depends only on ξ_1. However, the associated ODE

$$\psi''(s)s^2 + 2s\psi'(s) - 2\psi = 0, \quad s \in (0, +\infty),$$

is a standard Euler equation which admits a unique bounded solution $\psi = 0$. Therefore the maximal and minimal solutions are zero, and hence $z \equiv 0$.

As a consequence, whatever subsequence v_δ is converging in $C^1_{\mathrm{loc}}(\mathbb{R}^N_+)$, we find that $\partial_{\xi_1} v_\delta \to 0$, which yields, by the definition of v_δ and since $e_1 = \nabla d(x_0) = \nabla d(\delta\xi) + O(\delta)$,

$$\nabla v_\delta \cdot \nabla d(\delta\xi) = \nabla u(\delta\xi) \cdot \nabla d(\delta\xi) + \frac{1}{d(\delta\xi)} - \frac{\Delta d(x_0)}{2} \to 0. \tag{2.80}$$

Now observe that in a neighborhood of $\partial\Omega$ any point $x \in \Omega$ can be represented as $x = x_0 + \delta\nabla d(x_0)$ for some $x_0 \in \partial\Omega$, with $\delta = d(x)$. This corresponds to $\xi = (1, 0 \dots, 0)$ in the previous framework, and using (2.80) with $\xi = (1, 0 \dots, 0)$ we can deduce (2.73). $\square$

We can now complete the results stated in the introduction.

Proof of Theorem 1.2. The result is a consequence of Theorem 1.3 and Proposition 2.9, with the characterization of σ_1 in (2.71). $\square$

Proof of Theorem 1.1. Both (1.10) and (1.11) follow straightforward from Theorem 1.2, expanding

$$a(x) = -q|\nabla u(x)|^{q-2}\nabla u(x)$$

near the boundary of Ω. $\square$

Remark 2.10. Let us remark that if $q = 2$ the optimal control law $a(x)$ in general is not tangentially zero (differently than for $1 < q < 2$). Indeed, it is enough to take Ω as the unit ball in $\mathbb{R}^N$ and, in polar coordinates, define $u(x) = \psi(\rho) + \varphi(\theta)$, where ρ is the radial coordinate and θ is the angular one. If ψ is the radial explosive solution of $-\Delta\psi + \psi + |\nabla\psi|^2 = 0$, then u solves

$$-\Delta u + u + |\nabla u|^2 = f, \quad u(x) \to +\infty \quad \text{as } x \to \partial\Omega,$$

where $f = \frac{1}{\rho^2}\left[-\Delta_\theta\varphi + |\nabla_\theta\varphi|^2\right] + \varphi$. Here $a(x) = 2\nabla u(x)$, and its tangential component may not vanish on the boundary. Note that f can be taken as smooth as desired.

Similarly one can show that the results in Theorem 1.2 are optimal both in the normal and in the tangential components and that these latter ones have not in general a universal behavior (independent of f) as is the case for the normal derivative.

The idea of the proof of Proposition 2.9 can be used in order to improve the result of Theorem 1.3 concerning the normal derivative of u: Indeed from Theorem 3.2 we have that $u - S$ is Lipschitz continuous, so that we can look at the equation solved by the rescaled function

$$v_\delta(\xi) = \frac{u(\delta\xi) - S(\delta\xi) - \rho(x_0)}{\delta},$$

where $x_0 \in \partial\Omega$, $\rho(x_0) = (u(x) - S(x))|_{x_0}$, and we are using the same notations as in the above proof. Arguing as before, we can prove that $\partial_{\xi_1} v_\delta \to 0$ and hence that $\frac{\partial(u-S)}{\partial\nu}$ tends to zero at the boundary. This allows us to refine the estimate on the normal component of ∇u.

Corollary 2.11. *Let the assumptions of Theorem* 1.3 *hold true. Then we have*

$$\nabla(u - S)\nabla d \to 0 \quad \text{as } d(x) \to 0,$$

which yields in particular, for any $1 < q < 2$,

$$\begin{cases} \frac{\partial u(x)}{\partial \nu} = \frac{\alpha C^*}{d^{\alpha+1}} - \sum_{k=0}^{[\alpha]+1} \left[\frac{(k-\alpha)\sigma_k(x)}{d^{\alpha-k+1}(x)} - \frac{\nabla\sigma_{k-1}(x)\cdot\nu}{d^{\alpha-k+1}(x)} \right] + o(1) & \text{if } \alpha \notin \mathbb{N}, \\ \frac{\partial u(x)}{\partial \nu} = \frac{\alpha C^*}{d^{\alpha+1}} - \sum\limits_{k=1}^{\alpha-1} \left[\frac{(k-\alpha)\sigma_k(x)}{d^{\alpha-k+1}(x)} - \frac{\nabla\sigma_{k-1}(x)\cdot\nu}{d^{\alpha-k+1}(x)} \right] - \frac{(\sigma_\alpha - \nabla\sigma_{\alpha-1}\cdot\nu)}{d(x)} + o(1) & \text{if } \alpha \in \mathbb{N}. \end{cases}$$

3. Extensions and remarks.

3.1. The case of singular data f. Here we extend the results of Theorem 1.3 (and consequently those of Theorems 1.1 and 1.2) by considering data f possibly singular at the boundary. This is possible by taking full advantage of our use of weighted Bernstein-type estimates in the proof of Theorem 1.3, which allows us to treat data f with weighted regularity. A first result in this sense was essentially contained in the intermediate estimate (2.60) given in the proof of Theorem 1.3, which can be used to get Hölder-type estimates when a Lipschitz regularity cannot be expected.

Theorem 3.1. *Let $u(x)$ be the solution of* (1.1)–(1.2), *and for any* $0 < \beta < \alpha + 1$ *set*

$$m_\alpha^\beta = \left[\alpha - \frac{\beta}{2}\right] + 1.$$

Let

$$S = \begin{cases} S = \sum_{k=0}^{m_\alpha^\beta} \sigma_k d^{k-\alpha} & \text{if } \alpha \notin \mathbb{N}, \\ S = \sum_{k=0}^{\min(\alpha-1, m_\alpha^\beta)} \sigma_k d^{k-\alpha} + \sigma_\alpha^\beta \log d & \text{if } \alpha \in \mathbb{N},\ \alpha \geq 1, \\ S = -\log d & \text{if } \alpha = 0, \end{cases} \tag{3.1}$$

where the coefficients σ_k are as in Theorem 1.3, $\sigma_\alpha^\beta = \sigma_\alpha$ *if* $m_\alpha^\beta = \alpha$ *and* $\sigma_\alpha^\beta = 0$ *if* $m_\alpha^\beta \leq \alpha - 1$.

Assume that $\partial\Omega \in C^r$, $r = m_\alpha^\beta + 4$, *and* $f(x) \in W^{1,\infty}_{\text{loc}}(\Omega)$ *satisfies*

$$|\nabla f| \leq \frac{C}{d^{2+\frac{\beta}{2}}}. \tag{3.2}$$

Then we have

$$|\nabla u - \nabla S| \leq \frac{C}{d^{\frac{\beta}{2}}}, \tag{3.3 (i)}$$

$$\textit{If } \beta < 2, \textit{ then } u - S \textit{ belongs to } C^{0,1-\frac{\beta}{2}}(\overline{\Omega}). \tag{3.4 (ii)}$$

Proof. We come back to Step 1 of the proof of Theorem 1.3, with the same notations. Now, if f satisfies (3.2), we can construct (e.g., by convolution) a smooth approximation f_n satisfying a similar estimate

$$|\nabla f_n| \leq \frac{C}{d_n^{2+\frac{\beta}{2}}} \quad \text{in } \Omega,$$

where $d_n = d(x) + \frac{1}{n}$ as before. Then instead of (2.58) we obtain

$$2d_n^\beta |\nabla f_n \cdot \nabla z_n| \leq \varepsilon \frac{w_n}{d_n^2} + C_\varepsilon d_n^{2+\beta} |\nabla f_n|^2 \leq \varepsilon \frac{w_n}{d_n^2} + \frac{C_\varepsilon}{d_n^2}, \tag{3.5}$$

where $w_n = d_n^\beta |\nabla z_n|^2$. Note that now S is truncated so that F_n defined in (2.40) satisfies $F_n = O(d_n^{m_\alpha^\beta - \alpha - 1})$. Then similarly as in (2.59) we obtain, in a neighborhood Ω_{d_0} near the boundary, that u_n solves

$$-\Delta w_n + 2w_n + c_0 \frac{w_n}{d_n^2} + 2\beta \frac{\nabla w_n \cdot \nabla d}{d_n} + \frac{q|\xi|^{q-2}}{d_n} \nabla w_n \cdot \xi \leq C_0 \left(1 + d_n^{-2(\alpha - \frac{\beta}{2} - [\alpha - \frac{\beta}{2}])} + \frac{1}{d_n^2}\right)$$

equipped with the boundary condition

$$\frac{\partial w_n}{\partial \nu} \leq 0 \quad \text{on } \partial\Omega.$$

The maximum principle still allows one to conclude that $|\nabla z_n| \leq C d_n^{-\frac{\beta}{2}}$, which yields (3.3). In the particular case that $\frac{\beta}{2} < 1$, one can prove (see, e.g., [17]) that (3.3) implies the global Hölder regularity of $u - S$ stated in (ii). ◻

Note that, in terms of the asymptotic expansion for solutions of (1.1)–(1.2), (3.3) implies that $\nabla u = \nabla S + O(d^{-\frac{\beta}{2}})$, which is a sort of truncation of the complete expansion (1.16)–(1.17).

Let us see now that Theorem 1.3 can be extended even to cases when f blows up at the boundary.

THEOREM 3.2. *Let $u(x)$ be the solution of* (1.1)–(1.2). *Assume that $\partial\Omega \in C^r$, $r = [\alpha] + 5$, and $f(x)$ is a $W^{1,\infty}_{\mathrm{loc}}(\Omega)$ function such that*

$$|\nabla f| \leq \frac{\gamma(d)}{d^2}, \tag{3.6}$$

where $\gamma(s)$ is a positive nondecreasing function satisfying $\int_0^1 \frac{\gamma(s)}{s} ds < \infty$.

Let $S(x)$ be defined in (1.15). *Then we have that*

$$u - S \in W^{1,\infty}(\Omega).$$

In particular, the conclusion of Theorem 1.3 *holds true.*

Proof. We use the same notations and closely follow the proof of Theorem 1.3; hence z_n is solution of (2.38)–(2.39), where f_n is an approximation of f satisfying

$$|\nabla f_n| \le C\frac{\gamma(d_n)}{d_n^2}.$$

First note that the conclusion of Step 1 (estimate (2.60)) follows as in Theorem 3.1, since (3.6) implies (3.2) for any $\beta > 0$. Set now

$$w_n = e^{\theta(d_n)}|\nabla z_n|^2$$

as in Step 2 of the proof of Theorem 1.3. By choosing $\theta(s) = \int_0^s \frac{\gamma(t)}{t}dt$ we obtain from (2.65)

(3.7)

$$\begin{aligned}
-\Delta w_n &+ \left[\frac{(\alpha+2)\gamma(d_n)}{d_n^2} + \frac{\gamma'(d_n)d_n - \gamma(d_n)}{d_n^2} - \frac{\gamma(d_n)^2}{d_n^2} + \frac{\gamma(d_n)}{d_n}\Delta d - \frac{\omega(d_n)}{d_n^2}\right] w_n \\
&+ 2w_n + \frac{\alpha+2}{d_n^2}e^{\theta(d_n)}(\nabla d\cdot\nabla z_n)^2 + \frac{q|\xi|^{q-2}}{d_n}\nabla w_n\cdot\xi + 2\frac{\gamma(d_n)}{d_n}\nabla w_n\cdot\nabla d \\
&+2e^{\theta(d_n)}|D^2 z_n|^2 \le 2e^{\theta(d_n)}\left[\nabla f_n\cdot\nabla z_n + C\max\left(d_n^{2([\alpha]-\alpha)}, |\log(d_n)|\right)\right],
\end{aligned}$$

where $\max(d_n^{2([\alpha]-\alpha)}, |\log(d_n)|)$ depends only whether $\alpha\in\mathbb{N}$ or not. Observe that, since $\gamma(0) = 0$ and γ is nondecreasing,

$$\begin{aligned}
&\left[\frac{(\alpha+2)\gamma(d_n)}{d_n^2} + \frac{\gamma'(d_n)d_n - \gamma(d_n)}{d_n^2} - \frac{\gamma(d_n)^2}{d_n^2} + \frac{\gamma(d_n)}{d_n}\Delta d - \frac{\omega(d_n)}{d_n^2}\right] \\
&\qquad\ge \left[\frac{(\alpha+1+o(1))\gamma(d_n)}{d_n^2} - \frac{\omega(d_n)}{d_n^2}\right].
\end{aligned}$$

Now we use (3.6) to deduce

$$|2e^{\theta(d_n)}\nabla f_n\cdot\nabla z_n| \le \varepsilon\frac{\gamma(d_n)}{d_n^2}e^{\theta(d_n)}|\nabla z_n|^2 + C_\varepsilon|\nabla f_n|^2\frac{d_n^2}{\gamma(d_n)} \le \varepsilon\frac{\gamma(d_n)}{d_n^2}w_n + C_\varepsilon\frac{\gamma(d_n)}{d_n^2}.$$

Therefore, by choosing ε small we finally get from (3.7)

$$\begin{aligned}
-\Delta w_n + 2w_n &+ \left[\frac{(\alpha+1)\gamma(d_n)}{d_n^2}(1+o(1)) - \frac{\omega(d_n)}{d_n^2}\right] w_n + (\alpha+2)e^{\theta(d_n)}\left(\frac{\nabla d\cdot\nabla z_n}{d_n}\right)^2 \\
&+ 2\frac{\gamma(d_n)}{d_n}\nabla w_n\cdot\nabla d + \frac{q|\xi|^{q-2}}{d_n}\nabla w_n\cdot\xi \le C_0\left[\frac{\gamma(d_n)}{d_n^2} + \max\left(d_n^{2([\alpha]-\alpha)}, |\log(d_n)|\right)\right].
\end{aligned}$$

Without loss of generality we may assume that $\gamma(s) \ge \max(s^{2([\alpha]+1-\alpha)}, s^\eta)$ with $\eta < 1$: In particular, by using (2.66) as in Theorem 1.3, we have $\omega(d_n) = o(\gamma(d_n))$. Therefore we obtain for positive constants c_0, C_1:

$$-\Delta w_n + 2w_n + \frac{c_0\,\gamma(d_n)}{d_n^2}w_n + 2\frac{\gamma(d_n)}{d_n}\nabla w_n\cdot\nabla d + \frac{q|\xi|^{q-2}}{d_n}\nabla w_n\cdot\xi \le C_1\,\frac{\gamma(d_n)}{d_n^2}.$$

Moreover, since for s small, $\frac{\gamma(s)}{s} - 2\|D^2 d\|_\infty > 0$, Lemma 2.4 can be applied, and we have

$$\frac{\partial w_n}{\partial\nu} \le 0 \qquad \text{on } \partial\Omega.$$

Using the maximum principle we conclude that

$$w_n = e^{\theta(d_n)}|\nabla z_n|^2 \le C,$$

which yields $|\nabla z_n|^2 \le C$, and then in the limit as $n \to +\infty$ the Lipschitz regularity of $z = u - S$. $\square$

Remark 3.3. Several variants of the results above could be done by varying the weight function, i.e., by reasoning on $w = \Phi(d)|\nabla z|^2$ with possibly different functions Φ rather than those used above. An example is given by using $\Phi(d) = |\log d|^{-2}$: In this case, assuming f such that $|\nabla f| \le \frac{C}{d^2}$, one can prove that $|\nabla(u - s)| \le C|\log d|$.

Remark 3.4. A different approach to the results of Theorem 3.1 or 3.2 can be done if we can expand f as

$$f(x) = \sum_{j=0}^{m_f} \varphi_j d^{j-\gamma}, \tag{3.8}$$

where φ_j are smooth functions and $\gamma < \alpha + 2$. In this case we can modify the so-called "corrector term" S to get rid of the singular part of f. Indeed we can define S as

$$S = \sum_{k=0}^{m} \sigma_k d^{k-\alpha} + \sum_{j=0}^{m_f} \psi_j d^{j+2-\gamma}$$

and choose the functions ψ_j (depending on the φ_j's) to reduce the problem to the case of smooth data f. Actually requiring (3.8) means that $d^\gamma f$ is smooth, which is a stronger assumption than $|\nabla f| \le \frac{C}{d^{\gamma+1}}$: For instance, if $x_0 \in \partial\Omega$, then $f(x) = |x - x_0|^{-\gamma}$ satisfies $|\nabla f| \le C d^{-\gamma-1}$ but $d^\gamma f \notin C^0(\Omega)$.

In particular, if we assume the condition that $d(x)f$ is smooth (which is more demanding than (3.6) as regards the tangential behavior but possibly less strong in the normal component), then the result of Theorem 3.2 still holds true. This suggests that (3.6) may possibly be relaxed as far as the normal derivative of f is concerned, but we suspect it to be optimal due to tangential effects. On the other hand, we observe that the assumption (3.2) in Theorem 3.1 is optimal in order to get the intermediate estimate (3.3).

Remark 3.5. In the case of data f singular at the boundary, the method of proof of Theorem 1.3 may be used (although this extension is not trivial) to obtain a stabilization result as follows: Let u_1 and u_2 be solutions of (1.1)–(1.2) corresponding to different data f_1 and f_2 such that $|\nabla f_i| = O(d^{-\gamma})$, with $\gamma < \alpha + 2$. Then one has $\|\nabla(u_1 - u_2)\|_{L^\infty(\Omega)} \le C(1 + \|\nabla(f_1 - f_2)\|_{L^\infty(\Omega)})$.

3.2. Expansion of ∇u via blow-up argument. As pointed out in the introduction, the typical approach used in previous papers (see, e.g., the works [1], [3], [18], and references therein) for the asymptotic development of the gradient of explosive solutions consists in deriving the behavior of ∇u via scaling and blow-up once a precise estimate on the asymptotics of u has been established. In the present context, one could rephrase this method as follows: If one knows that

$$u(x) - S_k(x) = o(d^{k-\alpha}), \tag{3.9}$$

where $S_k(x)$ are the first $k+1$ terms in the formal expansion of u introduced in (1.15), then one can expect to prove that $\nabla u - \nabla S_k(x) = o(d^{k-\alpha-1})$. Indeed, this is the method used in [18] to get estimate (1.8) for solutions of (1.1)–(1.2), which the reader has also found in Lemma 2.7 in a similar version for the approximating solutions. Let us point out that this alternative approach works provided $\alpha > k$, i.e., if the translated function $u(x) - S_k(x)$ is still explosive. Recalling the value of $\alpha = \frac{2-q}{q-1}$, this means, for instance, that the second order terms of the expansion ($k = 1$ in (3.9)) can be recovered with this method provided $q < \frac{3}{2}$. Thus, although the result of Theorem 1.3 is clearly more general and complete, in some particular cases the approach via blow-up may still provide some intermediate results on the expansion of ∇u at the boundary, with the advantage of requiring slightly weaker assumptions on f (namely, no assumption on ∇f is required) and on $\partial\Omega$. It may be of interest to state explicitly this result.

THEOREM 3.6. *Let Ω be of class C^{k+3}, and let $u(x)$ be the unique solution of* (1.1)–(1.2). *Let f be such that $d^{\alpha+2-k}|f|$ is bounded and tends to zero at the boundary, and suppose $\alpha > k$, $k \in \mathbb{N}$, $k \geq 1$. Then*

$$d^{\alpha+1-k}(x)\left(\frac{\partial u(x)}{\partial \nu} - \frac{\partial S_k(x)}{\partial \nu}\right) \to 0 \quad \text{as} \quad d(x) \to 0, \tag{3.10}$$

and

$$d^{\alpha+1-k}(x)\left(\frac{\partial u(x)}{\partial \tau} - \frac{\partial S_k(x)}{\partial \tau}\right) \to 0 \quad \text{as} \quad d(x) \to 0, \tag{3.11}$$

where $S_k(x)$ are the first $k+1$ terms in the formal expansion introduced in (1.15).

The result of the above theorem generalizes, in some sense, the first order estimate proved in [18] (that is, (3.10)–(3.11) for $k = 0$). We also stress that from (3.10)–(3.11) with $k = 1$ we already can deduce an estimate on the control law which is weaker than (1.10) but requires weaker hypotheses, i.e., $\partial\Omega \in C^4$ and $f(x) = o(d^{-\alpha-1})$; namely, we have for $\alpha > 1$ (i.e. $q < \frac{3}{2}$)

$$a(x) = -\frac{q'}{d(x)}\,\nu(x) - \frac{q'(N-1)}{2} H(\overline{x})\,\nu(x) + o(1) \qquad \text{as} \qquad d(x) \to 0.$$

Sketch of the proof. Suppose $\alpha \notin \mathbb{N}$ (otherwise, the proof is slightly different). It is easy to prove, via sub- and supersolutions, that if $\alpha > k$, $k \in \mathbb{N}$, then (3.9) holds true. Note that, in order to get such an estimate, we use the regularity of the boundary and the assumption on $f(x)$. Then we use the same blow-up framework as in Proposition 2.9: We fix $x_0 \in \partial\Omega$, $d_0 > 0$ (small) and $P_0 = x_0 + d_0 \nabla d(x_0)$, and we denote by (η_1, η') a system of coordinates (obtained by translation and rotation) such that the origin $O \equiv x_0$ and the η_1-axis coincides with $\nabla d(x_0) = -\nu(x_0)$. Let $\eta \in B(P_0, d_0 - \delta)$, and, for any $0 < \delta < d_0$, define the rescaled variable $\xi = \frac{\eta - O_\delta}{\delta}$, where $O_\delta = (\delta, 0, \ldots, 0)$, and the rescaled function

$$v_\delta(\xi) = \frac{u(\eta) - S_{k-1}(\eta)}{\delta^{-\alpha+k}} = \frac{u(\delta\xi + O_\delta) - S_{k-1}(\delta\xi + O_\delta)}{\delta^{-\alpha+k}}.$$

Since u solves (1.1), by computing we find that v_δ solves for ξ in a domain $D_\delta \to \mathbb{R}^N_+$:

$$-\Delta v_\delta + \delta^2 v_\delta + H(\xi, \nabla v_\delta) = F_\delta(\delta\xi + O_\delta), \tag{3.12}$$

where (recall that $\eta = \delta\xi + O_\delta$)

$$H(\xi, p) = \delta^{\alpha+2-k}\left(\left|p\,\delta^{k-\alpha} + \nabla S_{k-1}(\eta)\right|^q - |\nabla S_{k-1}(\eta)|^q\right)$$

and

$$-\delta^{k-\alpha-2} F_\delta(\eta) = -\Delta S_{k-1}(\eta) + S_{k-1}(\eta) + |\nabla S_{k-1}(\eta)|^q - f(\eta).$$

Now by choosing σ_j as in Theorem 1.3 and using also the hypotesis on f, we have that the right-hand side is $O(d(\eta)^{k-2-\alpha})$; moreover, since $d(x)$ is smooth, we have $d(\eta) = \delta(1+\xi_1)+\delta^2 O(|\xi|^2)$, and consequently both $\frac{d(\eta)}{\delta}$ and $\frac{\delta}{d(\eta)}$ are locally uniformly bounded on $\mathbb{R}^N_+$. Therefore we deduce that $F_\delta(\eta)$ is a locally uniformly bounded function. Similarly, if we write

$$H(\xi, \nabla v_\delta) = \nabla v_\delta \cdot G_\delta(\xi), \quad \text{where} \quad G_\delta(\xi) = \int_0^1 H_p(\xi, t\nabla v_\delta(\xi))dt,$$

by computing $H_p(\xi, tp)$ and using that $|\nabla v_\delta| = o(\delta^{-1})$ (in consequence of (1.8)) and the growth of S_{k-1}, we have that $G_\delta(\xi)$ is locally bounded in $\mathbb{R}^N_+$. Since (by (3.9)) $v_\delta(\xi)$ is locally bounded as well, we can use the Calderon–Zygmund regularity for (3.12), and with a bootstrap argument and the Sobolev embedding theorem we deduce that $v_\delta(\xi)$ is compact in $C^1_{\mathrm{loc}}(\mathbb{R}^N_+)$. However, thanks to (3.9) we have that $(\frac{d(\eta)}{\delta})^{\alpha-k} v_\delta(\xi) \to \sigma_k(x_0)$; hence $\nabla v_\delta(\xi) \to (k-\alpha)(1+\xi_1)^{k-\alpha-1}\sigma_k(x_0)\,\nabla d(x_0)$. Estimates (3.10) and (3.11) are a consequence of this convergence. $\square$

3.3. Generalizations. With the approach of the present paper, it is possible to get similar results as in Theorem 1.3 for possibly more general examples than (1.1), as, for instance, if we consider

$$-\Delta u + \psi(x, \nabla u) + u = f(x) \qquad \text{in} \quad \Omega,$$

always coupled with explosive boundary condition (1.2). Let us mention that the method of proof of Theorem 1.3 still works if, for instance, we assume that $\psi(x,\xi) = |\xi|^q + H(x,\xi)$, where $H(x,\xi)$ is a C^1 function satisfying, for $|\xi|$ large enough, the growth conditions

$$\begin{aligned}
&|H_\xi| \le C|\xi|^r d(x)^{-\gamma},\\
&|H_\xi(x,\xi) - H_\xi(x,\eta)| \le (|\xi|+|\eta|)^{r-1}|\xi-\eta| d(x)^{-\gamma}, \qquad \text{with } \tfrac{r}{q-1}+\gamma<1,\\
&|H_x(x,\xi) - H_x(x,\eta)| \le (|\xi|+|\eta|)^{s}|\xi-\eta| d(x)^{-t-1}, \qquad \text{with } \tfrac{s}{q-1}+t<1,
\end{aligned}$$

provided we choose the corrector term S in a way that $-\Delta S + S + \psi(x, \nabla S)$ is not too singular at the boundary. Of course, this last step may turn out to be rather complicated and may change drastically the form of the expansion S depending on the structure of $\psi(x,\xi)$. However, if we have that $|H(x,\xi)| \le C|\xi|^{q-1}$, then one can take S of the same form as in (1.15) but for possibly different coefficients σ_k; this happens, for instance, in the example $\psi(x,\xi) = |\xi|^q + B(x)\cdot\xi$, with $B(x)$ smooth. In this latter case the influence of B on the solution can be seen even in the second term in the expansion of the gradient (which will not depend merely on the mean curvature). Similarly, we can also consider an example of an absorption-reaction term of the type $\psi(x,\xi) = |\xi|^q - |\xi|^p$: If $0 < p \le q-1$, then the same result as in Theorem 1.3 holds true, with the same corrector function S defined in (1.15). On the other hand, if we have $q-1 < p < q$, the same method of proof can be applied, but the construction of S needs to be suitably adapted (with quite long and tedious computations due to the superposition of the two terms).

Acknowledgment. We warmly thank J. Droniou for his kind advice concerning Lemma 2.7 and for pointing out to us the related reference [11].

REFERENCES

[1] C. Bandle and M. Essen, *On the solutions of quasilinear elliptic problems with boundary blow-up*, Symposia Math., 35 (1994), pp. 93–111.

[2] C. Bandle and E. Giarrusso, *Boundary blow-up for semilinear elliptic equations with nonlinear gradient terms*, Adv. Differential Equations, 1 (1996), pp. 133–150.

[3] C. Bandle and M. Marcus, *Asymptotic behaviour of solutions and their derivatives, for semilinear elliptic problems with blowup on the boundary*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 12 (1995), pp. 155–171.

[4] C. Bandle and M. Marcus, *Dependence of blow–up rate of large solutions of semilinear elliptic equations on the curvature of the boundary*, Complex Var. Theory Appl., 49 (2004), pp. 555–570.

[5] G. Barles, *A weak Bernstein method for fully nonlinear elliptic equations*, Differential Integral Equations, 4 (1991), pp. 241–262.

[6] G. Barles and F. Da Lio, *On the generalized Dirichlet problem for viscous Hamilton-Jacobi equations*, J. Math. Pures Appl., 9 (2004), pp. 53–75.

[7] G. Barles and E. Rouy, *A strong comparison result for the Bellman equation arising in stochastic exit time control problems and its applications*, Comm. Partial Differential Equations, 23 (1998), pp. 1995–2033.

[8] S. Bernstein, *Sur la généralization du probléme de Dirichlet* II, Math. Ann., 69 (1910), pp. 82–136.

[9] H. Brezis, *Analyse Fonctionnelle. Théorie et Applications*, Masson, Paris, 1983.

[10] M. Del Pino and R. Letelier, *The influence of domain geometry in boundary blow-up elliptic problems*, Nonlinear Anal., 48 (2002), pp. 897–904.

[11] T. Gallouët and A. Monier, *On the regularity of solutions to elliptic equations,* Rend. Mat. Appl., 19 (1999), pp. 471–488.

[12] D. Gilbarg and N. Trudinger, *Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin/New York, 1983.

[13] M. A. Katsoulakis, *Viscosity solutions of second order fully nonlinear elliptic equations with state constraints*, Indiana Univ. Math. J., 43 (1994), pp. 493–519.

[14] H. Ishii and P. Loreti, *A class of stochastic optimal control problems with state constraint*, Indiana Univ. Math. J., 51 (2002), pp. 1167–1196.

[15] J.-M. Lasry and P.-L. Lions, *Nonlinear elliptic equations with singular boundary conditions and stochastic control with state constraints.* I. *The model problem*, Math. Ann., 283 (1989), pp. 583–630.

[16] P.-L. Lions, *Résolution de problèmes elliptiques quasilinéaires*, Arch. Ration. Mech. Anal., 74 (1980), pp. 234–254.

[17] P.-L. Lions, *Quelques remarques sur les problémes elliptiques quasilinèaires du second ordre*, J. Anal. Math., 45 (1985), pp. 335–353.

[18] A. Porretta and L. Veron, *Asymptotic behaviour for the gradient of large solutions to some nonlinear elliptic equations*, Adv. Nonlinear Stud., 6 (2006), pp. 351–378.

[19] J. Serrin, *The problem of Dirichlet for quasilinear elliptic differential equations with many independent variables*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 264 (1969), pp. 413–496.

SIAM J. MATH. ANAL.
Vol. 39, No. 4, pp. 1328–1343

AN INVERSE PROBLEM FOR A DYNAMICAL LAMÉ SYSTEM WITH RESIDUAL STRESS*

VICTOR ISAKOV†, JENN-NAN WANG‡, AND MASAHIRO YAMAMOTO§

Abstract. In this paper we prove Hölder and Lipschitz stability estimates for determining all coefficients of a dynamical Lamé system with residual stress, including the density, Lamé parameters, and the residual stress, by three pairs of observations from the whole boundary or from a part of it. These estimates imply first uniqueness results for determination of all parameters in the residual stress systems from few boundary measurements. Our essential assumptions are that the Lamé system possesses a suitable pseudoconvex function, residual stress is small, and three sets of the initial data satisfy some independence condition.

Key words. elasticity system with residual stress, inverse problem, Carleman estimates

AMS subject classifications. 35R30, 74B10

DOI. 10.1137/060669115

1. Introduction. We consider an elasticity system with residual stress. This system is anisotropic; i.e., it exhibits elastic properties with different values when measured in various directions. The assumption about isotropy is too restrictive in most important applications, although it allows deeper mathematical analysis of direct, and especially inverse, problems. While the theory of the unique solvability of direct problems in a quite general anisotropic case is relatively well developed [3], almost nothing is known about determination of anisotropic elastic parameters from additional boundary value data (i.e., about inverse problems).

We handle the simplest anisotropy, known as the Lamé system with residual stress, which is a small perturbation of the classical isotropic Lamé system, by a scalar anisotropic second order operator. Smallness of perturbation is motivated by applications to material science [14]. Assuming that speeds of propagation of shear and compression waves in an unperturbed system satisfy some pseudoconvexity-type conditions (which exclude trapped elastic rays) and that three sets of initial conditions are in a certain sense independent, we obtain first uniqueness and stability results about identification of all nine elastic parameters of an isotropic medium with residual stress from lateral boundary observations. When observation time and the observed part of the boundary are arbitrary, we explicitly describe a domain where coefficients are guaranteed to be unique, and we give a Hölder stability estimate. When observation time is sufficiently large and observation is from the whole lateral boundary, we derive Lipschitz stability estimates. These estimates indicate the possibility of a numerical solution of an inverse problem with high resolution and therefore

*Received by the editors September 5, 2006; accepted for publication (in revised form) May 23, 2007; published electronically December 19, 2007.
http://www.siam.org/journals/sima/39-4/66911.html

†Department of Mathematics and Statistics, Wichita State University, Wichita, KS 67206 (victor.isakov@wichita.edu). The work of this author was partially supported by NSF grant DMS 04-05976.

‡Department of Mathematics, Taida Institute for Mathematical Sciences and NCTS (Taipei), National Taiwan University, Taipei 106, Taiwan (jnwang@math.ntu.edu.tw). The work of this author was partially supported by grant NSC 95-2115-M-002-003 of the National Science Council of Taiwan.

§Department of Mathematical Sciences, University of Tokyo, 3-8-1 Komaba, Meguro, Tokyo 153-8914, Japan (myama@ms.u-tokyo.ac.jp). The work of this author was partially supported by grant 15340027 from the Japan Society for the Promotion of Science and grant 15654015 from the Ministry of Education, Cultures, Sports, and Technology.

a substantial applied potential.

While our assumptions exclude zero initial data (most natural in many applications), recent progress in generating wave fields by interior sources in geophysics, material sciences, and medicine, and also by a substantial amount of historical seismic data from earthquakes (which are interior sources), make our assumptions more realistic.

Let Ω be an open bounded domain in $\mathbb{R}^3$ with boundary $\partial\Omega \in C^8$. The residual stress is modeled by a symmetric second-rank tensor $R(x) = (r_{jk}(x))_{j,k=1}^3 \in C^7(\overline{\Omega})$ which is divergence free

$$\text{div} R = 0 \quad \text{in} \quad \Omega \tag{1.1}$$

and satisfies the boundary condition

$$R\nu = 0 \quad \text{on} \quad \partial\Omega, \tag{1.2}$$

where $\text{div}R$ is a vector-valued function with components given by

$$(\text{div}R)_j = \sum_{k=1}^{3} \partial_k r_{jk}, \quad 1 \le j \le 3.$$

In this paper, $x = (x_1, x_2, x_3) \in \mathbf{R}^3$ and $\nu = (\nu_1, \nu_2, \nu_3)^\top$ is the unit outer normal vector to $\partial\Omega$. Here and below, differential operators ∇ and Δ, without subscripts, are with respect to x variables. Let $Q = \Omega \times (-T, T)$ and $\mathbf{u} = (u_1, u_2, u_3)^\top : Q \to \mathbb{R}^3$ be the displacement vector in Q. We note that $\epsilon(\mathbf{u}) = (\nabla\mathbf{u} + \nabla\mathbf{u}^\top)/2$ is the strain tensor. We consider the initial boundary value problem

$$\mathbf{A}_E\mathbf{u} := \rho\partial_t^2\mathbf{u} - \mu\Delta\mathbf{u} - (\lambda+\mu)\nabla(\text{div}\,\mathbf{u}) - (\nabla\lambda)\text{div}\,\mathbf{u} - 2\epsilon(\mathbf{u})\nabla\mu - \text{div}((\nabla\mathbf{u})R) = 0 \text{ in } Q, \tag{1.3}$$

$$\mathbf{u} = \mathbf{u}_0, \quad \partial_t\mathbf{u} = \mathbf{u}_1 \quad \text{on} \quad \Omega \times \{0\}, \tag{1.4}$$

$$\mathbf{u} = \mathbf{g} \quad \text{on} \quad \partial\Omega \times (-T, T), \tag{1.5}$$

where ρ is density and λ and μ are Lamé parameters depending only on x and satisfying inequalities

$$\varepsilon_1 < \mu, \quad \varepsilon_1 < \rho, \quad \varepsilon_1 < \lambda + \mu \quad \text{on} \quad \Omega \tag{1.6}$$

for some positive constant ε_1. Hereafter, we use E to represent the set of elastic coefficients in (1.3), i.e., $E = (\rho, \lambda, \mu, R)$. We will assume that $\rho \in C^6(\overline{\Omega})$ and $\lambda, \mu \in C^7(\overline{\Omega})$. The system (1.3) can be written as

$$\rho\, \partial_t^2\mathbf{u} - \text{div}\,\sigma(\mathbf{u}) = 0,$$

where $\sigma(\mathbf{u}) = \lambda(\text{tr}\,\epsilon)I + 2\mu\epsilon + R + (\nabla\mathbf{u})R$ is a stress tensor. Note that the term div R does not appear in (1.3) due to (1.1). Also, due to the same condition, we can see that

$$(\text{div}((\nabla\mathbf{u})R))_i = \sum_{j,k=1}^{3} r_{jk}\partial_j\partial_k u_i, \quad 1 \le i \le 3.$$

To make sure that problem (1.3) with (1.4) and (1.5) is well-posed, it suffices to assume that

$$\|R\|_{C^1(\overline{\Omega})} < \varepsilon_0 \tag{1.7}$$

for some (small) constant $\varepsilon_0 > 0$. Assumption (1.7) is also motivated by material science applications [14]. Indeed, residual stress of interest to engineers is due to past thermal changes in steel production which do not significantly change the elastic properties of steel. It is not hard to see that if ε_0 is sufficiently small, then the boundary value problem (1.3)–(1.5) is hyperbolic, and hence for any initial data $(\mathbf{u}_0, \mathbf{u}_1) \in H^1(\Omega) \times L^2(\Omega)$ and lateral Dirichlet data $\mathbf{g} \in C^1([-T,T]; H^1(\Omega))$, $\mathbf{u}_0 = \mathbf{g}$ on $\partial\Omega \times \{0\}$, there exists a unique solution $\mathbf{u}(\cdot; E; \mathbf{u}_0, \mathbf{u}_1, \mathbf{g}) \in C([-T,T]; H^1(\Omega))$ to (1.3)–(1.5).

In this paper we are interested in the following inverse problem: Let Γ be an open subset of $\partial\Omega$ with $\partial\Gamma \in C^1$. Determine density ρ, Lamé parameters λ, μ, and the residual stress R (a total of nine functions) from Cauchy-type data $(\mathbf{u}, \sigma(\mathbf{u})\nu)$ on $\Gamma \times (-T,T)$, where $\mathbf{u} = \mathbf{u}(\cdot; E; \mathbf{u}_0, \mathbf{u}_1, \mathbf{g})$, given for a finite number of pairs of initial data $(\mathbf{u}_0, \mathbf{u}_1)$.

We will address uniqueness and stability issues. Our main focus is on stability, since stability implies uniqueness. This work is a sequel to our recent paper [11], where we demonstrated uniqueness of only R assuming known constant ρ, λ, μ. Our method is based on Carleman estimates techniques initiated by Bukhgeim and Klibanov [2]. For works on Carleman estimates and related inverse problems for scalar equations, we refer to books [1] and [12] for further details and references. The method of [2] was modified for scalar equations in the paper of Imanuvilov and Yamamoto [6]. It was found by Imanuvilov, Isakov, and Yamamoto [8] that this modification allows one to obtain uniqueness and stability for coefficients of systems of equations; in particular, in [8] there is a first uniqueness result for all three elastic parameters ρ, λ, μ of isotropic elasticity. For further results on identification of the isotropic Lamé system we refer to [7]. For Carleman estimates and uniqueness of the continuation for the residual stress system (1.3) and for identification of the source term and R with given constant ρ, λ, μ, we refer to [10], [11], [13]. In the case of many boundary measurements and zero initial data, there are only partial results on identification of residual stress [5], [15]. In the present work we will show that we can determine all nine parameters in (1.3)–(1.5) by three pairs of Cauchy data. We will derive a Hölder stability estimate in the convex hull of the observation surface Γ and a Lipschitz stability estimate for (ρ, λ, μ, R) in Ω when $\Gamma = \partial\Omega$ and observation time T is large.

We are now ready to state our main results. Let $d = \inf |x|$ and $D = \sup |x|$ over $x \in \Omega$. We will assume that

$$0 < d. \tag{1.8}$$

Let θ be a positive number. For a function $c \in C^1(\overline{\Omega})$ we introduce the condition

$$\theta^2 < c \quad \text{and} \quad \theta^2 c + D\theta\sqrt{c}|\nabla c| + \frac{1}{2} c\, x \cdot \nabla c < c^2 \quad \text{on} \quad \overline{\Omega}. \tag{1.9}$$

Let $\varepsilon_0 > 0$ be given as in (1.7), $M > 0$ be arbitrarily fixed, and $\mathcal{E}_{\varepsilon_0, M}$ be the class of functions (elastic parameters) defined by

$$\begin{aligned}\mathcal{E}_{\varepsilon_0,M} = \quad & \{(\rho, \lambda, \mu, R) : \ \|\rho\|_{C^6(\overline{\Omega})} + \|\lambda\|_{C^7(\overline{\Omega})} + \|\mu\|_{C^7(\overline{\Omega})} + \|R\|_{C^6(\overline{\Omega})} < M : \\ & \rho, \lambda, \mu \text{ satisfy (1.6) and } c = \frac{\mu}{\rho},\ c = \frac{\lambda + 2\mu}{\rho} \text{ satisfy (1.9),} \\ & R \text{ is symmetric and satisfies (1.1), (1.2), and (1.7)}\}.\end{aligned}$$

To study the inverse problem, we need not only the well-posedness of (1.3)–(1.5) but also some extra regularity of the solution $\mathbf{u}$. To achieve the latter property, the initial and Dirichlet data $(\mathbf{u}_0, \mathbf{u}_1, \mathbf{g})$ are required to satisfy some smoothness and compatibility conditions. More precisely, we will assume that $\mathbf{u}_0 \in H^9(\Omega), \mathbf{u}_1 \in H^8(\Omega)$, and $\mathbf{g} \in C^8([-T,T]; H^1(\partial\Omega)) \cap C^5([-T,T]; H^4(\partial\Omega))$ and that they satisfy standard compatibility conditions of order 8 at $\partial\Omega \times \{0\}$. By using energy estimates [3] and Sobolev embedding theorems as in [8], one can show that

$$\|\partial_x^\alpha \partial_t^\beta \mathbf{u}\|_{C^0(\overline{Q})} \leq C \tag{1.10}$$

for $|\alpha| \leq 2$ and $0 \leq \beta \leq 5$. We will use three sets of initial data $(\mathbf{u}_0(\cdot;j), \mathbf{u}_1(\cdot;j))$, $j = 1,2,3$. To guarantee uniqueness in the inverse problem, we impose some nondegeneracy condition on the initial data. Namely, let $\mathbf{M}$ denote the 18×13 matrix

$$\begin{pmatrix}
\mu\Delta\mathbf{u}_0(\cdot;1) + (\lambda+\mu)\nabla(\operatorname{div}\mathbf{u}_0(\cdot;1)) & \operatorname{div}\mathbf{u}_0(\cdot;1)I_3 & 2\epsilon(\mathbf{u}_0(\cdot;1)) & \mathbf{R}(\mathbf{u}_0(\cdot;1)) \\
\mu\Delta\mathbf{u}_1(\cdot;1) + (\lambda+\mu)\nabla(\operatorname{div}\mathbf{u}_1(\cdot;1)) & \operatorname{div}\mathbf{u}_1(\cdot;1)I_3 & 2\epsilon(\mathbf{u}_1(\cdot;1)) & \mathbf{R}(\mathbf{u}_1(\cdot;1)) \\
\mu\Delta\mathbf{u}_0(\cdot;2) + (\lambda+\mu)\nabla(\operatorname{div}\mathbf{u}_0(\cdot;2)) & \operatorname{div}\mathbf{u}_0(\cdot;2)I_3 & 2\epsilon(\mathbf{u}_0(\cdot;2)) & \mathbf{R}(\mathbf{u}_0(\cdot;2)) \\
\mu\Delta\mathbf{u}_1(\cdot;2) + (\lambda+\mu)\nabla(\operatorname{div}\mathbf{u}_1(\cdot;2)) & \operatorname{div}\mathbf{u}_1(\cdot;2)I_3 & 2\epsilon(\mathbf{u}_1(\cdot;2)) & \mathbf{R}(\mathbf{u}_1(\cdot;2)) \\
\mu\Delta\mathbf{u}_0(\cdot;3) + (\lambda+\mu)\nabla(\operatorname{div}\mathbf{u}_0(\cdot;3)) & \operatorname{div}\mathbf{u}_0(\cdot;3)I_3 & 2\epsilon(\mathbf{u}_0(\cdot;3)) & \mathbf{R}(\mathbf{u}_0(\cdot;3)) \\
\mu\Delta\mathbf{u}_1(\cdot;3) + (\lambda+\mu)\nabla(\operatorname{div}\mathbf{u}_1(\cdot;3)) & \operatorname{div}\mathbf{u}_1(\cdot;3)I_3 & 2\epsilon(\mathbf{u}_1(\cdot;3)) & \mathbf{R}(\mathbf{u}_1(\cdot;3))
\end{pmatrix},$$

where I_3 is the 3×3 identity matrix, and where $\mathbf{R}(\mathbf{v})$ is a 3×6 matrix defined by

$$\mathbf{R}(\mathbf{v}) = \begin{pmatrix} \partial_1^2\mathbf{v} & 2\partial_1\partial_2\mathbf{v} & 2\partial_1\partial_3\mathbf{v} & \partial_2^2\mathbf{v} & 2\partial_2\partial_3\mathbf{v} & \partial_3^2\mathbf{v} \end{pmatrix}. \tag{1.11}$$

We will assume that

(1.12) there exists a 13×13 minor of $\mathbf{M}$ such that the absolute value of its determinant is greater than some constant $\varepsilon_2 > 0$ on $\overline{\Omega}$.

One can check that $\mathbf{u}_0(\cdot;1) = (x_1x_2, 0, 0)^\top$, $\mathbf{u}_1(\cdot;1) = (0,0,0)^\top$, $\mathbf{u}_0(\cdot;2) = (x_1, x_2, x_3)^\top$, $\mathbf{u}_1(\cdot;2) = (0, x_2, x_3)$, $\mathbf{u}_0(\cdot;3) = (x_1^2, x_2^2, x_3^2)^\top$, and $\mathbf{u}_1(\cdot;3) = (x_2x_3, x_1x_3, x_1x_2)^\top$ satisfy (1.12). Here, 13 row vectors from rows 2 and 7–18 are linearly independent on $\overline{\Omega}$. In fact, the direct calculations yield that the absolute value of the determinant of 13×13 minor is $2^{10}(\lambda(x) + \mu(x))$, and we can choose $\varepsilon_2 = 2^{10}\varepsilon_1$ in (1.12), where $\varepsilon_1 > 0$ is given in (1.6).

Condition (1.12) does not hold physically, but for the identification of the residual stress, the density, and the Lamé coefficients we have to set up the system by choosing initial values artificially, e.g., in a laboratory. The above example of such initial values suggests that there may be many choices for it.

We will use the following notation.

C, γ are generic constants depending only on $\Omega, T, \delta, \varepsilon_0, \varepsilon_1, \varepsilon_2, M, \mathbf{u}_0(\cdot;j), \mathbf{u}_1(\cdot;j)$, $j = 1,2,3$ (any other dependence will be indicated). $\|\cdot\|_{(k)}(Q)$ is the norm in the Sobolev space $H^k(Q)$. $Q(\varepsilon) = Q \cap \{\varepsilon < |x|^2 - \theta^2t^2 - d_1^2\}$ and $\Omega(\varepsilon) = \Omega \cap \{\varepsilon < |x|^2 - d_1^2\}$, where $d_1 \geq d$. Let $\mathbf{u}(;1;j)$ and $\mathbf{u}(;2;j)$ be solutions of (1.3), (1.4) with initial data $(\mathbf{u}_0(;j), \mathbf{u}_1(;j))$, for $j = 1,2,3$, corresponding to sets of coefficients $E_1 = (\rho_1, \lambda_1, \mu_1, R_1)$ and $E_2 = (\rho_2, \lambda_2, \mu_2, R_2)$, respectively. We will consider the Dirichlet data (displacements) as measurements (observations). We introduce the norm of the differences of the data as

$$F = \sum_{j=1}^{3}\sum_{\beta=2}^{4}(\|\partial_t^\beta(\mathbf{u}(;2;j) - \mathbf{u}(;1;j))\|_{(\frac{5}{2})}(\Gamma \times (-T,T))$$
$$+\|\partial_t^\beta\sigma(\mathbf{u}(;2;j) - \mathbf{u}(;1;j))\nu\|_{(\frac{3}{2})}(\Gamma \times (-T,T))).$$

This data norm includes the fourth order time derivatives and is technically necessary for our proof of the Lipschitz stability in the inverse problem. Because we have to obtain a suitable number of equations in 13 unknown functions ρ, λ, μ, r_{jk} ($1 \le j \le k \le 3$) in terms of data, we will use also $\partial_t^3 \mathbf{u}(\ ;k;j)$, $j = 1,2,3$ at $t = 0$ (see (3.9)). For that, we need L^2-estimates of $\partial_t^4 \mathbf{u}$ in (x,t) in the Carleman estimate, which yield estimates of $\partial_t^3 \mathbf{u}(\ ;k;j)$ at $t = 0$ (see (3.12)).

We first state the Hölder-type estimate for determining coefficients in $\Omega(\varepsilon)$.

THEOREM 1.1. *Assume that the domain Ω satisfies* (1.8) *and for some d_1 ($\ge d$),*

$$|x|^2 - d_1^2 < 0 \text{ when } x \in (\partial\Omega \setminus \Gamma) \text{ and } D^2 - \theta^2 T^2 - d_1^2 < 0. \tag{1.13}$$

Let the initial data $(\mathbf{u}_0(;j), \mathbf{u}_1(;j))$, $j = 1,2,3$, satisfy (1.12) *with $\lambda = \lambda_1, \mu = \mu_1$.*

Then there exist ε_0 and constants $C, \gamma \in (0,1)$ such that for $E_1, E_2 \in \mathcal{E}_{\varepsilon_0,M}$ with

$$\lambda_1 = \lambda_2 \quad \text{and} \quad \mu_1 = \mu_2 \quad \text{on} \quad \Gamma, \tag{1.14}$$

one has

$$\|\rho_1-\rho_2\|_{(0)}(\Omega(\varepsilon))+\|\lambda_1-\lambda_2\|_{(0)}(\Omega(\varepsilon))+\|\mu_1-\mu_2\|_{(0)}(\Omega(\varepsilon))+\|R_1-R_2\|_{(0)}(\Omega(\varepsilon)) \le CF^\gamma. \tag{1.15}$$

Remark 1.2. If $d_1 < D$, then the second condition in (1.13) implies that

$$\frac{D^2 - d_1^2}{\theta^2} < T^2.$$

In other words, the observation time T needs to be sufficiently large. In this case, we can determine elastic parameters in the domain $\Omega(\varepsilon)$. The domain $\Omega(\varepsilon)$ is discussed in [9, section 3.4].

If Γ is the whole lateral boundary and T is sufficiently large, then a much stronger (and in a certain sense the best possible) Lipschitz stability estimate holds.

THEOREM 1.3. *Let $d_1 = d$. Assume that*

$$\frac{D^2 - d^2}{\theta^2} < T^2 < \frac{d^2}{\theta^2}. \tag{1.16}$$

Let the initial data $(\mathbf{u}_0(;j), \mathbf{u}_1(;j))$, $j = 1,2,3$, satisfy (1.12) *with $\lambda = \lambda_1, \mu = \mu_1$, and $\Gamma = \partial\Omega$.*

Then there exists ε_0 in (1.7) *and C such that for $E_1, E_2 \in \mathcal{E}_{\varepsilon_0,M}$ satisfying the conditions*

$$\rho_1 = \rho_2,\ R_1 = R_2,\ \partial^\alpha \lambda_1 = \partial^\alpha \lambda_2, \text{and } \partial^\alpha \mu_1 = \partial^\alpha \mu_2 \text{ on } \Gamma \text{ when } |\alpha| \le 1, \tag{1.17}$$

one has

$$\|\rho_1 - \rho_2\|_{(0)}(\Omega) + \|\lambda_1 - \lambda_2\|_{(0)}(\Omega) + \|\mu_1 - \mu_2\|_{(0)}(\Omega) + \|R_1 - R_2\|_{(0)}(\Omega) \le CF. \tag{1.18}$$

Remark 1.4. Condition (1.16) is needed for pseudoconvexity of weight function φ in Carleman estimates in the next sections and generally cannot be removed. Existence of T is guaranteed by the condition $D^2 < 2d^2$. Under the additional assumption $e \cdot \nabla c(x) \le 0, x \in \Omega$ for some direction e, the condition $D^2 < 2d^2$ can be achieved by using translation $x = y + Le$ with large L.

As mentioned previously, the proofs of these theorems rely on Carleman estimates. We briefly describe the needed Carleman estimates in section 2. Using this estimate we will prove in section 3 the Hölder stability estimate (1.15). In section 4, we derive the Lipschitz stability estimate for our inverse problem.

2. Carleman estimate. In this section we will collect Carleman estimates needed to solve our inverse problem. Their proofs can be found in [10] and [11]. Let $\psi(x,t) = |x|^2 - \theta^2 t^2 - d_1^2$ and $\varphi(x,t) = \exp(\frac{\eta}{2}\psi(x,t))$. Due to conditions (1.9) and (1.13) and known sufficient conditions of pseudoconvexity [9, Theorem 3.4.1], we can fix (large) $\eta > 0$ so that the phase function φ is strongly pseudoconvex on $\overline{Q(0)}$ with respect to

$$\frac{\rho}{\mu}\partial_t^2 - \Delta, \qquad \frac{\rho}{\lambda+2\mu}\partial_t^2 - \Delta.$$

Similarly, (1.9) and the second inequality in (1.16) guarantee strong pseudoconvexity of φ on $\bar{Q}$.

THEOREM 2.1. *There are constants ε_0 and C such that under the conditions of Theorem 1.3 for $E \in \mathcal{E}_{\varepsilon_0,M}$,*

(2.1)
$$\int_Q (\tau|\nabla_{x,t}\mathbf{u}|^2 + \tau|\nabla_{x,t}\mathrm{div}\,\mathbf{u}|^2 + \tau|\nabla_{x,t}\mathrm{curl}\,\mathbf{u}|^2 + \tau^3|\mathbf{u}|^2 + \tau^3|\mathrm{div}\,\mathbf{u}|^2 + \tau^3|\mathrm{curl}\,\mathbf{u}|^2)e^{2\tau\varphi}$$
$$\leq C \int_Q (|\mathbf{A}_E\mathbf{u}|^2 + |\nabla(\mathbf{A}_E\mathbf{u})|^2)e^{2\tau\varphi}$$

for all $\mathbf{u} \in H_0^3(Q)$, and under the conditions of Theorem 1.1,

$$\int_{Q(0)} (\tau^2|\mathbf{u}|^2 + |\mathrm{div}\,\mathbf{u}|^2 + |\mathrm{curl}\,\mathbf{u}|^2 + \tau^{-1}|\nabla\mathbf{u}|^2)e^{2\tau\varphi} \leq C \int_{Q(0)} |\mathbf{A}_E\mathbf{u}|^2 e^{2\tau\varphi} \tag{2.2}$$

for all $\mathbf{u} \in H_0^2(Q(0))$.

The Carleman estimates of Theorem 2.1 form our basic tool for treating the inverse problem. The basic idea in proving Theorem 2.1 is to reduce (1.3) to an extended system of dimension 7 for $(\mathbf{u}, \mathrm{div}\,\mathbf{u}, \mathrm{curl}\,\mathbf{u})$. The resulting new system is not principally diagonalized. However, when the residual stress R is small, the second derivatives of $\mathbf{u}$ can be bounded by first derivatives of $\mathrm{div}\,\mathbf{u}$ and $\mathrm{curl}\,\mathbf{u}$. We refer to [10] and [11] for detailed computations. For the case considered here, we need only verify the strong pseudoconvexity of φ on $\overline{Q}$ or on $\bar{Q}$. Under conditions (1.9), (1.13) or (1.16) one can check that φ satisfies the required property when ε_0 is small and η is large (see [9] or [10]). An estimate similar to (2.2) was also derived in [8].

In order to use (2.1), it is required that the Cauchy data of the solution and the source term vanish on the lateral boundary. To handle nonvanishing Cauchy data, the following lemma is useful.

LEMMA 2.2. *For any pair of $(\mathbf{g}_0, \mathbf{g}_1) \in H^{\frac{5}{2}}(\partial\Omega \times (-T,T)) \times H^{\frac{3}{2}}(\partial\Omega \times (-T,T))$, we can find a vector-valued function $\mathbf{u}^* \in H^3(Q)$ such that*

$$\mathbf{u}^* = \mathbf{g}_0, \quad \sigma(\mathbf{u}^*)\nu = \mathbf{g}_1, \quad \mathbf{A}_E\mathbf{u}^* = 0 \quad \textit{on} \quad \partial\Omega \times (-T,T),$$

and

$$\|\mathbf{u}^*\|_{(3)}(Q) \leq C(\|\mathbf{g}_0\|_{(\frac{5}{2})}(\partial\Omega \times (-T,T)) + \|\mathbf{g}_1\|_{(\frac{3}{2})}(\partial\Omega \times (-T,T))) \tag{2.3}$$

for some $C > 0$, provided ε_0 in (1.7) is sufficiently small.

Proof. By standard extension theorems for any $\mathbf{g}_2 \in H^{\frac{1}{2}}(\partial\Omega \times (-T,T))$ we can find $\mathbf{u}^{**} \in H^3(Q)$ so that

$$\mathbf{u}^{**} = \mathbf{g}_0, \quad \sigma(\mathbf{u}^{**})\nu = \mathbf{g}_1, \quad \partial_\nu^2\mathbf{u}^{**} = \mathbf{g}_2 \quad \text{on} \quad \partial\Omega \times (-T,T)$$

and

$$\|\mathbf{u}^{**}\|_{(3)}(Q) \le C(\|\mathbf{g}_2\|_{(\frac{1}{2})}(\partial\Omega \times (-T,T)) + \|\mathbf{g}_1\|_{(\frac{3}{2})}(\partial\Omega \times (-T,T)) \\ + \|\mathbf{g}_0\|_{(\frac{5}{2})}(\partial\Omega \times (-T,T))).$$

Since $\partial\Omega \times (-T,T)$ is noncharacteristic with respect to $\mathbf{A}_E$ provided (1.6) holds and ε_0 is small, the condition $\mathbf{A}_E\mathbf{u}^{**} = 0$ on $\partial\Omega \times (-T,T)$ is equivalent to the fact that $\mathbf{g}_2$ can be written as a linear combination (with C^1 coefficients) of $\partial_t^2\mathbf{g}_0$ and tangential derivatives of $\mathbf{g}_0$ (up to second order) and of $\mathbf{g}_1$ (up to first order) along $\partial\Omega$. In particular,

$$\|\mathbf{g}_2\|_{(\frac{1}{2})}(\partial\Omega \times (-T,T)) \le C(\|\mathbf{g}_1\|_{(\frac{3}{2})}(\partial\Omega \times (-T,T)) + \|\mathbf{g}_0\|_{(\frac{5}{2})}(\partial\Omega \times (-T,T))).$$

Choosing $\mathbf{g}_2$ as this linear combination, we obtain (2.3). ∎

To handle $\nabla\lambda$ and $\nabla\mu$ in (1.3), we need other Carleman estimates. We first derive the estimate needed in the proof of Theorem 1.1. Let d_1 be given as in Theorem 1.1. Then we can see that $\partial\Omega(\varepsilon) = (\Gamma \cup \{|x|^2 = d_1^2 + \varepsilon\}) \cap \bar{\Omega}$.

Lemma 2.3. *Let $f \in C^1(\overline{\Omega})$ satisfy $f|_\Gamma = 0$. Then*

$$\tau \int_{\Omega(\varepsilon)} |f(x)|^2 e^{2\tau\varphi(x,0)} dx \le C \int_{\Omega(\varepsilon)} |\nabla f(x)|^2 e^{2\tau\varphi(x,0)} dx. \tag{2.4}$$

Proof. Denote $\varphi_0(x) = \varphi(x,0)$. Let $g = e^{\tau\varphi_0} f$; then $e^{\tau\varphi_0}\nabla f = \nabla g - \tau\nabla\varphi_0 g$. Note that $g|_\Gamma = 0$. We observe that $\nabla\varphi_0(x) = \eta x\varphi_0(x)$, and thus on $\partial\Omega(\varepsilon) \setminus \Gamma$ with the unit outer normal $\nu(= -x/|x|)$,

$$\partial_\nu\varphi_0(x) = \nabla\varphi_0 \cdot \nu = -\eta|x|\varphi_0(x). \tag{2.5}$$

Using integration by parts and (2.5), we have that

$$\begin{aligned}
&\int_{\Omega(\varepsilon)} |\nabla g - \tau\nabla\varphi_0 g|^2 \\
&\quad = \int_{\Omega(\varepsilon)} |\nabla g|^2 + \tau^2 \int_{\Omega(\varepsilon)} |\nabla\varphi_0 g|^2 - 2\tau \int_{\Omega(\varepsilon)} \nabla g \cdot \nabla\varphi_0 g \\
&\quad \ge -\tau \int_{\Omega(\varepsilon)} \nabla\varphi_0 \cdot \nabla(g^2) \\
&\quad = -\tau \int_{\partial\Omega(\varepsilon)\setminus\Gamma} \partial_\nu\varphi_0 g^2 + \tau \int_{\Omega(\varepsilon)} \Delta\varphi_0 g^2 \\
&\quad = \tau \int_{\partial\Omega(\varepsilon)\setminus\Gamma} \eta|x|\varphi_0(x) g^2(x) d\Gamma(x) + \tau \int_{\Omega(\varepsilon)} (3\eta + \eta^2|x|^2)\varphi_0 g^2(x) dx \\
&\quad \ge C \int_{\Omega(\varepsilon)} g^2,
\end{aligned}$$

which implies (2.4). ∎

The following estimate is useful in proving Theorem 1.3 (see also [8, Lemma 3.6]).

Corollary 2.4. *Let $f \in C^1(\overline{\Omega})$ and $f = 0$ on $\partial\Omega$. Then we have*

$$\tau \int_\Omega |f(x)|^2 e^{2\tau\varphi(x,0)} dx \le C \int_\Omega |\nabla f(x)|^2 e^{2\tau\varphi(x,0)} dx.$$

3. Hölder stability for the determination of coefficients. In this section we prove the first main result of the paper, Theorem 1.1. Let us denote $\mathbf{u}(;j) = \mathbf{u}(;2;j) - \mathbf{u}(;1;j)$ for $j = 1,2,3$, and $\mathbf{F} = (f_1, f_2, \ldots, f_9, R)^\top$, where $f_1 = \rho_1 - \rho_2$, $f_2 = \lambda_1 - \lambda_2$, $f_3 = \mu_1 - \mu_2$, $(f_4, f_5, f_6)^\top = \nabla f_2$, $(f_7, f_8, f_9)^\top = \nabla f_3$, and

$$R^\top = \begin{pmatrix} r_{11} \\ r_{12} \\ r_{13} \\ r_{22} \\ r_{23} \\ r_{33} \end{pmatrix} = \begin{pmatrix} r_{1,11} - r_{2,11} \\ r_{1,12} - r_{2,12} \\ r_{1,13} - r_{2,13} \\ r_{1,22} - r_{2,22} \\ r_{1,23} - r_{2,23} \\ r_{1,33} - r_{2,33} \end{pmatrix}.$$

Subtracting equations (1.3) for $\mathbf{u}(;1;j)$ from the equations for $\mathbf{u}(;2;j)$ yields

$$\mathbf{A}_{E_2}\mathbf{u}(;j) = \mathcal{A}(\mathbf{u}(;1;j))\mathbf{F} \quad \text{on} \quad Q, \tag{3.1}$$

where

$$\begin{aligned} \mathcal{A}(\mathbf{v})\mathbf{F} = \quad & -f_1\partial_t^2\mathbf{v} + (f_2 + f_3)\nabla(\operatorname{div}\mathbf{v}) + f_3\Delta\mathbf{v} + \operatorname{div}\mathbf{v}(f_4, f_5, f_6)^\top \\ & + 2\epsilon(\mathbf{v})(f_7, f_8, f_9)^\top + \sum_{j,k=1}^{3} r_{jk}\partial_j\partial_k\mathbf{v} \end{aligned}$$

and

$$\mathbf{u}(;j) = \partial_t\mathbf{u}(;j) = 0 \quad \text{on} \quad \Omega \times \{0\}. \tag{3.2}$$

Differentiating (3.1) in t and using the time-independence of the coefficients of the system, we get

$$\mathbf{A}_{E_2}\mathbf{U}(;j) = \mathcal{A}(\mathbf{U}(;1;j))\mathbf{F} \quad \text{on} \quad Q, \tag{3.3}$$

where

$$\mathbf{U}(;j) = \begin{pmatrix} \partial_t^2\mathbf{u}(;j) \\ \partial_t^3\mathbf{u}(;j) \\ \partial_t^4\mathbf{u}(;j) \end{pmatrix}, \quad \mathbf{U}(;1;j) = \begin{pmatrix} \partial_t^2\mathbf{u}(;1;j) \\ \partial_t^3\mathbf{u}(;1;j) \\ \partial_t^4\mathbf{u}(;1;j) \end{pmatrix},$$

and

$$\mathcal{A}(\mathbf{U}(;1;j)) = \begin{pmatrix} \mathcal{A}(\partial_t^2\mathbf{u}(;1;j)) \\ \mathcal{A}(\partial_t^3\mathbf{u}(;1;j)) \\ \mathcal{A}(\partial_t^4\mathbf{u}(;1;j)) \end{pmatrix}.$$

By extension theorems for Sobolev spaces, there exists $\mathbf{U}^*(;j) \in H^2(Q)$ such that

$$\mathbf{U}^*(;j) = \mathbf{U}(;j),\ \sigma(\mathbf{U}^*(;j))\nu = \sigma(\mathbf{U}(;j))\nu \quad \text{on} \quad \Gamma \times (-T,T) \tag{3.4}$$

and

$$\|\mathbf{U}^*(;j)\|_{(2)}(Q) \le C(\|\mathbf{U}(;j)\|_{(\frac{3}{2})}(\Gamma \times (-T,T)) + \|\sigma(\mathbf{U})(;j)\nu\|_{(\frac{1}{2})}(\Gamma \times (-T,T))) \le CF \tag{3.5}$$

for all $j = 1,2,3$ due to the definitions of $\mathbf{u}(;j), \mathbf{U}(;j)$, and F. We now introduce $\mathbf{V}(;j) = \mathbf{U}(;j) - \mathbf{U}^*(;j)$. Then

$$\mathbf{A}_{E_2}\mathbf{V}(;j) = \mathcal{A}(\mathbf{U}(;1;j))\mathbf{F} - \mathbf{A}_{E_2}\mathbf{U}^*(;j) \quad \text{on} \quad Q \tag{3.6}$$

and

$$\mathbf{V}(;j) = \sigma(\mathbf{V})(;j)\nu = 0 \quad \text{on} \quad \Gamma \times (-T,T). \tag{3.7}$$

To use the Carleman estimate (2.2), we introduce a cut-off function $\chi \in C^2(\mathbb{R}^4)$ such that $0 \le \chi \le 1$, $\chi = 1$ on $Q(\frac{\varepsilon}{2})$, and $\chi = 0$ on $Q \setminus Q(0)$. By the Leibniz formula,

$$\mathbf{A}_{E_2}(\chi\mathbf{V}(;j)) = \chi\mathbf{A}_{E_2}(\mathbf{V}(;j)) + \mathbf{A}_1\mathbf{V}(;j) = \chi\mathcal{A}\mathbf{F} - \chi\mathbf{A}_{E_2}\mathbf{U}^*(;j) + \mathbf{A}_1\mathbf{V}(;j)$$

due to (3.6). Here (and below) $\mathbf{A}_1$ denotes a first order matrix differential operator with coefficients uniformly bounded by $C(\varepsilon)$. By the choice of χ, $\mathbf{A}_1\mathbf{V}(;j) = 0$ on $Q(\frac{\varepsilon}{2})$. It is not hard to see that (3.7) implies that $\mathbf{V}(;j) = \partial_\nu\mathbf{V}(;j) = 0$ on $\Gamma\times(-T,T)$. Hence due to the first condition of (1.13), the function $\chi\mathbf{V}(;j) \in H_0^2(Q(0))$ (see, for example [4, Corollary 1.5.1.6, p. 39]). Observe, that $\chi = 0$ is zero near the non-C^8-smooth part of $\partial Q(0)$, and therefore we can use results for C^8-smooth domains by slightly extending $Q(0)$. So we can apply to $\chi\mathbf{V}(;j)$ the Carleman estimate (2.2) to get

$$\int_Q \tau|\chi\mathbf{V}(;j)|^2 e^{2\tau\varphi}$$

$$\le C\int_Q (|\mathbf{F}|^2 + |\mathbf{A}_{E_2}(\mathbf{U}^*(;j))|^2)e^{2\tau\varphi} + C\int_{Q\setminus Q(\frac{\varepsilon}{2})} |\mathbf{A}_1\mathbf{V}(;j)|^2 e^{2\tau\varphi}$$

$$\le C\left(\int_Q |\mathbf{F}|^2 e^{2\tau\varphi} + F^2 e^{2\tau\Phi} + C(\varepsilon)e^{2\tau\varepsilon_1}\right), \tag{3.8}$$

where $\Phi = \sup\varphi$ over Q and $\varepsilon_1 = e^{\frac{\eta\varepsilon}{4}}$. To get the last inequality, we used the bounds (3.5) and (1.10).

On the other hand, from (1.3), (3.1), (3.2) we have

$$\rho_2\partial_t^2\mathbf{u}(;j) = \mathcal{A}(\mathbf{u}(;1;j))\mathbf{F},$$

$$\rho_2\partial_t^3\mathbf{u}(;j) = \mathcal{A}(\partial_t\mathbf{u}(;1;j))\mathbf{F}$$

on $\Omega\times\{0\}$. We now want to rearrange the formulas above. Let $\mathbf{a}_{kj} = -\partial_t^{2+k}\mathbf{u}(0;1;j)$, $\mathbf{b}_{kj} = \nabla(\operatorname{div}\mathbf{u}_k(;j))$, $\mathbf{c}_{kj} = \Delta\mathbf{u}_k(;j)+\nabla\operatorname{div}\mathbf{u}_k(;j)$, $\mathbf{B}_{kj} = \operatorname{div}\mathbf{u}_k(;j)$, $\mathbf{C}_{kj} = 2\epsilon(\mathbf{u}_k(;j))$, and $\mathbf{R}_{kj} = \mathbf{R}(\mathbf{u}_k(;j))$ (see the definition of $\mathbf{R}$ in (1.11)), where $k = 0, 1$ and $j = 1, 2, 3$. Using that $\mathbf{u}(;1;j) = \mathbf{u}_0(;j), \partial_t\mathbf{u}(;1;j) = \mathbf{u}_1(;j)$ on $\Omega\times\{0\}$, we have

$$\begin{pmatrix} \mathbf{a}_{01} & \mathbf{B}_{01}I_3 & \mathbf{C}_{01} & \mathbf{R}_{01} \\ \mathbf{a}_{11} & \mathbf{B}_{11}I_3 & \mathbf{C}_{11} & \mathbf{R}_{11} \\ \mathbf{a}_{02} & \mathbf{B}_{02}I_3 & \mathbf{C}_{02} & \mathbf{R}_{02} \\ \mathbf{a}_{12} & \mathbf{B}_{12}I_3 & \mathbf{C}_{12} & \mathbf{R}_{12} \\ \mathbf{a}_{03} & \mathbf{B}_{03}I_3 & \mathbf{C}_{03} & \mathbf{R}_{03} \\ \mathbf{a}_{13} & \mathbf{B}_{13}I_3 & \mathbf{C}_{13} & \mathbf{R}_{13} \end{pmatrix} \begin{pmatrix} f_1 \\ f_4 \\ \vdots \\ f_9 \\ r_{11} \\ \vdots \\ r_{33} \end{pmatrix} = \rho_2 \begin{pmatrix} \partial_t^2\mathbf{u}(,0;1) \\ \partial_t^3\mathbf{u}(,0;1) \\ \partial_t^2\mathbf{u}(,0;2) \\ \partial_t^3\mathbf{u}(,0;2) \\ \partial_t^2\mathbf{u}(,0;3) \\ \partial_t^3\mathbf{u}(,0;3) \end{pmatrix} - \begin{pmatrix} \mathbf{b}_{01} & \mathbf{c}_{01} \\ \mathbf{b}_{11} & \mathbf{c}_{11} \\ \mathbf{b}_{02} & \mathbf{c}_{02} \\ \mathbf{b}_{12} & \mathbf{c}_{12} \\ \mathbf{b}_{03} & \mathbf{c}_{03} \\ \mathbf{b}_{13} & \mathbf{c}_{13} \end{pmatrix} \begin{pmatrix} f_2 \\ f_3 \end{pmatrix} \tag{3.9}$$

on Ω. From the system (1.3) at $t=0$ and from this system differentiated in t and taken at $t=0$, we obtain

$$
\begin{aligned}
\mathbf{a}_{kj} &= -\frac{\mu_1}{\rho_1}\Delta\mathbf{u}_k(;j) - \frac{\lambda_1+\mu_1}{\rho_1}\nabla(\operatorname{div}\mathbf{u}_k(;j)) - \operatorname{div}\mathbf{u}_k(;j)\frac{\nabla\lambda_1}{\rho_1}\\
&\quad -2\epsilon(\mathbf{u}_k(;j))\frac{\nabla\mu_1}{\rho_1} - \sum_{\ell,m=1}^{3} r_{1,\ell m}\partial_\ell\partial_m\mathbf{u}_k(;j)\\
&= -\frac{\mu_1}{\rho_1}\Delta\mathbf{u}_k(;j) - \frac{\lambda_1+\mu_1}{\rho_1}\nabla(\operatorname{div}\mathbf{u}_k(;j)) - \mathbf{B}_{kj}\frac{\nabla\lambda_1}{\rho_1}\\
&\quad -\mathbf{C}_{kj}\frac{\nabla\mu_1}{\rho_1} - \sum_{\ell,m=1}^{3} r_{1,\ell m}\partial_\ell\partial_m\mathbf{u}_k(;j)
\end{aligned}
\tag{3.10}
$$

when $k=0,1$ and $j=1,2,3$.

We now consider the matrix on the left-hand side of (3.9). Using (3.10), one can add to the first column the remaining columns multiplied by suitable factors such that $-\operatorname{div}\mathbf{u}_k(;j)\frac{\nabla\lambda_1}{\rho_1}$, $-2\epsilon(\mathbf{u}_k(;j))\frac{\nabla\mu_1}{\rho_1}$, and $-\sum_{\ell,m=1}^{3} r_{1,\ell m}\partial_\ell\partial_m\mathbf{u}_k(;j)$ are eliminated from the first column of this matrix. Then we multiply the first column of the new matrix by $-\rho_1$. We end up with the matrix $\mathbf{M}$ defined in section 1. Obviously, determinants of corresponding minors of the matrix on the left side of (3.9) and of the matrix $\mathbf{M}$ are the same. It follows from condition (1.12) and bounds (1.10) that

$$
|\mathbf{F}|^2 \le C\left(\sum_{j=1}^{3}\sum_{\beta=2}^{3}|\partial_t^\beta\mathbf{u}(0;j)|^2 + |f_2|^2 + |f_3|^2\right) \quad \text{on} \quad \overline{\Omega}. \tag{3.11}
$$

Since $\chi(\cdot,T)=0$,

$$
\int_\Omega |\chi\partial_t^\beta\mathbf{u}(;j)|^2(x,0)e^{2\tau\varphi(x,0)}dx = -\int_0^T \partial_t\left(\int_\Omega |\chi\partial_t^\beta\mathbf{u}(;j)|^2(x,t)e^{2\tau\varphi(x,t)}dx\right)dt
$$

$$
\le \int_Q 2\chi^2(|\partial_t^{\beta+1}\mathbf{u}(;j)||\partial_t^\beta\mathbf{u}(;j)| + \tau|\partial_t\varphi||\partial_t^\beta\mathbf{u}(;j)|^2)e^{2\tau\varphi} + 2\int_{Q\setminus Q(\frac{\varepsilon}{2})}|\partial_t^\beta\mathbf{u}(;j)|^2\chi|\partial_t\chi|e^{2\tau\varphi},
$$

where $\beta=2,3$. The right side does not exceed

$$
C\left(\int_Q \tau|\chi\mathbf{U}(;j)|^2e^{2\tau\varphi} + C(\varepsilon)\int_{Q\setminus Q(\frac{\varepsilon}{2})}|\mathbf{U}(;j)|^2e^{2\tau\varphi}\right)
$$

$$
\le C\left(\int_Q \tau|\chi\mathbf{V}(;j)|^2e^{2\tau\varphi} + C(\varepsilon)\int_{Q\setminus Q(\frac{\varepsilon}{2})}|\mathbf{U}(;j)|^2e^{2\tau\varphi} + \tau\int_Q|\mathbf{U}^*(;j)|^2e^{2\tau\varphi}\right)
$$

because $\mathbf{U}(;j)=\mathbf{V}(;j)+\mathbf{U}^*(;j)$. Using that $\chi=1$ on $\Omega(\frac{\varepsilon}{2})$, $\varphi<\varepsilon_1$ on $Q\setminus Q(\frac{\varepsilon}{2})$, and $\varphi<\Phi$ on Q from these inequalities, from (3.8), (3.5), and (1.10) we set

$$
\int_{\Omega(\frac{\varepsilon}{2})}|\partial_t^\beta\mathbf{u}(0;j)|^2e^{2\tau\varphi(,0)} \le C\left(\int_Q|\mathbf{F}|^2e^{2\tau\varphi} + C(\varepsilon)e^{2\tau\varepsilon_1} + \tau e^{2\tau\Phi}F^2\right) \tag{3.12}
$$

for $\beta=2,3$ and $j=1,2,3$. Using that $\chi=1$ on $\Omega(\frac{\varepsilon}{2})$, from (3.11) and (3.12) we obtain

$$\int_{\Omega(\frac{\varepsilon}{2})} |\mathbf{F}|^2 e^{2\tau\varphi(,0)} \leq C\left(\int_{Q(\frac{\varepsilon}{2})} |\mathbf{F}|^2 e^{2\tau\varphi} + \tau e^{2\tau\Phi} F^2 + C(\varepsilon) e^{2\tau\varepsilon_1} + \int_{\Omega(\frac{\varepsilon}{2})} (|f_2|^2 + |f_3|^2) e^{2\tau\varphi(,0)} \right), \tag{3.13}$$

where we also split Q in the right side of (3.12) into $Q(\frac{\varepsilon}{2})$ and its complement and used that $|\mathbf{F}| \leq C$ and $\varphi < \varepsilon_1$ on the complement.

To eliminate the first integral in the right side of (3.13), we observe that

$$\int_{Q(\frac{\varepsilon}{2})} |\mathbf{F}|^2(x) e^{2\tau\varphi(x,t)} dxdt \leq \int_{\Omega(\frac{\varepsilon}{2})} |\mathbf{F}|^2(x) e^{2\tau\varphi(x,0)} \left(\int_{-T}^{T} e^{2\tau(\varphi(x,t)-\varphi(x,0))} dt \right) dx.$$

Due to our choice of function φ, we have $\varphi(x,t) - \varphi(x,0) < 0$ when $t \neq 0$. Hence by the Lebesgue theorem the inner integral (with respect to t) converges to 0 as τ goes to infinity. By reasons of continuity of φ, this convergence is uniform with respect to $x \in \Omega$. Choosing $\tau > C$ we therefore can absorb the integral over $Q(\frac{\varepsilon}{2})$ in the right side of (3.13) by the left side and arrive at the inequality

$$\int_{\Omega(\frac{\varepsilon}{2})} |\mathbf{F}|^2 e^{2\tau\varphi(,0)} \leq C \left(\tau e^{2\tau\Phi} F^2 + C(\varepsilon) e^{2\tau\varepsilon_1} + \int_{\Omega(\frac{\varepsilon}{2})} (|f_2|^2 + |f_3|^2) e^{2\tau\varphi(,0)} \right). \tag{3.14}$$

On the other hand, to eliminate the last integral on the right side of (3.14), we use Lemma 2.3 with condition (1.14) to get

$$\int_{\Omega(\frac{\varepsilon}{2})} (|f_2|^2 + |f_3|^2) e^{2\tau\varphi(,0)} \leq \frac{C}{\tau} \int_{\Omega(\frac{\varepsilon}{2})} (|\nabla f_2|^2 + |\nabla f_3|^2) e^{2\tau\varphi(,0)}. \tag{3.15}$$

Using (3.15) with large τ and the inequality $\tau \leq e^{\tau}$, we absorb the last integral in the right side of (3.14) into the left side and obtain

$$\int_{\Omega(\frac{\varepsilon}{2})} |\mathbf{F}|^2 e^{2\tau\varphi(,0)} \leq C(e^{2\tau(\Phi_1+1)} F^2 + C(\varepsilon) e^{2\tau\varepsilon_1}).$$

Letting $\varepsilon_2 = e^{\frac{\eta\varepsilon}{2}} \leq \varphi$ on $\Omega(\varepsilon)$ and dividing both parts by $e^{2\tau\varepsilon_2}$ yields

$$\int_{\Omega(\varepsilon)} |\mathbf{F}|^2 \leq C(\tau e^{2\tau(\Phi+1-\varepsilon_2)} F^2 + e^{-2\tau(\varepsilon_2-\varepsilon_1)}) \leq C(\varepsilon)(e^{2\tau(\Phi+1)} F^2 + e^{-2\tau(\varepsilon_2-\varepsilon_1)}) \tag{3.16}$$

since $\tau e^{-2\tau\varepsilon_2} < C(\varepsilon)$. If $\frac{1}{C} \leq F$, then bound (1.15) is obvious because the left side in (1.15) is less than C. So to prove (1.15) it suffices to assume that $F < \frac{1}{C}$. Then $\tau = \frac{-logF}{\Phi+1+\varepsilon_2-\varepsilon_1} > C$, and we can use this τ in (3.16). Due to the choice of τ,

$$e^{-2\tau(\varepsilon_2-\varepsilon_1)} = e^{2\tau(\Phi+1)} F^2 = F^{2\frac{\varepsilon_2-\varepsilon_1}{\Phi+1+\varepsilon_2-\varepsilon_1}},$$

and from (3.16) we obtain (1.15) with $\gamma = \frac{\varepsilon_2-\varepsilon_1}{\Phi+1+\varepsilon_2-\varepsilon_1}$. The proof of Theorem 1.1 is now complete.

4. Lipschitz stability for the determination of coefficients. In this section we will prove Theorem 1.3. The key ingredient is the following Lipschitz stability estimate for the Cauchy problem for the system $\mathbf{A}_E\mathbf{u} = \mathbf{f}$.

THEOREM 4.1. *Suppose that* Ω *and* T *satisfy the assumptions of Theorem* 1.3. *Let* $\mathbf{u} \in (H^3(Q))^3$ *solve the Cauchy problem*

$$\begin{cases} \mathbf{A}_E\mathbf{u} = \mathbf{f} & \text{in} \quad Q, \\ \mathbf{u} = \sigma_\nu(\mathbf{u}) = 0 & \text{on} \quad \partial\Omega \times (-T, T) \end{cases} \tag{4.1}$$

with $\mathbf{f} \in L^2((-T,T); H^1(\Omega))$ *and* $\mathbf{f} = 0$ *on* $\partial\Omega \times (-T,T)$. *Furthermore, assume that* (1.7) *holds for sufficiently small* ε_0.

Then there exists a constant $C > 0$ *such that*

$$\|\mathbf{u}\|^2_{H^1(Q)} + \|\operatorname{div}\mathbf{u}\|^2_{H^1(Q)} + \|\operatorname{curl}\mathbf{u}\|^2_{H^1(Q)} \le C\|\mathbf{f}\|^2_{L^2((-T,T);H^1(\Omega))}. \tag{4.2}$$

This estimate was proved in [11].

By virtue of (4.2) and an equivalence of the norms $\|\mathbf{u}\|_{(1)}(\Omega)$ and of

$$\|\operatorname{div}\mathbf{u}\|_{(0)}(\Omega) + \|\operatorname{curl}\mathbf{u}\|_{(0)}(\Omega) + \|\mathbf{u}\|_{(0)}(\Omega)$$

in $H^1_0(\Omega)$ (e.g., [3, pp. 358–359]), it is not hard to derive the following.

COROLLARY 4.2. *Under the conditions of Theorem* 4.1,

$$\|\mathbf{u}\|_{(0)}(Q) + \|\nabla_{x,t}\mathbf{u}\|_{(0)}(Q) + \|\partial_t\nabla\mathbf{u}\|_{(0)}(Q) \le C\|\mathbf{f}\|_{L^2((-T,T);H^1(\Omega))}. \tag{4.3}$$

Now we are ready to prove Theorem 1.3. We will use the notations in section 3. Recall that

$$\mathbf{A}_{E_2}\mathbf{U}(;1;j) = \mathcal{A}(\mathbf{U}(;1;j))\mathbf{F},$$

where

$$\begin{aligned} \mathcal{A}(\mathbf{U}(;1;j))\mathbf{F} \quad = \quad & -f_1\partial_t^2\mathbf{U}(;1;j) + (f_2+f_3)\nabla(\operatorname{div}\mathbf{U}(;1;j)) + f_3\Delta\mathbf{U}(;1;j) \\ & + \operatorname{div}\mathbf{U}(;1;j)(f_4,f_5,f_6)^\top + 2\epsilon(\mathbf{U}(;1;j))(f_7,f_8,f_9)^\top \\ & + \sum_{j,k=1}^{3} r_{jk}\partial_j\partial_k\mathbf{U}(;1;j). \end{aligned}$$

So, from (1.17) we have

$$\mathbf{A}_{E_2}\mathbf{U}(;j) = 0 \quad \text{on} \quad \partial\Omega \times (-T,T). \tag{4.4}$$

Furthermore, in view of Lemma 2.2, there exists $\mathbf{U}^*(;j) \in H^3(Q)$ such that

$$\mathbf{U}^*(;j) = \mathbf{U}(;j), \quad \sigma(\mathbf{U}^*(;j))\nu = \sigma(\mathbf{U}(;j))\nu, \quad \mathbf{A}_{E_2}\mathbf{U}^*(;j) = 0 \quad \text{on} \quad \partial\Omega \times (-T,T) \tag{4.5}$$

and

$$\|\mathbf{U}^*(;j)\|_{(3)}(Q) \le C(\|\mathbf{U}(;j)\|_{(\frac{5}{2})}(\partial\Omega\times(-T,T)) + \|\sigma(\mathbf{U})(;j)\nu\|_{(\frac{3}{2})}(\partial\Omega\times(-T,T))) \le CF \tag{4.6}$$

due to the definition of F. As before, we set $\mathbf{V}(;j) = \mathbf{U}(;j) - \mathbf{U}^*(;j)$. Due to (4.4) and (4.5), we get

$$\mathbf{V}(;j) = \sigma(\mathbf{V})(;j)\nu = 0, \quad \mathbf{A}_{E_2}\mathbf{V}(;j) = 0 \quad \text{on} \quad \partial\Omega \times (-T,T). \tag{4.7}$$

With (4.7), applying Corollary 4.2 to (3.6), (3.7) and using (4.6) gives

$$(4.8)\quad \|\mathbf{V}(;j)\|^2_{(0)}(Q)+\|\nabla_{x,t}\mathbf{V}(;j)\|^2_{(0)}(Q)+\|\partial_t\nabla\mathbf{V}(;j)\|^2_{(0)}(Q)\le C(\|\mathbf{F}\|_{(1)}(\Omega)^2+F^2)$$

for $j=1,2,3$.

On the other hand, as in the proof of Theorem 1.1 we will bound the right side of (4.8) by $\mathbf{V}$. To use the Carleman estimate (2.1) we need to cut off $\mathbf{V}(;j)$ near $t=T$ and $t=-T$. We first observe that from the definition,

$$1\le\varphi(x,0),\qquad x\in\Omega,$$

and from condition (1.16),

$$\varphi(x,T)=\varphi(x,-T)<1\quad\text{when}\quad x\in\overline{\Omega}.$$

So there exists a $\delta>\frac{1}{C}$ such that

$$(4.9)\qquad 1-\delta<\varphi\quad\text{on}\quad\Omega\times(0,\delta),\qquad\varphi<1-2\delta\quad\text{on}\quad\Omega\times(T-2\delta,T).$$

We now choose a smooth cut-off function $0\le\chi_0(t)\le 1$ such that $\chi_0(t)=1$ for $-T+2\delta<t<T-2\delta$ and $\chi(t)=0$ for $|t|>T-\delta$. As in the argument before (3.8) (see also Lemma A.1 in [13]), from (4.7) we derive that $\mathbf{V}(;j)=\partial_\nu\mathbf{V}(;j)=0$ on $\partial\Omega\times(-T,T)$. Then since $\partial\Omega\times(-T,T)$ is not characteristic with respect to $\mathbf{A}_E$ the third equation in (4.7) implies that $\partial_\nu^2\mathbf{V}(;j)=0$ on $\partial\Omega\times(-T,T)$. Summing up, $\mathbf{V}(;j)=\partial_\nu\mathbf{V}(;j)=\partial_\nu^2\mathbf{V}(;j)=0$ on $\partial\Omega\times(-T,T)$. Now from known results about traces in Sobolev spaces [4], as above we conclude that $\chi_0\mathbf{V}(;j)\in H_0^3(Q)$. Using the Leibniz formula

$$\mathbf{A}_{E_2}(\chi_0\mathbf{V}(;j))=\chi_0\mathcal{A}(\mathbf{U}(;1;j))\mathbf{F}-\chi_0\mathbf{A}_{E_2}\mathbf{U}^*(;j)+2\rho_2(\partial_t\chi_0)\partial_t\mathbf{V}(;j)+\rho_2(\partial_t^2\chi_0)\mathbf{V}(;j)$$

and Carleman estimate (2.1) yields

$$\int_Q\chi_0^2(\tau^3|\mathbf{V}(;j)|^2+\tau|\nabla\mathbf{V}(;j)|^2)e^{2\tau\varphi}$$

$$\le C\left(\int_Q(|\mathbf{F}|^2+|\nabla\mathbf{F}|^2+|\mathbf{A}_{E_2}\mathbf{U}^*(;j)|^2+|\nabla(\mathbf{A}_{E_2}\mathbf{U}^*)(;j)|^2)e^{2\tau\varphi}\right.$$

$$\left.+\int_{\Omega\times\{T-2\delta<|t|<T\}}(|\mathbf{V}(;j)|^2+|\nabla_{x,t}\mathbf{V}(;j)|^2+|\partial_t\nabla\mathbf{V}(;j)|^2\right)e^{2\tau\varphi})$$

$$\le C\left(\int_Q(|\mathbf{F}|^2+|\nabla\mathbf{F}|^2)e^{2\tau\varphi}+e^{2\tau\Phi}F^2+e^{2\tau(1-2\delta)}\int_\Omega(|\mathbf{F}|^2+|\nabla\mathbf{F}|^2)\right),$$

where we let $\Phi = \sup_Q \varphi$ and used (4.6), (4.8), (4.9). Since $\mathbf{U}(;j) = \mathbf{V}(;j) + \mathbf{U}^*(;j)$, from (4.6) we obtain

$$\int_Q \chi_0^2(\tau^3|\mathbf{U}(;j)|^2 + \tau|\nabla\mathbf{U}(;j)|^2)e^{2\tau\varphi}$$

$$\leq C\left(\tau^3 e^{2\tau\Phi}F^2 + \int_\Omega \left(\int_{-T}^T e^{2\tau\varphi(x,t)}dt + e^{2\tau(1-2\delta)}\right)(|\mathbf{F}|^2 + |\nabla\mathbf{F}|^2)(x)\right)dx. \tag{4.10}$$

Utilizing (3.2) and (1.12), similarly to deriving (3.11), we get from (3.9) that

$$|\mathbf{F}|^2 + |\nabla\mathbf{F}|^2 \leq C\left(\sum_{j=1}^3 \sum_{\substack{\beta=2,3;\\ k=0,1}} |\partial_t^\beta \nabla^k \mathbf{u}(0;j)|^2 + \sum_{k=0,1}(|\nabla^k f_2|^2 + |\nabla^k f_3|^2)\right). \tag{4.11}$$

Therefore, by (4.11) and Corollary 2.4 (with conditions (1.17) for Lamé coefficients), we have

$$\int_\Omega (|\mathbf{F}|^2 + |\nabla\mathbf{F}|^2)e^{2\tau\varphi(,0)}$$

$$\leq C\left(\int_\Omega \sum_{j=1}^3 \sum_{\substack{\beta=2,3;\\ k=0,1}} |\partial_t^\beta \nabla^k \mathbf{u}(0;j)|^2 e^{2\tau\varphi(,0)} + \int_\Omega \sum_{k=0,1}(|\nabla^k f_2|^2 + |\nabla^k f_3|^2)e^{2\tau\varphi(,0)}\right)$$

$$\leq -C\int_0^T \partial_t\left(\int_\Omega \sum_{j=1}^3 \sum_{\substack{\beta=2,3;\\ k=0,1}} \chi_0^2 |\partial_t^\beta \nabla^k \mathbf{u}(;j)|^2(x,t)e^{2\tau\varphi(x,t)}dx\right)dt$$
$$+\frac{C}{\tau}\int_\Omega (|\mathbf{F}|^2 + |\nabla\mathbf{F}|^2)e^{2\tau\varphi(;0)}.$$

Choosing τ large, we eliminate the last term and obtain

$$\int_\Omega (|\mathbf{F}|^2 + |\nabla\mathbf{F}|^2)e^{2\tau\varphi(,0)}$$

$$\leq C\int_Q \chi_0^2 \sum_{j=1}^3 \sum_{\substack{\beta=2,3;\\ k=0,1}} (|\partial_t^\beta \nabla^k \mathbf{u}(;j)||\partial_t^{\beta+1}\nabla^k \mathbf{u}(;j)| + \tau|\partial_t\varphi||\partial_t^\beta \nabla^k \mathbf{u}(;j)|^2)e^{2\tau\varphi}$$

$$+C\int_{\Omega\times(T-2\delta,T)} \chi_0|\partial_t\chi_0| \sum_{j=1}^3 \sum_{\beta=2,3;k=0,1} |\partial_t^\beta \nabla^k \mathbf{u}(;j)|^2 e^{2\tau\varphi}.$$

Now as in the proofs of section 3, the right side is less than

$$C\left(\int_Q \tau\chi_0^2(|\mathbf{U}(;j)|^2+|\nabla\mathbf{U}(;j)|^2)e^{2\tau\varphi}+\int_{\Omega\times(T-2\delta,T)}(|\mathbf{U}(;j)|^2+|\nabla\mathbf{U}(;j)|^2)e^{2\tau\varphi}\right)$$

$$\leq C\left(\int_Q \tau\chi_0^2(|\mathbf{U}(;j)|^2+|\nabla\mathbf{U}(;j)|^2)e^{2\tau\varphi}+e^{2\tau(1-2\delta)}(\|\mathbf{F}\|^2_{(1)}(\Omega)+F^2)\right),$$

where we used equality $\mathbf{U}(;j)=\mathbf{U}^*(;j)+\mathbf{V}(;j)$ and (4.6), (4.8). From the two previous bounds and (4.10) we conclude that

$$\int_\Omega(|\mathbf{F}|^2+|\nabla\mathbf{F}|^2)e^{2\tau\varphi(,0)} \leq C\left(\tau^3e^{2\tau\Phi}F^2+\int_\Omega\left(\int_{-T}^{T}e^{2\tau\varphi(,t)}dt+e^{2\tau(1-2\delta)}\right)(|\mathbf{F}|^2+|\nabla\mathbf{F}|^2)\right). \tag{4.12}$$

Due to our choice of φ, $1\leq\varphi(,0), \varphi(,t)-\varphi(,0)<0$ when $t\neq 0$. Thus by the Lebesgue theorem as in the proofs of section 3, we have

$$2C\left(\int_{-T}^{T}e^{2\tau\varphi(,t)}dt+e^{2\tau(1-\delta)}\right)\leq e^{2\tau\varphi(,0)}$$

uniformly on Ω when $\tau>C$. Hence choosing and fixing such large τ, we eliminate the second term on the right side of (4.12). The proof of Theorem 1.3 is now complete.

5. Conclusion. While natural in some applications, the assumption about the smallness of residual stress is restrictive. In our opinion it can be relaxed by using the methods of papers [8], [11], and this paper. More restrictive and much more difficult to remove is the condition that the initial data are not zero. At present, even for scalar isotropic hyperbolic equations, global uniqueness of the speed of propagation or of the potential from few lateral boundary measurements is an open and outstanding research problem (see, for example, [9]). Moreover, in the case of zero initial data, general anisotropic hyperbolic operators (and hence systems) cannot be uniquely determined by all lateral boundary measurements (Dirichlet-to-Neumann map). In fact, a large gauge transformation group changes equations inside Ω without affecting the lateral boundary data. Hence (special) nonzero initial data are necessary for the complete identification of such equations and systems.

Of substantial interest is uniqueness in inverse problems for more general anisotropic systems, for example, for dynamical elasticity systems with transversal isotropy. For such systems there are no Carleman estimates or uniqueness of the continuation results. On the other hand, such systems are quite important for applications to geophysics, material science, and medicine, and they are notorious mathematical challenges.

REFERENCES

[1] A. L. Bukhgeim, *Introduction to the Theory of Inverse Problems*, VSP, Utrecht, The Netherlands, 2000.

[2] A. L. Bukhgeim and M. V. Klibanov, *Global uniqueness of a class of multidimensional inverse problems*, Soviet Math. Dokl., 24 (1981), pp. 244–247.

[3] G. Duvaut and J.-L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, New York, 1976.

[4] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[5] S. Hansen and G. Uhlmann, *Propagation of polarization in elastodynamics with residual stress and travel times*, Math. Ann., 326 (2003), pp. 536–587.

[6] O. Imanuvilov and M. Yamamoto, *Determination of a coefficient in an acoustic equation with a single measurement*, Inverse Problems, 19 (2003), pp. 157–171.

[7] O. Imanuvilov and M. Yamamoto, *Carleman estimates for the nonstationary Lamé system and the application to an inverse problem*, ESAIM Control Optim. Calc. Var., 11 (2005), pp. 1–56.

[8] O. Imanuvilov, V. Isakov, and M. Yamamoto, *An inverse problem for the dynamical Lamé system with two sets of boundary data*, Comm. Pure Appl. Math., 56 (2003), pp. 1–17.

[9] V. Isakov, *Inverse Problems for Partial Differential Equations*, 2nd ed., Springer-Verlag, New York, 2006.

[10] V. Isakov, G. Nakamura, and J.-N. Wang, *Uniqueness and stability in the Cauchy problem for the elasticity system with residual stress*, in Inverse Problems: Theory and Applications, Contemp. Math. 333, AMS, Providence, RI, 2003, pp. 99–113.

[11] V. Isakov, J.-N. Wang, and M. Yamamoto, *Uniqueness and stability of determining the residual stress by one measurement*, Comm. Partial Differential Equations, 32 (2007), pp. 833–848.

[12] M. V. Klibanov and A. Timonov, *Carleman Estimates for Coefficient Inverse Problems and Numerical Applications*, VSP, Utrecht, The Netherlands, 2004.

[13] C.-L. Lin and J.-N. Wang, *Uniqueness in inverse problems for an elasticity system with residual stress*, Inverse Problems, 19 (2003), pp. 807–820.

[14] C.-S. Man, *Hartig's law and linear elasticity with initial stress*, Inverse Problems, 14 (1998), pp. 313–320.

[15] L. Rachelle, *Uniqueness in inverse problems for elastic media with residual stress*, Comm. Partial Differential Equations, 28 (2003), pp. 1787–1806.

SIAM J. MATH. ANAL.
Vol. 39, No. 4, pp. 1344–1365

EXISTENCE AND UNIQUENESS OF GLOBAL STRONG SOLUTIONS FOR ONE-DIMENSIONAL COMPRESSIBLE NAVIER–STOKES EQUATIONS*

A. MELLET† AND A. VASSEUR†

Abstract. We consider Navier–Stokes equations for compressible viscous fluids in one dimension. It is a well-known fact that if the initial datum are smooth and the initial density is bounded by below by a positive constant, then a strong solution exists locally in time. In this paper, we show that under the same hypothesis, the density remains bounded by below by a positive constant uniformly in time, and that strong solutions therefore exist globally in time. Moreover, while most existence results are obtained for positive viscosity coefficients, the present result holds even if the viscosity coefficient vanishes with the density. Finally, we prove that the solution is unique in the class of weak solutions satisfying the usual entropy inequality. The key point of the paper is a new entropy-like inequality introduced by Bresch and Desjardins for the shallow water system of equations. This inequality gives additional regularity for the density (provided such regularity exists at initial time).

Key words. Navier–Stokes, compressible, nonconstant viscosity coefficients, strong solutions, uniqueness, mono-dimensional

AMS subject classifications. 76N10, 35Q30

DOI. 10.1137/060658199

1. Introduction.

This paper is devoted to the existence of global strong solutions of the following Navier–Stokes equations for compressible isentropic flow:

$$\rho_t + (\rho u)_x = 0, \tag{1.1}$$

$$(\rho u)_t + (\rho u^2)_x + p(\rho)_x = (\mu(\rho)\, u_x)_x, \qquad (x,t) \in \mathbb{R} \times \mathbb{R}_+, \tag{1.2}$$

with possibly degenerate viscosity coefficient.

Throughout the paper, we will assume that the pressure $p(\rho)$ obeys a gamma type law

$$p(\rho) = \rho^\gamma, \qquad \gamma > 1, \tag{1.3}$$

(though more general pressure laws could be taken into account).

The viscosity coefficient $\mu(\rho)$ is often assumed to be a positive constant. However, it is well known that the viscosity of a gas depends on the temperature, and thus on the density (in the isentropic case). For example, the Chapman–Enskog viscosity law for hard sphere molecules predicts that $\mu(\rho)$ is proportional to the square root of the temperature (see [7]). In the case of monoatomic gas ($\gamma = 5/3$), this leads to $\mu(\rho) = \rho^{1/3}$. More generally, $\mu(\rho)$ is expected to vanish as a power of the ρ on the vacuum. In this paper, we consider degenerate viscosity coefficients that vanish for $\rho = 0$ at most like ρ^α for some $\alpha < 1/2$. In particular, the cases $\mu(\rho) = \nu$ and $\mu(\rho) = \nu\rho^{1/3}$ (with ν positive constant) are included in our result (see conditions (2.3)–(2.4) for details).

*Received by the editors April 25, 2006; accepted for publication (in revised form) May 14, 2007; published electronically January 4, 2008.

http://www.siam.org/journals/sima/39-4/65819.html

†Department of Mathematics, University of Texas at Austin, 1 University Station C1200, Austin, TX, 78712-0257 (mellet@math.utexas.edu, vasseur@math.utexas.edu). The first author was partially supported by NSF grant DMS-0456647. The second author was partially supported by NSF grant DMS-060705.

One-dimensional Navier–Stokes equations have been studied by many authors when the viscosity coefficient μ is a positive constant. The existence of weak solutions was first established by Kazhikhov and Shelukhin [16] for smooth enough data close to the equilibrium (bounded away from zero). The case of discontinuous data (still bounded away from zero) was addressed by Shelukhin [24, 25, 26] and then by Serre [22, 23] and Hoff [12]. First results concerning vanishing initial density were also obtained by Shelukhin [27]. In [14], Hoff proved the existence of global weak solutions with large discontinuous initial data, possibly having different limits at $x = \pm\infty$. Moreover, he proved that the constructed solutions have strictly positive densities (vacuum states cannot form in finite time). In dimension greater than two, similar results were obtained by Matsumura and Nishida [19] for smooth data and Hoff [13] for discontinuous data close to the equilibrium. The first global existence result for initial density that are allowed to vanish was due to Lions (see [17]). The result was later improved by Feireisl [10] and Feireisl, Novotný, and Petzeltová [11].

Another question is that of the regularity and uniqueness of the solutions. This problem was first analyzed by Solonnikov [29] for smooth initial data and for small time. However, the regularity may blow-up as the solution gets close to vacuum. This leads to another interesting question of whether vacuum may arise in finite time. Hoff and Smoller [15] show that any weak solution of the Navier–Stokes equations in one space dimension do not exhibit vacuum states, provided that no vacuum states are present initially. More precisely, they showed that if the initial data satisfies

$$\int_E \rho_0(x)\,dx > 0$$

for all open subsets $E \subset \mathbb{R}$, then

$$\int_E \rho(x,t)\,dx > 0$$

for every open subset $E \subset \mathbb{R}$ and for every $t \in [0,T]$.

The main theorem of this paper states that the strong solutions constructed by Solonnikov in [29] remain bounded away from zero uniformly in time (i.e., vacuum never arises) and are thus defined globally in time. This result can be seen as the equivalent of the result of Hoff in [13] for strong solutions instead of weak solutions. Another interest of this paper is the fact that unlike all the references mentioned above, the result presented here is valid with degenerate viscosity coefficients.

Note that compressible Navier–Stokes equations with degenerate viscosity coefficients have been studied before (see, for example, [18, 21, 31, 32]. All of those papers, however, are devoted to the case of compactly supported initial data and to the description of the evolution of the free boundary. We are interested here in the opposite situation in which vacuum never arises.

The new tool that allows us to obtain those results is an entropy inequality that was derived by Bresch and Desjardins in [2] for the multi-dimensional Korteweg system of equations (which corresponds to the case $\mu(\rho) = \rho$ and with an additional capillary term) and later generalized by the same authors (see [4]) to include other density-dependent viscosity coefficients. In the one dimensional case, a similar inequality was introduced earlier by Vaĭgant [30] for flows with constant viscosity (see also Shelukhin [28]).

The main interest of this inequality is to provide further regularity for the density. When $\mu(\rho) = \rho$, for instance, it implies that the gradient of $\sqrt{\rho}$ remains bounded for all time, provided it was bounded at time $t = 0$. This has very interesting consequences for many hydrodynamic equations. In [28], Shelukhin establishes the existence of a unique weak solution for one-dimensional flows with constant viscosity coefficient. In higher dimension, Bresch, Desjardin, and Lin use this inequality to establish the stability of weak solutions for the Korteweg system of equations in [6] and Bresch and Desjardin use this inequality for the compressible Navier–Stokes equations with an additional quadratic friction term in [3]. In [20], we establish the stability of weak solutions for the compressible isentropic Navier–Stokes equations in dimension 2 and 3 (without any additional terms). We also refer to [5] for recent developments concerning the full system of compressible Navier–Stokes equations (for heat conducting fluids).

At this point, we want to stress the fact that in dimension 2 and higher, this inequality holds only when the two viscosity coefficients satisfy a relation that considerably restricts the range of admissible coefficients (and in particular implies that one must have $\mu(0) = 0$). This necessary condition disappears in dimension 1, as the two viscosity coefficients become one (the derivation of the inequality is also much simpler in one dimension).

Another particularity of the dimension 1, is that the inequality gives control on some negative powers of the density (this is not true in higher dimensions). This will allow us to show that vacuum cannot arise if it was not present at time $t = 0$.

Finally, we point out that the present result is very different from that of [20] where the density was allowed to vanish (and the difficulty was to control the velocity u on the vacuum). Naturally, a result similar to that of [20] holds in dimension one, though it is not the topic of this paper.

Our main result is made precise in the next section. Section 3 is devoted to the derivation of the fundamental entropy inequalities and a priori estimates. The existence part of Theorem 2.1 is proved in section 4. The uniqueness is addressed in section 5.

2. The result.

Following Hoff in [14], we work with positive initial data having (possibly different) positive limits at $x = \pm\infty$: We fix constant velocities u_+ and u_- and constant positive density $\rho_+ > 0$ and $\rho_- > 0$, and we let $\overline{u}(x)$ and $\overline{\rho}(x)$ be two smooth monotone functions satisfying

$$\overline{\rho}(x) = \rho_\pm \quad \text{when } \pm x \geq 1, \quad \overline{\rho}(x) > 0 \quad \forall x \in \mathbb{R}, \tag{2.1}$$

and

$$\overline{u}(x) = u_\pm \quad \text{when } \pm x \geq 1. \tag{2.2}$$

We recall that the pressure satisfies $p(\rho) = \rho^\gamma$ for some $\gamma > 1$ and we assume that there exists a constant $\nu > 0$ such that the viscosity coefficient $\mu(\rho)$ satisfies

$$\begin{aligned} &\mu(\rho) \geq \nu\rho^\alpha && \forall \rho \leq 1 && \text{for some } \alpha \in [0, 1/2), \\ &\mu(\rho) \geq \nu && \forall \rho \geq 1, && \end{aligned} \tag{2.3}$$

and

$$\mu(\rho) \leq C + Cp(\rho) \qquad \forall \rho \geq 0. \tag{2.4}$$

Note that (2.4) is only a restriction on the growth of μ for large ρ. Examples of admissible viscosity coefficients include $\mu(\rho) = \nu$ and $\mu(\rho) = \rho^{1/3}$.

Our main theorem is the following.

THEOREM 2.1. *Assume that the initial data $\rho_0(x)$ and $u_0(x)$ satisfy*

$$0 < \underline{\kappa}_0 \le \rho_0(x) \le \overline{\kappa}_0 < \infty, \quad \rho_0 - \overline{\rho} \in H^1(\mathbb{R}), \quad u_0 - \overline{u} \in H^1(\mathbb{R}), \tag{2.5}$$

for some constants $\underline{\kappa}_0$ and $\overline{\kappa}_0$. Assume also that $\mu(\rho)$ verifies (2.3) *and* (2.4)*. Then there exists a global strong solution (ρ, u) of* (1.1)–(1.3) *on $\mathbb{R}^+ \times \mathbb{R}$ such that for every $T > 0$:*

$$\rho - \overline{\rho} \in L^\infty(0, T; H^1(\mathbb{R})),$$
$$u - \overline{u} \in L^\infty(0, T; H^1(\mathbb{R})) \cap L^2(0, T; H^2(\mathbb{R})).$$

Moreover, for every $T > 0$, there exist constants $\underline{\kappa}(T)$ and $\overline{\kappa}(T)$ such that

$$0 < \underline{\kappa}(T) \le \rho(x,t) \le \overline{\kappa}(T) < \infty \qquad \forall (t,x) \in (0,T) \times \mathbb{R}.$$

Finally, if $\mu(\rho) \ge \nu > 0$ for all $\rho \ge 0$, if μ is uniformly Lipschitz and if $\gamma \ge 2$, then this solution is unique in the class of weak solutions satisfying the usual entropy inequality (3.7)*.*

Note that the assumption (2.5) on the initial data implies, in particular, that the initial entropy (or relative entropy) is finite.

When the viscosity coefficient $\mu(\rho)$ satisfies

$$\mu(\rho) \ge \nu > 0 \qquad \forall \rho \ge 0, \tag{2.6}$$

the existence of a smooth solution for small time is a well-known result. More precisely, we have the following proposition.

PROPOSITION 2.2 (see [29]). *Let (ρ_0, u_0) satisfy* (2.5) *and assume that μ satisfies* (2.6)*, then there exists $T_0 > 0$ depending on $\underline{\kappa}_0$, $\overline{\kappa}_0$, $||\rho_0 - \overline{\rho}||_{H^1}$, and $||u_0 - \overline{u}||_{H^1}$ such that* (1.1)–(1.3) *has a unique solution (ρ, u) on $(0, T_0)$ satisfying*

$$\rho - \overline{\rho} \in L^\infty(0, T_1, H^1(\mathbb{R})), \qquad \partial_t \rho \in L^2((0,T_1) \times \mathbb{R}),$$
$$u - \overline{u} \in L^2(0, T_1; H^2(\mathbb{R})), \qquad \partial_t u \in L^2((0,T_1) \times \mathbb{R})$$

for all $T_1 < T_0$.
Moreover, there exist some $\underline{\kappa}(t) > 0$ and $\overline{\kappa}(t) < \infty$ such that $\underline{\kappa}(t) \le \rho(x,t) \le \overline{\kappa}(t)$ for all $t \in (0, T_0)$.

In view of this proposition, we see that if we introduce a truncated viscosity coefficient $\mu_n(\rho)$,

$$\mu_n(\rho) = \max(\mu(\rho), 1/n),$$

then there exist approximated solutions (ρ_n, u_n) defined for small time $(0, T_0)$ (T_0 possibly depending on n). To prove Theorem 2.1, we only have to show that (ρ_n, u_n) satisfies the following bounds uniformly with respect to n and T large:

$$\underline{\kappa}(T) \le \rho_n \le \overline{\kappa}(T) \quad \forall t \in [0,T],$$
$$\rho_n - \overline{\rho} \in L^\infty(0, T; H^1(\mathbb{R})),$$
$$u_n - \overline{u} \in L^\infty(0, T; H^1(\mathbb{R})).$$

In the next section, we derive the entropy inequalites that will be used to obtain the necessary bounds on ρ_n and u_n.

3. Entropy inequalities. In its conservative form, (1.1)–(1.3) can be written as

$$\partial_t U + \partial_x[A(U)] = \begin{bmatrix} 0 \\ (\mu(\rho)u_x)_x \end{bmatrix}$$

with the state vector

$$U = \begin{bmatrix} \rho \\ \rho\, u \end{bmatrix} = \begin{bmatrix} \rho \\ m \end{bmatrix}$$

and the flux

$$A(U) = \begin{bmatrix} \rho\, u \\ \rho\, u^2 + \rho^\gamma \end{bmatrix} = \begin{bmatrix} m \\ \dfrac{m^2}{\rho} + \rho^\gamma \end{bmatrix}.$$

It is well known that

$$\mathscr{H}(U) = \rho\frac{u^2}{2} + \frac{1}{\gamma-1}\rho^\gamma = \frac{m^2}{2\rho} + \frac{1}{\gamma-1}\rho^\gamma$$

is an entropy for the system of equations (1.1)–(1.3). More precisely, if (ρ, u) is a smooth solution, then we have

$$\partial_t \mathscr{H}(U) + \partial_x\big[F(U) - \mu(\rho)uu_x\big] + \mu(\rho)u_x^2 = 0 \tag{3.1}$$

with

$$F(U) = \rho u\frac{u^2}{2} + \frac{\gamma}{\gamma-1}u\rho^\gamma.$$

In particular, by integrating (3.1) with respect to x, we immediately see that

$$\frac{d}{dt}\int_{\mathbb{R}} \left[\rho\frac{u^2}{2} + \frac{1}{\gamma-1}\rho^\gamma\right] dx + \int_{\mathbb{R}} \mu(\rho)|u_x|^2\, dx \le 0. \tag{3.2}$$

However, since we are looking for solutions $\rho(x,t)$ and $u(x,t)$ which converge to $\rho_\pm$ and $u_\pm$ at $\pm\infty$, we do not expect the entropy to be integrable. It is thus natural to work with the relative entropy instead of the entropy.

The relative entropy is defined for any functions U and $\tilde{U}$ by

$$\begin{aligned} \mathscr{H}(U|\tilde{U}) &= \mathscr{H}(U) - \mathscr{H}(\tilde{U}) - D\mathscr{H}(\tilde{U})(U - \tilde{U}) \\ &= \rho(u-\tilde{u})^2 + p(\rho|\tilde{\rho}), \end{aligned}$$

where $p(\rho|\tilde{\rho})$ is the relative entropy associated to $\frac{1}{\gamma-1}\rho^\gamma$:

$$p(\rho|\tilde{\rho}) = \frac{1}{\gamma-1}\rho^\gamma - \frac{1}{\gamma-1}\tilde{\rho}^\gamma - \frac{\gamma}{\gamma-1}\tilde{\rho}^{\gamma-1}(\rho-\tilde{\rho}).$$

Note that, since p is strictly convex, $p(\rho|\tilde{\rho})$ is nonnegative for every ρ and $p(\rho|\tilde{\rho}) = 0$ if and only if $\rho = \tilde{\rho}$.

We recall that $\overline{\rho}(x)$ and $\overline{u}(x)$ are smooth functions satisfying (2.1) and (2.2), and we denote

$$\overline{U} = \begin{bmatrix} \overline{\rho} \\ \overline{\rho}\,\overline{u} \end{bmatrix}.$$

It is easy to check that there exists a positive constant C (depending on $\inf \overline{\rho}$) such that for every ρ and for every $x \in \mathbb{R}$, we have

$$\rho + \rho^\gamma \leq C[1 + p(\rho|\overline{\rho})], \tag{3.3}$$

$$\liminf_{\rho \to 0} p(\rho|\overline{\rho}) \geq C^{-1}. \tag{3.4}$$

The first inequality we will use in the proof of Theorem 2.1 is the usual relative entropy inequality for compressible Navier–Stokes equations.

LEMMA 3.1. *Let ρ, u be a solution of* (1.1)–(1.3) *satisfying the entropy inequality*

$$\partial_t \mathscr{H}(U) + \partial_x[F(U) - \mu(\rho) u u_x] + \mu(\rho)|u_x|^2 \leq 0. \tag{3.5}$$

Assume that the initial data (ρ_0, u_0) satisfies

$$\int_{\mathbb{R}} \mathscr{H}(U_0|\overline{U})\, dx = \int_{\mathbb{R}} \left[\rho_0 \frac{(u_0 - \overline{u})^2}{2} + p(\rho_0|\overline{\rho})\right] dx < +\infty. \tag{3.6}$$

Then, for every $T > 0$, there exists a positive constant $C(T)$ such that

$$\sup_{[0,T]} \int_{\mathbb{R}} \left[\rho \frac{(u - \overline{u})^2}{2} + p(\rho|\overline{\rho})\right] dx + \int_0^T \int_{\mathbb{R}} \mu(\rho)|u_x|^2\, dx\, dt \leq C(T). \tag{3.7}$$

The constant $C(T)$ depends only on $T > 0$, $\overline{U}$, the initial value U_0, γ, and the constant C appearing in (2.4).

Note that when both $\overline{\rho}$ and ρ_0 are bounded above and below away from zero, it is easy to check that

$$p(\rho_0|\overline{\rho}) \leq C(\rho_0 - \overline{\rho})^2,$$

and thus (3.6) holds under the assumptions of Theorem 2.1.

Proof of Lemma 3.1. First, we have (by a classical but tedious computation; see [8]) that

$$\begin{aligned}
\partial_t \mathscr{H}(U|\overline{U}) = {} & \big[\partial_t \mathscr{H}(U) + \partial_x(F(U) - \mu(\rho) u \partial_x u)\big] - \partial_t \mathscr{H}(\overline{U}) \\
& - \partial_x[F(U) - \mu(\rho) u \partial_x u] + \partial_x[DF(\overline{U})(U - \overline{U})] \\
& - D^2\mathscr{H}(\overline{U})[\partial_t \overline{U} + \partial_x A(\overline{U})](U - \overline{U}) \\
& - D\mathscr{H}(\overline{U})[\partial_t U + \partial_x A(U)] \\
& + D\mathscr{H}(\overline{U})[\partial_t \overline{U} + \partial_x A(\overline{U})] \\
& + D\mathscr{H}(\overline{U}) \partial_x[A(U|\overline{U})],
\end{aligned}$$

where the relative flux is defined by

$$\begin{aligned}
A(U|\overline{U}) &= A(U) - A(\overline{U}) - DA(\overline{U}) \cdot (U - \overline{U}) \\
&= \begin{bmatrix} 0 \\ \rho(u - \overline{u})^2 + (\gamma - 1) p(\rho|\overline{\rho}) \end{bmatrix}.
\end{aligned}$$

Since U is a solution of (1.1)–(1.3) and satisfies the entropy inequality, and using the fact that $\overline{U} = (\overline{\rho}, \overline{\rho u})$ satisfies (2.1) and (2.2) (and, in particular, $\partial_t \overline{U} = 0$), we deduce that

$$
\begin{aligned}
\partial_t \mathscr{H}(U|\overline{U}) \leq & -\mu(\rho)|\partial_x u|^2 \\
& -D^2\mathscr{H}(\overline{U})[\partial_x A(\overline{U})](U-\overline{U}) \\
& -D_2\mathscr{H}(\overline{U})[\partial_x(\mu(\rho)\partial_x u)] \\
& +D\mathscr{H}(\overline{U})\partial_x[A(U|\overline{U})] \\
& +D\mathscr{H}(\overline{U})[\partial_x A(\overline{U})] \\
& -\partial_x[F(U)-\mu(\rho)uu_x]+\partial_x[DF(\overline{U})(U-\overline{U})],
\end{aligned}
$$

where $D_2\mathscr{H}(\overline{U}) = \overline{u}$. We now integrate with respect to $x \in \mathbb{R}$, using the fact that supp $(\partial_x\overline{U}) \in [-1,1]$, and we get

$$
\begin{aligned}
& \frac{d}{dt}\int_{\mathbb{R}} \mathscr{H}(U|\overline{U})\,dx + \int_{\mathbb{R}} \mu(\rho)|\partial_x u|^2 \\
& \quad \leq -\int_{-1}^{1} D^2\mathscr{H}(\overline{U})[\partial_x A(\overline{U})](U-\overline{U})\,dx \\
& \qquad + \int_{-1}^{1} (\partial_x \overline{u})\mu(\rho)\partial_x u\,dx \\
& \qquad - \int_{-1}^{1} \partial_x[D\mathscr{H}(\overline{U})]A(U|\overline{U})\,dx \\
& \qquad - \int_{-1}^{1} \partial_x[D\mathscr{H}(\overline{U})]A(\overline{U})\,dx.
\end{aligned}
$$

Writing

$$
\left|\int_{-1}^{1} (\partial_x\overline{u})\mu(\rho)\partial_x u\,dx\right| \leq \|\partial_x \overline{u}\|_{L^\infty}^2 \int_{-1}^{1} \mu(\rho)\,dx + \frac{1}{2}\int_{-1}^{1} \mu(\rho)|\partial_x u|^2\,dx\,,
$$

it follows that there exists a constant C depending on $||\overline{U}||_{W^{1,\infty}}$ such that

$$
\begin{aligned}
& \frac{d}{dt}\int_{\mathbb{R}} \mathscr{H}(U|\overline{U})\,dx + \frac{1}{2}\int_{\mathbb{R}} \mu(\rho)\,|\partial_x u|^2 \\
& \quad \leq C + C\int_{-1}^{1} |U-\overline{U}|\,dx + C\int_{-1}^{1} |A(U|\overline{U})|\,dx \\
& \qquad + C\int_{-1}^{1} \mu(\rho)\,dx. \qquad (3.8)
\end{aligned}
$$

To conclude, we need to show that the right-hand side can be controlled by $\mathscr{H}(U|\overline{U})$. First, we note that

$$
|A(U|\overline{U})| \leq \max(1,(\gamma-1))\mathscr{H}(U|\overline{U}),
$$

and that (3.3) and (2.4) yield

$$
\int_{-1}^{1} \mu(\rho)\,dx \leq C + \int_{\mathbb{R}} p(\rho|\overline{\rho})\,dx. \qquad (3.9)
$$

Next, using (3.3) we get

$$
\begin{aligned}
\int_{-1}^{1} |U-\overline{U}|\,dx &\le \int_{-1}^{1} |\rho-\overline{\rho}|\,dx + \int_{-1}^{1} \rho|u-\overline{u}|\,dx + \int_{-1}^{1} |\overline{u}(\rho-\overline{\rho})|\,dx \\
&\le C\int_{-1}^{1} (1+p(\rho|\overline{\rho}))\,dx \\
&\quad + \left(\int_{-1}^{1} \rho\,dx\right)^{1/2} \left(\int_{-1}^{1} \rho(u-\overline{u})^2\,dx\right)^{1/2} \\
&\le C\int_{-1}^{1} (1+p(\rho|\overline{\rho}))\,dx \\
&\quad + \left(\int_{-1}^{1} (1+p(\rho|\overline{\rho}))\,dx\right)^{1/2} \left(\int_{-1}^{1} \mathscr{H}(U|\overline{U})\,dx\right)^{1/2} \\
&\le C\int_{-1}^{1} \mathscr{H}(U|\overline{U})\,dx + C.
\end{aligned}
$$

So (3.8) becomes

$$
\frac{d}{dt}\int_{\mathbb{R}} \mathscr{H}(U|\overline{U})\,dx + \frac{1}{2}\int_{\mathbb{R}} \mu(\rho)\,|\partial_x u|^2 \le C + C\int_{-1}^{1} |\mathscr{H}(U|\overline{U})|\,dx,
$$

and Gronwall's lemma gives Lemma 3.1.

For further reference, we note that this also implies that

$$
\frac{d}{dt}\int_{\mathbb{R}} \mathscr{H}(U|\overline{U})\,dx \le C(T). \qquad \square \tag{3.10}
$$

Unfortunately, it is a well-known fact that the estimates provided by Lemma 3.1 are not enough to prove the stability of the solutions of (1.1)–(1.3). The key tool of this paper is thus the following lemma.

Lemma 3.2. *Assume that $\mu(\rho)$ is a C^2 function, and let (ρ, u) be a solution of* (1.1)–(1.3) *such that*

$$
u-\overline{u} \in L^2((0,T);H^2(\mathbb{R})), \quad \rho-\overline{\rho} \in L^\infty((0,T);H^1(\mathbb{R})), \quad 0 < m \le \rho \le M. \tag{3.11}
$$

Then there exists $C(T)$ such that the following inequality holds:

$$
\begin{aligned}
&\sup_{[0,T]} \int \left[\frac{1}{2}\rho\big|(u-\overline{u}) + \partial_x(\varphi(\rho))\big|^2 + p(\rho|\overline{\rho})\right] dx \\
&\qquad + \int_0^T \int_{\mathbb{R}} \partial_x(\varphi(\rho))\partial_x(\rho^\gamma)\,dx\,dt \le C(T),
\end{aligned} \tag{3.12}
$$

with φ such that

$$
\varphi'(\rho) = \frac{\mu(\rho)}{\rho^2}. \tag{3.13}
$$

The constant $C(T)$ depends only on $T > 0$, $(\overline{\rho}, \overline{u})$, the initial value U_0, γ, and the constant C appearing in (2.4).

Since the viscosity coefficient $\mu(\rho)$ is nonnegative, (3.13) implies that $\varphi(\rho)$ is increasing. The lemma thus implies that

$$\sup_{[0,T]} \int \left[\frac{1}{2}\rho \big| (u-\overline{u}) + \partial_x(\varphi(\rho)) \big|^2 + p(\rho|\overline{\rho}) \right] dx \leq C(T),$$

which, together with Lemma 3.1, yields

$$||\sqrt{\rho}\partial_x(\varphi(\rho))||_{L^\infty(0,T;L^2(\Omega))} = 2||\mu(\rho)(\rho^{-1/2})_x||_{L^\infty(0,T;L^2(\Omega))} \leq C(T).$$

This inequality will be the cornerstone of the proof of Theorem 2.1 which is detailed in the next section.

As mentioned in the introduction, Lemma 3.2 relies on a new mathematical entropy inequality that was first derived by Bresch and Desjardins in [2] and [4] in dimension 2 and higher. Of course, the computations are much simpler in dimension 1.

We stress the fact that it is important to know exactly what regularity is needed on ρ and u to establish this inequality. Indeed, unlike inequality (3.7) which is quite classical, there is no obvious way to regularize the system (1.1)–(1.3) while preserving the structure necessary to derive (3.12). Fortunately, it turns out that (3.11), which is the natural regularity for strong solutions, is enough to justify the computations, as we will see in the proof.

Proof. We have to show that

$$\frac{d}{dt}\int \left[\frac{1}{2}\rho|u-\overline{u}|^2 + \rho(u-\overline{u})(\varphi(\rho))_x + \frac{1}{2}\rho(\varphi(\rho))_x^2 \right] dx + \frac{d}{dt}\int p(\rho|\overline{\rho})\, dx$$

is bounded

Step 1. From the proof of the previous lemma (see (3.10)), we already know that

$$\frac{d}{dt}\int \left[\frac{1}{2}\rho|u-\overline{u}|^2 \right] dx + \frac{d}{dt}\int p(\rho|\overline{\rho})\, dx \leq C(T). \tag{3.14}$$

Step 2. Next we show that

$$\begin{aligned} \frac{d}{dt}\int \rho\frac{(\varphi(\rho))_x^2}{2}\, dx = &-\int \rho^2\varphi'(\rho)(\varphi(\rho))_x u_{xx}\, dx \\ &-\int (2\rho\varphi'(\rho) + \rho^2\varphi''(\rho))\rho_x(\varphi(\rho))_x u_x\, dx. \end{aligned} \tag{3.15}$$

This follows straightforwardly from (1.1) when the second derivatives of the density are bounded in $L^2((0,T)\times\mathbb{R})$. In that case, it is worth mentioning that the right-hand side can be rewritten as

$$\int \rho^2\varphi'(\rho)(\varphi(\rho))_{xx} u_x\, dx.$$

However, we do not have any bounds on ρ_{xx}. It is thus important to justify the derivation of (3.15).

First, we point out that (3.15) makes sense when (ρ,u) satisfies only (3.11) (we recall that since $\overline{\rho}$ and $\overline{u}$ are constant outside $(-1,1)$, (3.11) implies that $\rho_x \in L^\infty((0,T);L^2(\mathbb{R}))$ and $u_x \in L^2((0,T);H^1(\mathbb{R}))$). Moreover, we note that the computation only makes use of the continuity equation. The rigorous derivation of (3.15) (under assumption (3.11)) can thus be achieved by carefully regularizing the continuity equation. The details are presented in the appendix (see Lemma A.1).

Step 3. Next, we evaluate the derivative of the cross-product:

$$\begin{aligned}\frac{d}{dt}\int \rho(u-\overline{u})\partial_x(\varphi(\rho))\,dx\\ &=\int \partial_x(\varphi(\rho))\partial_t(\rho(u-\overline{u}))\,dx+\int \rho(u-\overline{u})\partial_t\partial_x(\varphi(\rho))\,dx\\ &=\int \partial_x(\varphi(\rho))\partial_t(\rho(u-\overline{u}))\,dx-\int(\rho(u-\overline{u}))_x\varphi'(\rho)\partial_t\rho\,dx\,.\end{aligned}\tag{3.16}$$

Multiplying (1.2) by $\partial_x\varphi(\rho)$, we get

$$\begin{aligned}\int \partial_x(\varphi(\rho))\partial_t(\rho(u-\overline{u}))\,dx &=\int \partial_x(\varphi(\rho))\partial_t(\rho u)\,dx-\int \partial_x(\varphi(\rho))(\partial_t\rho)\overline{u}\,dx\\ &=\int(\varphi(\rho))_x(\mu(\rho)u_x)_x\,dx\\ &\quad-\int \partial_x(\varphi(\rho))\partial_x(\rho^\gamma)\,dx\\ &\quad-\int \partial_x(\varphi(\rho))\partial_x(\rho u^2)\,dx\\ &\quad+\int \varphi'(\rho)(\rho u)_x\rho_x\overline{u}\,dx.\end{aligned}$$

The continuity equation easily yields

$$\int(\rho(u-\overline{u}))_x\varphi'(\rho)\partial_t\rho\,dx=-\int((\rho u)_x)^2\varphi'(\rho)\,dx+\int(\rho\overline{u})_x(\rho u)_x\varphi'(\rho)\,dx.$$

Note that those equalities hold as soon as ρ and u satisfy (3.11).

Step 4. If φ and μ satisfy (3.13), then we have

$$\begin{aligned}\int(\varphi(\rho))_x(\mu(\rho)u_x)_x\,dx &=\int \rho^2\varphi'(\rho)(\varphi(\rho))_xu_{xx}\,dx\\ &\quad\int(2\rho\varphi'(\rho)+\rho^2\varphi''(\rho))\rho_x(\varphi(\rho))_xu_x\,dx,\end{aligned}$$

so (3.15) and (3.16) yield

$$\begin{aligned}\frac{d}{dt}\left\{\int \rho(u-\overline{u})\partial_x\varphi(\rho)+\rho\frac{|\partial_x\varphi(\rho)|^2}{2}\,dx\right\}+\int \partial_x\varphi(\rho)\partial_x(\rho^\gamma)\,dx\\ =-\int \partial_x(\varphi(\rho))\partial_x(\rho u^2)\,dx+\int((\rho u)_x)^2\varphi'(\rho)\,dx\\ +\int(\rho u)_x\varphi'(\rho)[\rho_x\overline{u}-(\rho\overline{u})_x]\,dx\\ =\int \varphi'(\rho)[-\rho_x(\rho u^2)_x+((\rho u)_x)^2]\,dx-\int(\rho u)_x\varphi'(\rho)\rho\overline{u}_x\,dx\\ =\int \rho^2\varphi'(\rho)u_x^2\,dx-\int(\rho u)_x\varphi'(\rho)\rho\overline{u}_x\,dx,\end{aligned}$$

and using (3.13), we deduce

$$\frac{d}{dt}\left\{\int \rho(u-\overline{u})\partial_x\varphi(\rho)+\rho\frac{|\partial_x\varphi(\rho)|^2}{2}\,dx\right\}+\int \partial_x\varphi(\rho)\partial_x(\rho^\gamma)\,dx$$
$$=\int \mu(\rho)(u_x)^2\,dx-\int \mu(\rho)u_x\overline{u}_x\,dx-\int \rho\,\partial_x(\varphi(\rho))u\,\overline{u}_x\,dx. \tag{3.17}$$

Moreover, since $\overline{u}_x$ has support in $(-1,1)$ and using the bounds given by Lemma 3.1 and inequality (3.9), it is readily seen that the right-hand side in this equality is bounded by

$$C\int_{\mathbb{R}}\mu(\rho)|u_x|^2\,dx+C\int_{-1}^{1}\mu(\rho)\,dx+C\int\rho|\partial_x\varphi(\rho)|^2\,dx+C\int_{-1}^{1}\rho u^2\,dx$$
$$\leq C\int\mu(\rho)|u_x|^2\,dx+C\int_{\mathbb{R}}p(\rho,\overline{\rho})\,dx+C\int_{\mathbb{R}}\rho|(u-\overline{u})+\partial_x\varphi(\rho)|^2\,dx+C(T).$$

Putting (3.17) and (3.14) together, we deduce

$$\frac{d}{dt}\int_{\mathbb{R}}\left[\frac{1}{2}\rho\big|(u-\overline{u})+\partial_x(\varphi(\rho))\big|^2+p(\rho|\overline{\rho})\right]dx+\int_{\mathbb{R}}\partial_x(\varphi(\rho))\partial_x(\rho^\gamma)\,dx$$
$$\leq C\int_{\mathbb{R}}\mu(\rho)|u_x|^2\,dx+C\int_{\mathbb{R}}\left[\frac{1}{2}\rho|(u-\overline{u})+\partial_x\varphi(\rho)|^2+p(\rho,\overline{\rho})\right]dx+C(T).$$

Finally, using the bounds on the viscosity from Lemma 3.1 and Gronwall's inequality we easily deduce (3.12). ∎

4. Proof of Theorem 2.1. In this section, we prove the existence part of Theorem 2.1.

The proof relies on the following proposition.

PROPOSITION 4.1. *Assume that the viscosity coefficient μ satisfies* (2.3)–(2.4) *and consider initial data (ρ_0,u_0) satisfying* (2.5). *Then for all $T>0$ there exist some constants $C(T)$, $\underline{\kappa}(T)$, and $\overline{\kappa}(T)$ such that for any strong solution $(\rho,\,u)$ of* (1.1)–(1.3) *with initial data (ρ_0,u_0), defined on $(0,T)$ and satisfying*

$$\rho-\overline{\rho}\in L^\infty(0,T,H^1(\mathbb{R}))\,,\qquad \partial_t\rho\in L^2((0,T)\times\mathbb{R})\,,$$
$$u-\overline{u}\in L^2(0,T;H^2(\mathbb{R}))\,,\qquad \partial_t u\in L^2((0,T)\times\mathbb{R}),$$

with ρ and ρ^{-1} bounded, the following bounds hold:

$$0<\underline{\kappa}(T)\leq\rho(t)\leq\overline{\kappa}(T)\quad \forall t\in[0,T],$$
$$\|\rho-\overline{\rho}\|_{L^\infty(0,T;H^1(\mathbb{R}))}\leq C(T),$$
$$\|u-\overline{u}\|_{L^\infty(0,T;H^1(\mathbb{R}))}\leq C(T).$$

Moreover, the constants $C(T)$, $\underline{\kappa}(T)$, and $\overline{\kappa}(T)$ depend on μ only through the constant C arising in (2.3) *and* (2.4).

Proof of Theorem 2.1. We define $\mu_n(\rho)$ to be the following positive approximation of the viscosity coefficient:

$$\mu_n(s)=\max(\mu(s),1/n).$$

Notice that μ_n verifies

$$\mu\leq\mu_n\leq\mu+1.$$

In particular, μ_n satisfies (2.3) and (2.4) with some constants that are independent on n.

Next, for all $n > 0$, we let (ρ_n, u_n) be the strong solution of (1.1)–(1.3) with $\mu = \mu_n$:

$$\begin{aligned}&\rho_t + (\rho u)_x = 0,\\ &(\rho u)_t + (\rho u^2)_x + p(\rho)_x = (\mu_n(\rho)\, u_x)_x.\end{aligned}$$

This solution exists at least for small time $(0, T_0)$ thanks to Proposition 2.2 (note that T_0 may depend on n). Proposition 4.1 then implies that for all $T > 0$ there exists $C(T)$, $\overline{\kappa}(T)$, and $\underline{\kappa}(T) > 0$, independent on n, such that

$$\begin{aligned}&\underline{\kappa}(T) \le \rho_n(t) \le \overline{\kappa}(T) \quad \forall t \in [0, T],\\ &\|\rho_n - \overline{\rho}\|_{L^\infty(0,T;H^1(\mathbb{R}))} \le C(T),\\ &\|u_n - \overline{u}\|_{L^\infty(0,T;H^1(\mathbb{R}))} \le C(T).\end{aligned}$$

In particular, we can take $T_0 = \infty$ in Proposition 2.2 (for all n). Moreover, since the bound from below for the density is uniform in n for any $T > 0$, by taking n large enough (namely $n \ge 1/\underline{\kappa}(T)$), it is readily seen that (ρ_n, u_n) is a solution of (1.1)–(1.3) on $[0, T]$ with the nontruncated viscosity coefficient $\mu(\rho)$. From the uniqueness of the solution of Proposition 2.2, we see that, passing to the limit in n, we get the desired global solution of (1.1)–(1.3). □

The rest of this section is thus devoted to the proof of Proposition 4.1. First, we will show that ρ is bounded from above and from below uniformly by some positive constants. Then we will investigate the regularity of the velocity by some standard arguments for parabolic equations.

4.1. A priori estimates. Since the initial datum (ρ_0, u_0) satisfies (2.5), we have

$$\int \rho_0 (u_0 - \overline{u})^2\, dx < \infty \quad \text{and} \quad \int_\Omega \rho_0 |\partial_x(\varphi(\rho_0))|^2\, dx < +\infty.$$

Moreover, $(\rho,\, u)$ satisfies (3.11), so we can use the inequalities stated in Lemmas 3.1 and 3.2. We deduce the following estimates, which we shall use throughout the proof of Proposition (4.1):

$$\begin{aligned}&||\sqrt{\rho}(u - \overline{u})||_{L^\infty(0,T;L^2(\Omega))} \le C(T),\\ &||\rho||_{L^\infty(0,T;L^1_{\text{loc}} \cap L^\gamma_{\text{loc}}(\Omega))} \le C(T),\\ &||\rho - \overline{\rho}||_{L^\infty(0,T;L^1(\Omega))} \le C(T),\\ &||\sqrt{\mu(\rho)}(u)_x||_{L^2(0,T;L^2(\Omega))} \le C(T).\end{aligned} \tag{4.1}$$

and

$$\begin{aligned}&||\mu(\rho)\partial_x(\rho^{-1/2})||_{L^\infty(0,T;L^2(\Omega))} \le C(T),\\ &||\sqrt{\mu(\rho)}\partial_x(\rho^{\gamma/2-1/2})||_{L^2(0,T;L^2(\Omega))} \le C(T).\end{aligned} \tag{4.2}$$

4.2. Uniform bounds for the density. The first proposition shows that no vacuum states can arise.

PROPOSITION 4.2. *For every $T > 0$, there exists a constant $\underline{\kappa}(T) > 0$ such that*

$$\rho(x, t) \ge \underline{\kappa}(T) \qquad \forall (x, t) \in \mathbb{R} \times [0, T].$$

The proof of this proposition will follow from two lemmas. The first is as follows.

LEMMA 4.3. *For every $T > 0$, there exist $\delta > 0$ and $R(T)$ such that for every $x_0 \in \mathbb{R}$ and $t_0 > 0$, there exists $x_1 \in [x_0 - R(T),\ x_0 + R(T)]$ with*

$$\rho(x_1, t_0) > \delta.$$

This nice result can be found in [14]. We give a proof of it for the sake of completeness.

Proof. Let $\delta > 0$ be such that

$$p(\rho|\overline{\rho}) \geq \frac{C^{-1}}{2} \qquad \forall \rho < \delta$$

(such a δ exists thanks to (3.4)). Then, if

$$\sup_{x \in [x_0 - R, x_0 + R]} \rho(x, t_0) < \delta,$$

we then have

$$\int p(\rho|\overline{\rho})\, dx \geq C^{-1} R,$$

and since the integral in the left-hand side is bounded by a constant (see Lemma 3.1), a suitable choice of R leads to a contradiction. ◻

LEMMA 4.4. *Let*

$$w(x,t) = \inf(\rho(x,t), 1) = 1 - (1 - \rho(x,t))_+.$$

Then there exists $\varepsilon > 0$ and a constant $C(T)$ such that

$$||\partial_x w^{-\varepsilon}||_{L^\infty(0,T;L^2(\mathbb{R}))} \leq C(T).$$

Proof. We have

$$\partial_x w = \partial_x \rho 1_{\{\rho \leq 1\}}.$$

In particular, (4.2) gives

$$\left\| \frac{\mu(w)}{w^{3/2}} w_x \right\|_{L^\infty(0,T;L^2(\Omega))} \leq C$$

so using (2.3) we deduce that

$$||w^{\alpha - 3/2} \partial_x w||_{L^\infty(0,T;L^2(\Omega))} = ||\partial_x w^{\alpha - 1/2}||_{L^\infty(0,T;L^2(\Omega))} \leq C,$$

and the result follows with $\varepsilon = 1/2 - \alpha > 0$. ◻

Proof of Proposition 4.2. Together with Sobolev–Poincaré inequality, Lemmas 4.3 and 4.4 yield that $w^{-\varepsilon}$ is bounded in $L^\infty((0,T) \times \mathbb{R})$:

$$w^{-\varepsilon}(x,t) \leq C(T) \qquad \forall (x,t) \in \mathbb{R} \times (0,T).$$

This yields Proposition 4.2 with $\underline{\kappa}(T) = C(T)^{-1/\varepsilon}$. ◻

Next, we find a bound for the density in L^∞.

PROPOSITION 4.5. *For every $T > 0$, there exist a constant $\overline{\kappa}(T)$ such that*

$$\rho(x,t) \le \overline{\kappa}(T) \qquad \forall (x,t) \in \mathbb{R} \times (0,T).$$

Proof. Let $s = (\gamma - 1)/2$, then (3.12) with (2.3) and (3.13) yields $\partial_x(\rho^s)$ bounded in $L^2((0,T)\times\mathbb{R})$. Moreover, for every compact subset K of $\mathbb{R}$, we have

$$\begin{aligned}\int_K |\partial_x \rho^s|\,dx &= \int_K |\rho^{s-1}\partial_x\rho|\,dx \\ &\le \left(\int_K \rho^{1+2s}\,dx\right)^{1/2}\left(\int_K \frac{1}{\rho^3}(\partial_x\rho)^2\,dx\right)^{1/2} \\ &\le \left(\int_K \rho^{\gamma}\,dx\right)^{1/2}\left(\int_K \rho\varphi'(\rho)^2(\partial_x\rho)^2\,dx\right)^{1/2}\end{aligned}$$

and so using (3.3) we get

$$\int_K |\partial_x\rho^s|\,dx \le C\left(|K| + \int_K p(\rho|\overline{\rho})\,dx\right)^{1/2}\left(\int_K \rho\varphi'(\rho)^2(\partial_x\rho)^2\,dx\right)^{1/2}.$$

Since

$$\rho^s \le 1 + \rho^{\gamma},$$

we deduce that

$$\rho^s \text{ is bounded in } L^{\infty}\left(0,T; W^{1,1}_{\mathrm{loc}}(\mathbb{R})\right),$$

and the $W^{1,1}(K)$ norm of $\rho^s(t,\cdot)$ only depends on $|K|$. Sobolev imbedding thus yields Proposition 4.5. ∎

PROPOSITION 4.6. *There exists a constant $C(T)$ such that*

$$\|\rho(x,t) - \overline{\rho}(x)\|_{L^{\infty}(0,T;H^1(\mathbb{R}))} \le C(T).$$

Proof. Proposition 4.5 yields

$$\begin{aligned}\int (\partial_x\rho)^2\,dx &\le \overline{\kappa}^3 \int \frac{1}{\rho^3}(\partial_x\rho)^2\,dx \\ &\le \overline{\kappa}^3 \int \frac{\rho}{(\mu(\rho))^2}(\phi'(\rho))^2(\partial_x\rho)^2\,dx \\ &\le \frac{\nu\overline{\kappa}^3}{\inf(1,\underline{\kappa}^{2\alpha})}\int \rho(\partial_x\phi(\rho))^2\,dx \\ &\le C(T),\end{aligned}$$

and the result follows. ∎

4.3. Uniform bounds for the velocity.

PROPOSITION 4.7. *There exists a constant $C(T)$ such that*

$$\|u - \overline{u}\|_{L^2(0,T;H^2(\mathbb{R}))} \le C(T)$$

and

$$\|\partial_t u\|_{L^2(0,T;L^2(\mathbb{R}))} \le C(T).$$

In particular, $u-\overline{u}\in C^0(0,T;H^1(\mathbb{R}))$.

Proof. First, we show that $u-\overline{u}$ is bounded in $L^2(0,T;H^1(\mathbb{R}))$. Since $\rho\geq\kappa>0$, and using (2.3), it is readily seen that there exists a constant $\nu'>0$ such that

$$\mu(\rho(x,t))\geq\nu' \qquad \forall(x,t)\in\mathbb{R}\times[0,T],$$

and so (3.7) gives

$$\partial_x u \text{ is bounded in } L^2((0,T)\times\mathbb{R})$$

and

$$u-\overline{u} \text{ is bounded in } L^\infty(0,T;L^2(\mathbb{R})).$$

Therefore $u-\overline{u}$ is bounded in $L^2(0,T;H^1(\mathbb{R}))$.

Note that this implies that $\partial_t\rho$ is bounded in $L^2((0,T)\times\mathbb{R})$. Since $\rho-\overline{\rho}$ is bounded in $L^\infty(0,T;H^1)$, it follows (see [1], for example) that

$$\rho\in\mathcal{C}^{s_0}((0,T)\times\mathbb{R})$$

for some $s_0\in(0,1)$.

Next, we rewrite (1.2) as follows:

$$\partial_t u-\left(\frac{\mu(\rho)}{\rho}u_x\right)_x=-\gamma\rho^{\gamma-2}\rho_x+\overline{u}u_x+(\partial_x(\varphi(\rho))-(u-\overline{u}))u_x, \tag{4.3}$$

where we recall that φ, which is defined by $\varphi'(\rho)=\mu(\rho)/\rho^2$, is the function arising in the new entropy inequality (see Lemma 3.2).

In order to deduce some bounds on u, we need to control the right-hand side of (4.3). The first term, $\rho^{\gamma-2}\rho_x$, is bounded in $L^\infty(0,T;L^2(\mathbb{R}))$ (thanks to Proposition 4.6). The second term is bounded in $L^2((0,T)\times\mathbb{R})$ since $\overline{u}$ is in L^∞. For the last part, we write (using Hölder inequality and interpolation inequality):

$$\begin{aligned}
&||\partial_x(\varphi(\rho))-(u-\overline{u})u_x||_{L^2(0,T;L^{4/3}(\mathbb{R}))}\\
&\quad\leq||\partial_x(\varphi(\rho))-(u-\overline{u})||_{L^\infty(0,T;L^2(\mathbb{R}))}||u_x||_{L^2(0,T;L^4(\mathbb{R}))}\\
&\quad\leq||\partial_x(\varphi(\rho))-(u-\overline{u})||_{L^\infty(L^2)}||u_x||^{2/3}_{L^2(L^2)}||u_x||^{1/3}_{L^2(0,T;W^{1,4/3}(\mathbb{R}))}\\
&\quad\leq C||u_x||^{1/3}_{L^2(0,T;W^{1,4/3}(\mathbb{R}))}
\end{aligned}$$

(here we make use of (3.12) and Proposition 4.2). So regularity results for parabolic equation of the form (4.3) (note that the diffusion coefficient is in $\mathcal{C}^{s_0}((0,T)\times\mathbb{R})$) yield

$$||u_x||_{L^2(0,T;W^{1,4/3}(\mathbb{R}))}\leq C||u_x||^{1/3}_{L^2(0,T;W^{1,4/3}(\mathbb{R}))}+C,$$

and so

$$||u_x||_{L^2(0,T;W^{1,4/3}(\mathbb{R}))}\leq C.$$

Using Sobolev inequalities, it follows that u_x is bounded in $L^2(0,T;L^\infty(\mathbb{R}))$.

Finally, we can now see that the right-hand side in (4.3) is bounded in $L^2(0,T;L^2(\mathbb{R}))$, and classical regularity results for parabolic equations give

$$u-\overline{u} \text{ is bounded in } L^2(0,T;H^2(\mathbb{R}))$$

and

$$\partial_t u \text{ is bounded in } L^2(0,T;L^2(\mathbb{R})),$$

which concludes the proof. ∎

It is now readily seen that Proposition 4.1 follows from Propositions 4.2, 4.5, 4.6, and 4.7.

5. Uniqueness. In this last section, we establish the uniqueness of the global strong solution in a large class of weak solutions satisfying the usual entropy inequality. This result can be rewritten as follows.

PROPOSITION 5.1. *Assume that*

$$\mu(\rho) \geq \nu > 0 \quad \forall \rho \geq 0,$$

and that there exists a constant C such that

$$|\mu(\rho) - \mu(\tilde{\rho})| \leq C|\rho - \tilde{\rho}| \quad \forall \rho, \tilde{\rho} \geq 0.$$

Assume moreover that $\gamma \geq 2$, and let (ρ, u) be the solution of (1.1)–(1.3) *given by Theorem* 2.1.

If $(\tilde{\rho}, \tilde{u})$ is a weak solution of (1.1)–(1.3) *with initial data (ρ_0, u_0) and satisfies the entropy inequality* (3.5) *and relative entropy bound* (3.7)*, and if*

$$\lim_{x\to\pm\infty} (\tilde{\rho} - \rho_\pm) = 0\,, \qquad \lim_{x\to\pm\infty} (\tilde{u} - u_\pm) = 0,$$

then

$$(\tilde{\rho}, \tilde{u}) = (\rho, u).$$

Notice that we do not need to assume that $\tilde{\rho}$ does not vanish; this proposition will be a consequence of the following lemma.

LEMMA 5.2. *Let $\tilde{U} = (\tilde{\rho}, \tilde{\rho}\tilde{u})$ be a weak solution of* (1.1)–(1.3) *satisfying the inequality* (3.5)*, and let $U = (\rho, \rho u)$ be a strong solution of* (1.1)–(1.3) *satisfying the equality* (3.1)*. Assume, moreover, that $\tilde{U}$ and U are such that*

$$\lim_{x\to\pm\infty} (\tilde{\rho} - \rho) = 0\,, \qquad \lim_{x\to\pm\infty} (\tilde{u} - u) = 0. \tag{5.1}$$

Then we have

$$\begin{aligned} \frac{d}{dt}\int_{\mathbb{R}} \mathscr{H}(\tilde{U}|U)\,dx &+ \int_{\mathbb{R}} \mu(\tilde{\rho})\,[\partial_x(\tilde{u}-u)]^2 \\ &\leq C\int |\partial_x u|\mathscr{H}(\tilde{U}|U)\,dx \\ &\quad - \int_{\mathbb{R}} \partial_x u[\mu(\tilde{\rho}) - \mu(\rho)]\,[\partial_x(\tilde{u}-u)]\;dx \\ &\quad + \int_{\mathbb{R}} \frac{\partial_x(\mu(\rho)\partial_x u)}{\rho}\,(\tilde{\rho}-\rho)\,(u-\tilde{u})\,dx. \end{aligned} \tag{5.2}$$

The proof of this lemma relies only on the structure of the equation and not on the properties of the solutions. We postpone it to the end of this section.

Proof of Proposition 5.1. In order to prove Proposition 5.1, we have to show that the last two terms in (5.2) can be controlled by the relative entropy $\mathscr{H}(\tilde{U}|U)$ and the viscosity. Since $\gamma \geq 2$ and $\rho \geq \overline{\kappa} > 0$, we note that there exists C such that

$$p(\tilde{\rho}|\rho) \geq C|\tilde{\rho}-\rho|^2 \quad \forall \tilde{\rho} \geq 0.$$

Then, we can write

$$\begin{aligned}
&\left|\int_{\mathbb{R}} \partial_x u[\mu(\tilde{\rho})-\mu(\rho)]\,[\partial_x(\tilde{u}-u)]\;dx\right| \\
&\qquad \leq C\|\partial_x u\|_{L^\infty(\mathbb{R})}\left(\int_{\mathbb{R}}|\tilde{\rho}-\rho|^2\,dx\right)^{1/2}\left(\int_{\mathbb{R}}|\partial_x(\tilde{u}-u)|^2\,dx\right)^{1/2} \\
&\qquad \leq C\|\partial_x u\|^2_{L^\infty(\mathbb{R})}\int_{\mathbb{R}}|\tilde{\rho}-\rho|^2\,dx+\frac{1}{4}\int_{\mathbb{R}}\mu(\tilde{\rho})|\partial_x(\tilde{u}-u)|^2\,dx \\
&\qquad \leq C\|\partial_x u\|^2_{L^\infty(\mathbb{R})}\int_{\mathbb{R}}\mathscr{H}(\tilde{U}|U)\,dx+\frac{1}{4}\int_{\mathbb{R}}\mu(\tilde{\rho})|\partial_x(\tilde{u}-u)|^2\,dx
\end{aligned}$$

which does the trick for the first of the last two terms in (5.2). For the last term, we see that if we had $\partial_x(\mu(\rho)\partial_x u)$ bounded in $L^\infty((0,T)\times\mathbb{R})$, a similar computation would apply. However, writing

$$\partial_x(\mu(\rho)\partial_x u) = \mu'(\rho)(\partial_x\rho)\,(\partial_x u)+\mu(\rho)\partial_{xx}u$$

it is readily seen that $\partial_x(\mu(\rho)\partial_x u)$ is only bounded in $L^2((0,T)\times\mathbb{R})$. For that reason, we need to control $|\tilde{u}-u|$ in L^∞, which is made possible by the following lemma.

LEMMA 5.3. *Let $\tilde{\rho} \geq 0$ be such that $\int p(\tilde{\rho}|\overline{\rho})\,dx < +\infty$. Then there exists a constant C (depending on $\int p(\tilde{\rho}|\overline{\rho})\,dx$) such that for any regular function h:*

$$\|h\|_{L^\infty(\mathbb{R})} \leq C\left(\int_{\mathbb{R}}\tilde{\rho}|h|^2\,dx\right)^{1/2}+C\left(\int_{\mathbb{R}}|h_x|^2\,dx\right)^{1/2}.$$

Using Lemma 5.3 with $h=\tilde{u}-u$, we deduce that

$$\begin{aligned}
&\int_{\mathbb{R}}\frac{\partial_x(\mu(\rho)\partial_x u)}{\rho}\,(\tilde{\rho}-\rho)\,(u-\tilde{u})\,dx \\
&\leq \left\|\frac{\partial_x(\mu(\rho)\partial_x u)}{\rho}\right\|_{L^2(\mathbb{R})}\|\tilde{\rho}-\rho\|_{L^2(\mathbb{R})}\|u-\tilde{u}\|_{L^\infty(\mathbb{R})} \\
&\leq C\left\|\frac{\partial_x(\mu(\rho)\partial_x u)}{\rho}\right\|_{L^2(\mathbb{R})}\mathscr{H}(\tilde{U}|U)^{\frac{1}{2}}\left(\mathscr{H}(\tilde{U}|U)^{\frac{1}{2}}+\left(\int_R|\partial_x(u-\tilde{u})|^2\,dx\right)^{\frac{1}{2}}\right) \\
&\leq C\left\|\frac{\partial_x(\mu(\rho)\partial_x u)}{\rho}\right\|^2_{L^2(\mathbb{R})}\mathscr{H}(\tilde{U}|U)+\frac{1}{4}\int_R\mu(\tilde{\rho})|\partial_x(u-\tilde{u})|^2\,dx.
\end{aligned}$$

So (5.2) becomes

$$\frac{d}{dt}\int_{\mathbb{R}}\mathscr{H}(\tilde{U}|U)\,dx+\frac{1}{2}\int_{\mathbb{R}}\mu(\tilde{\rho})\,[\partial_x(\tilde{u}-u)]^2 \leq C(t)\int\mathscr{H}(\tilde{U}|U)\,dx,$$

where $C(t)\in L^1(0,T)$. The Gronwall lemma, together with the fact that

$$\mathscr{H}(\tilde{U}|U)(t=0)=0,$$

yields Proposition 5.1. ◻

Proof of Lemma 5.3. Using (3.4), we see that there exists some $\delta > 0$ and C such that

$$|\{x \in \mathbb{R};\ \tilde{\rho} \leq \delta\}| \leq C \int_{\mathbb{R}} p(\tilde{\rho}|\overline{\rho})\, dx.$$

We take $R = C \int p(\tilde{\rho}|\overline{\rho})\, dx + 1$. Then, for every x_0 in $\mathbb{R}$, we know that in the interval $(x_0 - R/2, x_0 + R/2)$, $\tilde{\rho}$ is larger than δ, which is a set of measure at least 1. We denote by ω this set:

$$\omega = (x_0 - R/2, x_0 + R/2) \cap \{\tilde{\rho} \geq \delta\}.$$

Then, for all $x \in \omega$, we have

$$|h(x_0)| \leq |h(x)| + \int_{x_0}^{x} |h_x(y)|\, dy \leq |h(x)| + R^{1/2} \left(\int_{\mathbb{R}} |h_x(y)|^2\, dy \right)^{1/2}.$$

Integrating with respect to x in ω, we deduce that

$$\begin{aligned} |h(x_0)| &\leq \frac{1}{|\omega|} \int_{\omega} |h|\, dx + R^{1/2} \left(\int_{\mathbb{R}} |h_x|^2\, dx \right)^{1/2} \\ &\leq \frac{1}{|\omega|^{1/2}} \left(\int_{\omega} |h|^2\, dx \right)^{1/2} + R^{1/2} \left(\int_{\mathbb{R}} |h_x|^2\, dx \right)^{1/2}. \end{aligned}$$

Finally, since $\tilde{\rho} \geq \delta$ in ω, we have

$$|h(x_0)| \leq \frac{1}{\delta^{1/2} |\omega|^{1/2}} \left(\int_{\omega} \tilde{\rho} |h|^2\, dx \right)^{1/2} + R^{1/2} \left(\int_{\mathbb{R}} |h_x|^2\, dx \right)^{1/2},$$

and since $|\omega| \geq 1$, the result follows. ◻

Proof of Lemma 5.2. To prove the lemma, it is convenient to note that the system (1.1)–(1.3) can be rewritten in the form

$$\partial_t U_i + \partial_x A_i(U) = \partial_x \big[B_{ij}(U) \partial_x \left(D_j \mathscr{H}(U) \right) \big],$$

where $B(U)$ is a positive symmetric matrix and $D\mathscr{H}$ denotes the derivative (with respect to U) of the entropy $\mathscr{H}(U)$ associated with the flux $A(U)$. The existence of such an entropy is equivalent to the existence of an entropy flux function F such that

$$D_j F(U) = \sum_i D_i \mathscr{H}(U) D_j A_i(U) \tag{5.3}$$

for all U. Then strong solutions of (1.1)–(1.3) satisfy

$$\partial t \mathscr{H}(U) + \partial_x F(U) - \partial_x (B(U) \partial_x D\mathscr{H}(U)) D\mathscr{H}(U) = 0.$$

In our case, we have

$$A(U) = \begin{bmatrix} m \\ \dfrac{m^2}{\rho} + \rho^\gamma \end{bmatrix} = \begin{bmatrix} \rho u \\ \rho u^2 + \rho^\gamma \end{bmatrix}$$

and

$$B(U) = \mu(\rho) \left[\begin{array}{cc} 0 & 0 \\ 0 & 1 \end{array} \right].$$

Then, a careful computation (using (5.3)) yields

$$\begin{aligned}
&\partial_t \mathscr{H}(\tilde{U}|U) \\
&\quad = \left[\partial_t \mathscr{H}(\tilde{U}) + \partial_x(F(\tilde{U})) - \partial_x(B(\tilde{U})\partial_x D\mathscr{H}(\tilde{U}))D\mathscr{H}(\tilde{U})\right] \\
&\qquad - \left[\partial_t \mathscr{H}(U) + \partial_x F(U) - \partial_x(B(U)\partial_x D\mathscr{H}(U))D\mathscr{H}(U)\right] \\
&\qquad -\partial_x[F(\tilde{U}) - F(U)] \\
&\qquad -D^2\mathscr{H}(U)[\partial_t U + \partial_x A(U) - \partial_x\left(B(U)\partial_x\left(D\mathscr{H}(U)\right)\right)](\tilde{U} - U) \\
&\qquad -D\mathscr{H}(U)[\partial_t \tilde{U} + \partial_x A(\tilde{U}) - \partial_x(B(\tilde{U})\partial_x(D\mathscr{H}(\tilde{U})))] \\
&\qquad +D\mathscr{H}(U)[\partial_t U + \partial_x A(U) - \partial_x(B(U)\partial_x\left(D\mathscr{H}(U)\right))] \\
&\qquad +\partial_x[DF(U)(\tilde{U} - U)] \\
&\qquad +D\mathscr{H}(U)\partial_x[A(\tilde{U}|U)] \\
&\qquad +\partial_x[B(\tilde{U})\partial_x D\mathscr{H}(\tilde{U}) - B(U)\partial_x D\mathscr{H}(U)][D\mathscr{H}(\tilde{U}) - D\mathscr{H}(U)] \\
&\qquad +\partial_x(B(U)\partial_x D\mathscr{H}(U))D\mathscr{H}(\tilde{U}|U),
\end{aligned}$$

where the relative flux is defined by

$$A(\tilde{U}|U) = A(\tilde{U}) - A(U) - DA(U)\cdot(\tilde{U} - U).$$

Using the fact that $\tilde{U}$ and U are solutions satisfying the natural entropy inequality and equality, we deduce that

$$\begin{aligned}
&\partial_t \mathscr{H}(\tilde{U}|U) \\
&\quad \leq -\partial_x[F(\tilde{U}) - F(U)] + \partial_x[DF(U)(\tilde{U} - U)] \\
&\qquad +D\mathscr{H}(U)\partial_x[A(\tilde{U}|U)] \\
&\qquad +\partial_x[B(\tilde{U})\partial_x D\mathscr{H}(\tilde{U}) - B(U)\partial_x D\mathscr{H}(U)][D\mathscr{H}(\tilde{U}) - D\mathscr{H}(U)] \\
&\qquad +\partial_x(B(U)\partial_x D\mathscr{H}(U))D\mathscr{H}(\tilde{U}|U).
\end{aligned}$$

Integrating with respect to x and using (5.1), we deduce that

$$\begin{aligned}
&\frac{d}{dt}\int_{\mathbb{R}} \mathscr{H}(\tilde{U}|U)\,dx \\
&\leq -\int \partial_x[D\mathscr{H}(U)]A(\tilde{U}|U)\,dx \\
&\quad -\int_{\mathbb{R}}[B(\tilde{U})\partial_x D\mathscr{H}(\tilde{U}) - B(U)\partial_x D\mathscr{H}(U)]\partial_x[D\mathscr{H}(\tilde{U}) - D\mathscr{H}(U)]dx \\
&\quad +\int_{\mathbb{R}} \partial_x[B(U)\partial_x D\mathscr{H}(U)]D\mathscr{H}(\tilde{U}|U)\,dx.
\end{aligned}$$

Finally, we check that

$$\partial_x[D\mathscr{H}(U)]A(\tilde{U}|U) = (\partial_x u)[\rho(u-\tilde{u})^2 + (\gamma-1)p(\rho|\tilde{\rho})]\,,$$

$$\partial_x[B(U)\partial_x D\mathscr{H}(U)]D\mathscr{H}(\tilde{U}|U) = \frac{\partial_x(\mu(\rho)\partial_x u)}{\rho}\,(\tilde{\rho}-\rho)\,(u-\tilde{u}),$$

and

$$\begin{aligned}
&[B(\tilde U)\partial_x D\mathscr{H}(\tilde U) - B(U)\partial_x D\mathscr{H}(U)]\partial_x[D\mathscr{H}(\tilde U) - D\mathscr{H}(U)]\\
&\qquad = [\mu(\tilde\rho)\partial_x\tilde u - \mu(\rho)\partial_x u]\,\partial_x\,[\tilde u - u]\\
&\qquad = \mu(\tilde\rho)\,[\partial_x\tilde u - \partial_x u]^2 + \partial_x u[\mu(\tilde\rho) - \mu(\rho)]\partial_x\,[\tilde u - u]\,.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\frac{d}{dt}\int_{\mathbb{R}} \mathscr{H}(\tilde U|U)\,dx + \int_{\mathbb{R}} \mu(\tilde\rho)\,[\partial_x(\tilde u - u)]^2\\
&\quad \le C\int |\partial_x u|\mathscr{H}(\tilde U|U)\,dx\\
&\qquad - \int_{\mathbb{R}} \partial_x u[\mu(\tilde\rho) - \mu(\rho)]\,[\partial_x(\tilde u - u)]\;dx\\
&\qquad + \int_{\mathbb{R}} \frac{\partial_x(\mu(\rho)\partial_x u)}{\rho}\,(\tilde\rho - \rho)\,(u - \tilde u)\,dx.
\end{aligned}$$

which gives the lemma. ∎

Appendix. Proof of equality (3.15).

LEMMA A.1. *Let* (ρ, u) *satisfy* (3.11) *and*

$$\begin{cases} \partial_t\rho + \partial_x(\rho\, u) = 0,\\ \rho(x,0) = \rho_0(x). \end{cases} \tag{A.1}$$

Then (ρ, u) *satisfies* (3.15).

Proof. We denote by h_ε the convolution of any function h by a mollifier. Convoluting (A.1) by the mollifier, we get

$$\partial_t\rho_\varepsilon + \partial_x(\rho_\varepsilon u) = r_\varepsilon,$$

where

$$r_\varepsilon = \partial_x(\rho_\varepsilon u) - \partial_x(\rho u)_\varepsilon.$$

Since ρ_ε is now a smooth function, a straightforward computation yields

$$\begin{aligned}
&\frac{d}{dt}\int \rho_\varepsilon\frac{(\varphi(\rho_\varepsilon)_x)^2}{2}\,dx\\
&\qquad - \int (\rho_\varepsilon)^2\varphi'(\rho_\varepsilon)\varphi(\rho_\varepsilon)_x u_{xx}\,dx\\
&\qquad - \int (2\rho_\varepsilon\varphi'(\rho_\varepsilon) + (\rho_\varepsilon)^2\varphi''(\rho_\varepsilon))(\rho_\varepsilon)_x\varphi(\rho_\varepsilon)_x u_x\,dx\\
&\qquad\qquad = \int \rho_\varepsilon\varphi(\rho_\varepsilon)_x(\varphi'(\rho_\varepsilon)r_\varepsilon)_x\,dx.
\end{aligned} \tag{A.2}$$

In order to pass to the limit $\varepsilon \to 0$, we note that

$$\rho_\varepsilon - \overline{\rho} \longrightarrow \rho - \overline{\rho} \quad \text{ in } L^\infty(0,T;H^1(\mathbb{R})) \text{ strong},$$

which is enough to take the limit in the left-hand side of (A.2) (note that it implies the strong convergence in $L^\infty(0,T;L^\infty(\mathbb{R}))$). To show that the right-hand side goes to zero, we only need to show that r_ε goes to zero in $L^2(0,T;H^1(\mathbb{R}))$ strong (and thus in $L^2(0,T;L^\infty(\mathbb{R}))$). We write

$$\begin{aligned}\partial_x r_\varepsilon = 2[&\partial_x\rho_\varepsilon\partial_x u - (\partial_x\rho\,\partial_x u)_\varepsilon]\\ &+\rho_\varepsilon\partial_{xx}u - (\rho\,\partial_{xx}u)_\varepsilon\\ &+\partial_{xx}\rho_\varepsilon u - (\partial_{xx}\rho\,u)_\varepsilon.\end{aligned}$$

The first two terms converge to zero thanks to the strong convergence of ρ_ε in $L^\infty(0,T;H^1(\mathbb{R}))$. For the last term, we note that

$$\partial_x\rho\in L^\infty(0,T;L^2(\mathbb{R}))\quad\text{ and }\quad u\in L^2(0,T;W^{1,\infty}(\mathbb{R}))$$

so the strong convergence to zero in $L^2((0,T)\times\mathbb{R})$ follows from Lemma II.1 in [9] from DiPerna and Lions. $\square$

REFERENCES

[1] H. Amann, *Compact embeddings of vector-valued Sobolev and Besov spaces*, Glas. Mat. Ser. III, 35 (2000), pp. 161–177. Dedicated to the memory of Branko Najman.

[2] D. Bresch and B. Desjardins, *Sur un modèle de Saint-Venant visqueux et sa limite quasi-géostrophique*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 1079–1084.

[3] D. Bresch and B. Desjardins, *Existence of global weak solutions for a 2D viscous shallow water equations and convergence to the quasi-geostrophic model*, Comm. Math. Phys., 238 (2003), pp. 211–223.

[4] D. Bresch and B. Desjardins, *Some diffusive capillary models of Korteweg type.*, C. R. Math. Acad. Sci. Paris, Section Mécanique, 332 (2004), pp. 881–886.

[5] D. Bresch and B. Desjardins, *On the existence of global weak solutions to the Navier-Stokes equations for viscous compressible and heat conducting fluids*, J. Math. Pures Appl. (9), 87 (2007), pp. 57–90.

[6] D. Bresch, B. Desjardins, and C.-K. Lin, *On some compressible fluid models: Korteweg, lubrication, and shallow water systems*, Comm. Partial Differential Equations, 28 (2003), pp. 843–868.

[7] S. Chapman and T. G. Cowling, *The mathematical theory of non-uniform gases. An account of the kinetic theory of viscosity, thermal conduction and diffusion in gases*, 3rd ed., prepared in co-operation with D. Burnett, Cambridge University Press, London, 1970.

[8] C. M. Dafermos, *The second law of thermodynamics and stability*, Arch. Rational Mech. Anal., 70 (1979), pp. 167–179.

[9] R. J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[10] E. Feireisl, *On the motion of a viscous, compressible, and heat conducting fluid*, Indiana Univ. Math. J., 53 (2004), pp. 1705–1738.

[11] E. Feireisl, A. Novotný, and H. Petzeltová, *On the existence of globally defined weak solutions to the Navier-Stokes equations*, J. Math. Fluid Mech., 3 (2001), pp. 358–392.

[12] D. Hoff, *Global existence for 1D, compressible, isentropic Navier-Stokes equations with large initial data*, Trans. Amer. Math. Soc., 303 (1987), pp. 169–181.

[13] D. Hoff, *Global solutions of the Navier-Stokes equations for multidimensional compressible flow with discontinuous initial data*, J. Differential Equations, 120 (1995), pp. 215–254.

[14] D. Hoff, *Global solutions of the equations of one-dimensional, compressible flow with large data and forces, and with differing end states*, Z. Angew. Math. Phys., 49 (1998), pp. 774–785.

[15] D. Hoff and J. Smoller, *Non-formation of vacuum states for compressible Navier-Stokes equations*, Comm. Math. Phys., 216 (2001), pp. 255–276.

[16] A. V. Kazhikhov and V. V. Shelukhin, *Unique global solution with respect to time of initial-boundary value problems for one-dimensional equations of a viscous gas*, Prikl. Mat. Meh., 41 (1977), pp. 282–291.

[17] P.-L. Lions, *Mathematical Topics in Fluid Mechanics, Vol. 2, Compressible Models*, Oxford Lecture Series in Mathematics and Its Applications 10, The Clarendon Press, Oxford University Press, New York, 1998.

[18] T.-P. Liu, Z. Xin, and T. Yang, *Vacuum states for compressible flow*, Discrete Contin. Dynam. Systems, 4 (1998), pp. 1–32.

[19] A. Matsumura and T. Nishida, *The initial value problem for the equations of motion of compressible viscous and heat-conductive fluids*, Proc. Japan Acad. Ser. A Math. Sci., 55 (1979), pp. 337–342.

[20] A. Mellet and A. Vasseur, *On the barotropic compressible Navier-Stokes equation*, Comm. Partial Differential Equations, 32 (2007), pp. 431–452.

[21] M. Okada, Š. Matušů-Nečasová, and T. Makino, *Free boundary problem for the equation of one-dimensional motion of compressible gas with density-dependent viscosity*, Ann. Univ. Ferrara Sez. VII (N.S.), 48 (2002), pp. 1–20.

[22] D. Serre, *Solutions faibles globales des équations de Navier-Stokes pour un fluide compressible*, C. R. Acad. Sci. Paris Sér. I Math., 303 (1986), pp. 639–642.

[23] D. Serre, *Sur l'équation monodimensionnelle d'un fluide visqueux, compressible et conducteur de chaleur*, C. R. Acad. Sci. Paris Sér. I Math., 303 (1986), pp. 703–706.

[24] V. V. Shelukhin, *Motion with a contact discontinuity in a viscous heat conducting gas*, Dinamika Sploshn. Sredy, 57 (1982), pp. 131–152.

[25] V. V. Shelukhin, *Evolution of a contact discontinuity in the barotropic flow of a viscous gas*, Prikl. Mat. Mekh., 47 (1983), pp. 870–872.

[26] V. V. Shelukhin, *On the structure of generalized solutions of the one-dimensional equations of a polytropic viscous gas*, Prikl. Mat. Mekh., 48 (1984), pp. 912–920.

[27] V. V. Shelukhin, *Boundary value problems for equations of a barotropic viscous gas with nonnegative initial density*, Dinamika Sploshn. Sredy, 74 (1986), pp. 108–125, 162–163.

[28] V. V. Shelukhin, *A shear flow problem for the compressible Navier-Stokes equations*, Internat. J. Non-Linear Mech., 33 (1998), pp. 247–257.

[29] V. A. Solonnikov, *The solvability of the initial-boundary value problem for the equations of motion of a viscous compressible fluid*, Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI), 56 (1976), pp. 128–142, 197.

[30] V. A. Vaĭgant, *Nonhomogeneous boundary value problems for equations of a viscous heat-conducting gas*, Dinamika Sploshn. Sredy, 97 (1990), pp. 3–21, 212.

[31] T. Yang, Z.-a. Yao, and C. Zhu, *Compressible Navier-Stokes equations with density-dependent viscosity and vacuum*, Comm. Partial Differential Equations, 26 (2001), pp. 965–981.

[32] T. Yang and C. Zhu, *Compressible Navier-Stokes equations with degenerate viscosity coefficient and vacuum*, Comm. Math. Phys., 230 (2002), pp. 329–363.

SIAM J. MATH. ANAL.
Vol. 39, No. 5, pp. 1367–1401

ASYMPTOTIC ANALYSIS OF PHASE FIELD FORMULATIONS OF BENDING ELASTICITY MODELS*

XIAOQIANG WANG†

Abstract. In this paper, we give the asymptotic analysis of sharp interface analysis of the phase field function in some phase field models for Willmore's problem and equilibrium lipid bilayer cell membrane problems. We derive the explicit expression of the asymptotic expansion of the phase field functions minimizing the Willmore energy. Based on the structure of the phase field functions obtained via the asymptotic analysis, we can then demonstrate the consistency of phase field models and the sharp interface models. Also some error estimates of energy and Euler number formulae are further analyzed. Some numerical experiments are performed to verify our assumptions and results. The results of this paper lead to a better understanding of the structure of the phase field functions in the phase field models for Willmore's problem and the equilibrium configurations of the lipid vesicle membranes.

Key words. phase field, Willmore's problem, elastic bending energy, lipid membrane, asymptotic expansion, spectral method, numerical simulation, approximation

AMS subject classifications. 35B40, 74K15, 92B05

DOI. 10.1137/060663519

1. Introduction.

Willmore's problem [26] is a classical problem to minimize the mean curvature square energy of a compact surface Γ:

$$W = \int_\Gamma H^2 \, dS, \tag{1.1}$$

where $H = (k_1 + k_2)/2$ is the mean curvature and k_1 and k_2 are the two principal curvatures. For the equilibrium surface Γ minimizing the Willmore energy, the mean curvature H and the Gaussian curvature $K = k_1 k_2$ of Γ satisfy

$$\Delta_\Gamma H + 2H(H^2 - K) = 0, \tag{1.2}$$

which is the Euler–Lagrange equation of Willmore's problem.

Willmore's problem is closely related to the equilibrium shapes of biological lipid bilayer membranes, as shown in the bending elasticity model, first developed by Canham, Evans, and Helfrich [15, 19, 21]. Biomembranes are mostly composed by lipid molecules, which form a special bilayer structure of the membrane [20]. Due to the bending property of the lipid bilayer structure, in the bending elasticity model, the elastic bending energy is formulated by

$$W_c = \frac{\kappa}{2} \int_\Gamma (H - c)^2 \, dS, \tag{1.3}$$

where H is the mean curvature of the membrane surface Γ, c is the spontaneous curvature used to describe some physical or chemical difference between the inside and the outside of the membrane, and κ is the bending rigidity of the bilayer membrane.

*Received by the editors June 21, 2006; accepted for publication (in revised form) June 5, 2007; published electronically January 9, 2008.
http://www.siam.org/journals/sima/39-5/66351.html
†School of Computational Science and Department of Mathematics, Florida State University, Tallahassee, FL 32306-4120 (xwang@scs.fsu.edu).

The equilibrium membrane configurations are the minimizers of the elastic bending energy (1.3). Some constraints may apply, such as a given surface area and fixed cell volume to account for the effects of density change and osmotic pressure [7].

Some analytical and computational methods for computing the equilibrium shapes have been developed in the past few years [1, 2, 4, 16, 22, 27]. The computational methods include the direct methods such as front tracking [23, 14], volume of fluid (VOF) [18], and level set methods [3]. Another important computational method was developed recently in [7] and [11, 12, 8, 24, 25], using a phase field function to build up a general energetic variational framework. There are several advantages to using the phase field method. It can easily handle the topological changes of the shapes, and it can be formulated within a unified energetic framework, easy to understand and implement.

The basic idea of phase field formulations is to introduce a phase field function φ defined on a physical (computational) domain Ω to label the inside and the outside of the membrane Γ. We visualize that the level set $\{x : \varphi(x) = 0\}$ gives the surface Γ, while $\{x : \varphi(x) > 0\}$ represents the inside of Γ and $\{x : \varphi(x) \leq 0\}$ the outside. Define the following modified Willmore energy:

$$W(\varphi) = \int_\Omega \frac{\epsilon}{2} \left(\Delta\varphi - \frac{1}{\epsilon^2}\varphi(\varphi^2 - 1) \right)^2 dx, \tag{1.4}$$

where ϵ is a *transition* parameter that is taken to be very small. $W(\varphi)$ is a phase field translation of mean curvature energy (1.1) or the elastic bending energy (1.3) without the spontaneous curvature. For the elastic bending energy (1.3) with spontaneous curvature, the phase field translation is [11]

$$W_c(\varphi) = \int_\Omega \frac{\epsilon}{2} \left(\Delta\varphi - \frac{1}{\epsilon^2}(\varphi^2 - 1)(\varphi + \sqrt{2}c\epsilon) \right)^2 dx, \tag{1.5}$$

and the goal of the phase field model is to minimize $W(\varphi)$. Some constraints such as the fixed inside volume and the surface area may apply. In [6, 7, 8], some numerical schemes and discretization techniques have been developed for the phase field model. The theoretical analysis to the convergence of those numerical schemes is given in [5]. They have been successfully implemented and used in the numerical simulation of the membrane deformation. Various equilibrium solution branches and energy diagrams have been obtained, including interesting new three-dimensional solutions. Also, the shapes found in numerical experiments are coincident with the biophysical experiments and other theoretical methods for this problem.

There are basic issues about this phase field model requiring further consideration. One of them is to prove the consistency of the phase field model with the sharp interface Willmore's problem or the bending elasticity model for small ϵ. In other words, it is needed to verify that when phase field function φ minimizes $W(\varphi)$, the zero level set of φ is the desired surface Γ minimizing the Willmore energy (1.1) or the elastic bending energy (1.3), as the interfacial width parameter $\epsilon \to 0$. In [10], some positive results have been presented regarding the consistency. Some preliminary analyses for this consistency problem have been provided there, based on an assumption that the phase field function can be asymptotically expanded by

$$\varphi(x) = q(d(x)/\epsilon) + \epsilon h + g, \tag{1.6}$$

where $q \in C^2(\mathbf{R})$, $h \in C^2(\Omega)$ are independent of ϵ, and $\|\nabla^k g\|_{L^\infty} = o(\epsilon)$ for $k = 0, \ldots, 4$.

Our recent numerical experiments indicate that the above ansatz assumption is too restrictive in general. The purpose of this paper is to weaken the asymptotic expansion assumptions in formula (1.6) and derive the detailed expressions of the terms up to the third order. For example, we prove that the expansion to the second order term is

$$\varphi(x) = \tanh\left(\frac{d(x)}{\sqrt{2}\epsilon}\right) + 4\epsilon^2(H^2 - K)p\left(\frac{d(x)}{\epsilon}\right) + O(\epsilon^3), \tag{1.7}$$

where p is a given function independent of ϵ and Γ. Obviously, the second order term does not satisfy $\|\nabla^k g\|_{L^\infty} = o(\epsilon)$ for $k = 0, \ldots, 4$. Further expansion to the third order term reveals more details of the structure of the phase field function φ so that the consistency of the phase field model and the sharp interface model can still be verified.

The asymptotic analysis in this paper provides a firm mathematical foundation to the phase field models. With this theory, not only do we have the consistency of a phase field model with sharp interface models, but we can also perform error estimates of the phase field models, which leads to some very interesting identities. More importantly, it leads to a better understanding of the phase field formulations, which may shed light on the phase field modeling of other kinds of problems.

The paper is organized as follows. In the second section, we first make some assumptions on the asymptotic expansion of the phase field function, and the exact formulation for the leading, first, second, and third order terms are obtained. Based on the third order approximation, in the third section we can easily get the equilibrium state equation and the Willmore flow force, which can be used to prove the consistency of the phase field model and the sharp interface model. We generalize the problem to the spontaneous curvature case in section four. The error estimations of the phase field models are given in the fifth section. An interesting relationship between some Euler number formulae and the elastic bending energy is also derived in this section. Some numerical experiments are performed in the sixth section to verify our theory. Finally we draw the conclusion and give some further considerations.

2. Asymptotic analysis of phase field functions. In this section, we will give the asymptotic analysis of phase field functions for Willmore's problem. First, we will introduce some assumptions for the asymptotic expansions of phase field functions based on some geometric notation. Then some very useful lemmas will be introduced, from which we derive the forms of the expansion term by term.

2.1. Geometric notation and phase field model assumptions. Suppose the surface Γ is smooth and compact. We can find a constant c, such that within the domain $D_\Gamma = \{x|\mathrm{dist}(x, \Gamma) < c\}$, we can have a set of smooth surfaces $\{\Gamma_l\}_{i \in A}$ such that

$$D_\Gamma \subset \bigcup_{l \in A} \Gamma_l, \qquad \Gamma_a \cap \Gamma_b = \emptyset \quad \text{if} \quad a \neq b.$$

Therefore, for any point $x \in D_\Gamma$, we can get $a \in A$ such that $x \in \Gamma_a$. We can define a *normal direction* $\vec{n}(x)$ by the normal direction of Γ_a at point x. Further, we denote the *integral curve* along the normal direction through x by γ_x, and we assume that the constant c is small enough such that every integral curve has a unique intersection with any Γ_l. The *distance* from any point $x \in D_\Gamma$ to Γ_i is defined by the length of the integral curve γ_x between Γ_a and Γ_i. We call Γ_a and Γ_b *parallel* if the distance between any point of Γ_a and Γ_b is a constant, and we denote the *distance* of Γ_a and Γ_b by this constant.

Without proof, we can always find a set of parallel surfaces of Γ. Denote Γ by Γ_0. A *sign* is the distance between Γ and any parallel surface, positive to the inside surfaces and negative to the outside surfaces. We use the signed distance to denote the index of the surface. For example, Γ_d has a distance d to Γ_0.

Suppose $x \in \Gamma_d$, and the *distance function* $d(x)$ is defined by d, the distance between Γ_d and Γ, or the signed length from point x along the integral curve γ_x of normal direction to Γ_0. With this definition, the normal direction

$$\vec{n}(x) = \nabla d(x).$$

For any point $x \in \Gamma_l$, the *mean curvature* $H(x)$ and *Gaussian curvature* $K(x)$ are defined by the mean curvature and Gaussian curvature of Γ_l at point x. It is known that

$$H(x) = -\frac{1}{2}\Delta d(x),$$

and $H(x)$ is a smooth function in domain D_Γ.

We also define $x|_\Gamma = \gamma_x \cap \Gamma$, the *projection* of point x on Γ along the integral curve of ∇d.

As we are mainly concerned with the phase field function theory, some of the above geometry facts are used without any proof.

Remark 2.1. Note that we do not define the distance function $d(x)$ by the Euclid distance between x and Γ.

For a phase field function φ defined on a computational domain Ω, for any $\epsilon > 0$, denote

$$f(\varphi) = \Delta\varphi - \frac{1}{\epsilon^2}(\varphi^2 - 1)\varphi. \tag{2.1}$$

The variational problem is to minimize

$$W(\varphi) = \frac{\epsilon}{2}\int_\Omega f(\varphi(x))^2\, dx. \tag{2.2}$$

Denote the minimizer of (2.2) by φ. Denote Γ_0 by $\{x|\varphi(x) = 0\}$.

Now we make the following assumptions:

(A1) Γ is a smooth compact surface, and the set of parallel surfaces Γ_l exists.

(A2) φ is a function that can be expanded by

$$\varphi(x) = \sum_{n=0}^{\infty} \epsilon^n q_n\left(\frac{d(x)}{\epsilon}, x|_\Gamma\right),$$

where $q_i \in C^\infty(\bar{\mathbf{R}} \times \Gamma)$ independent of ϵ and bounded.

(A3) $\Phi(\epsilon, t, x) = \sum_{n=0}^{\infty} \epsilon^n q_n(t, x) \in C^\infty(\mathbf{R}^2 \times \Gamma)$.

(A4) $\lim_{\epsilon\to 0}\varphi(x) = 1$ if $d(x) > 0$; $\lim_{\epsilon\to 0}\varphi(x) = -1$ if $d(x) < 0$; and $\lim_{\epsilon\to 0} \nabla^n\varphi(x) = 0$ for $x \in \partial D_\Gamma$ and any $n \geq 0$.

Remark 2.2. Assumption (A2) is the main point. It describes how φ depends on ϵ. Assumption (A1) is natural. (A3) ensures boundedness of the partial derivatives of φ on x. (A4) gives the boundary condition for doing integration by parts.

These assumptions weaken the ansatz in [10], which assumes that q_0 is only a function of d/ϵ, and q_1 and q_2 depend only on x. Here we assume that q_0, q_1, and

q_2 depend on both d/ϵ and x, as in a standard two-scale analysis. Our result in section 2.3 shows that q_2 does depend on both, and not x alone.

For simplicity, we ignore the parameters $\frac{d(x)}{\epsilon}$ and $x|_\Gamma$ when doing so would cause no confusion, but one should always keep them in mind. This may also apply to the notation $O(\epsilon^k)$ and $o(\epsilon^k)$. For example,

$$q_n = q_n\left(\frac{d(x)}{\epsilon}, x|_\Gamma\right). \tag{2.3}$$

We also denote the kth directive

$$q_n^{(k)} = (\partial_1^k q_n)\left(\frac{d(x)}{\epsilon}, x|_\Gamma\right). \tag{2.4}$$

Sometimes, we denote the error term functions

$$e_i = \sum_{n=i}^{\infty} \epsilon^{n-i} q_n. \tag{2.5}$$

Using the above notation, we have the following two useful identities:

$$\nabla q_n = \frac{1}{\epsilon} q_n' \nabla d + \nabla_\Gamma q_n = \frac{1}{\epsilon} q_n' \nabla d + O(1), \tag{2.6}$$

$$\begin{aligned} \Delta q_n &= \frac{1}{\epsilon^2} q_n'' + \frac{1}{\epsilon} q_n' \Delta d + \frac{1}{\epsilon} \nabla_\Gamma q_n' \cdot \nabla d + \Delta_\Gamma q_n \\ &= \frac{1}{\epsilon^2} q_n'' + \frac{1}{\epsilon} q_n' \Delta d + \Delta_\Gamma q_n = \frac{1}{\epsilon^2} q_n'' - \frac{2H}{\epsilon} q_n' + O(1). \end{aligned} \tag{2.7}$$

2.2. Several lemmas. Now we introduce several useful lemmas.

LEMMA 2.3.

$$\nabla H \cdot \nabla d = 2H^2 - K, \tag{2.8}$$

$$\nabla K \cdot \nabla d = 2HK, \tag{2.9}$$

$$\Delta H = \Delta_\Gamma H + 4H(H^2 - K). \tag{2.10}$$

Proof. Suppose two principal curvatures are $k_1(x)$ and $k_2(x)$. It is easy to know that $\nabla k_i \cdot \nabla d = k_i^2$ for $i = 1, 2$. Therefore

$$\nabla H \cdot \nabla d = \frac{1}{2}(k_1^2 + k_2^2) = 2H^2 - K,$$

$$\nabla K \cdot \nabla d = k_1 k_2^2 + k_1^2 k_2 = 2HK,$$

$$\begin{aligned} \Delta H &= \Delta_\Gamma H + \nabla \cdot ((\nabla H \cdot \nabla d)\nabla d) \\ &= \Delta_\Gamma H + \nabla \cdot ((2H^2 - K)\nabla d) \\ &= \Delta_\Gamma H + \nabla(2H^2 - K) \cdot \nabla d + (2H^2 - K)\Delta d \\ &= \Delta_\Gamma H + 4H(2H^2 - K) - 2HK - 2H(2H^2 - K) \\ &= \Delta_\Gamma H + 4H(H^2 - K). \qquad \square \end{aligned}$$

LEMMA 2.4. *For two smooth functions f and g, assume $f(d(x)/\epsilon, x|_\Gamma)|_{\partial\Omega} = f'(d(x)/\epsilon, x|_\Gamma)|_{\partial\Omega} = g(d(x)/\epsilon, x|_\Gamma)|_{\partial\Omega} = g'(d(x)/\epsilon, x|_\Gamma)|_{\partial\Omega} = 0$. We have*

$$\int_\Omega fg'\,dx = -\int_\Omega (f' - 2Hf\epsilon)g\,dx, \tag{2.11}$$

$$\int_\Omega Hfg'\,dx = \int_\Omega (Kf\epsilon - Hf')g\,dx, \tag{2.12}$$

$$\int_\Omega Kfg'\,dx = -\int_\Omega Kf'g\,dx, \tag{2.13}$$

$$\int_\Omega fg''\,dx = \int_\Omega (f'' - 4Hf'\epsilon + 2Kf\epsilon^2)g\,dx, \tag{2.14}$$

$$\int_\Omega Hfg''\,dx = \int_\Omega (Hf'' - 2Kf'\epsilon)g\,dx, \tag{2.15}$$

$$\int_\Omega H^2fg'\,dx = -\int_\Omega \big(2H(H^2-K)f\epsilon + H^2f'\big)g\,dx, \tag{2.16}$$

where $H(x)$ is the mean curvature and $K(x)$ is the Gaussian curvature.

Remark 2.5. *Here we hide the parameters $d(x)/\epsilon$ and $x|_\Gamma$ for f, g, and their derivatives.*

Proof. Because the proofs are similar, here we give the proof for only the first three equations.

Since $\nabla g = \frac{1}{\epsilon}g'\nabla d + \nabla_\Gamma g$ and $\nabla_\Gamma g \cdot \nabla d = 0$, we know that

$$g' = \epsilon(\nabla g \cdot \nabla d - \nabla_\Gamma g \cdot \nabla d) = \epsilon \nabla g \cdot \nabla d.$$

Then

$$\begin{aligned}\int_\Omega fg'\,dx &= \epsilon\int_\Omega f\nabla g\cdot\nabla d\,dx\\ &= -\epsilon\int_\Omega g\nabla f\cdot\nabla d + gf\Delta d\,dx\\ &= -\int_\Omega (f' - 2Hf\epsilon)g\,dx.\end{aligned}$$

From Lemma 2.3, we have

$$\begin{aligned}\int_\Omega Hfg'\,dx &= \epsilon\int_\Omega Hf\nabla g\cdot\nabla d\,dx\\ &= -\epsilon\int_\Omega \nabla(Hf)\cdot\nabla dg - 2H(Hfg)\,dx\\ &= -\int_\Omega (\nabla H\cdot\nabla d)fg\epsilon + Hf'g - 2H^2fg\epsilon\,dx\\ &= -\int_\Omega (2H^2-K)fg\epsilon + Hf'g - 2H^2fg\epsilon\,dx\\ &= \int_\Omega (Kf\epsilon - Hf')g\,dx,\end{aligned}$$

$$\begin{aligned}\int_\Omega Kfg'\,dx &= \epsilon\int_\Omega Kf\nabla g\cdot\nabla d\,dx\\ &= -\epsilon\int_\Omega \nabla(Kf)\cdot\nabla dg - 2H(Kfg)\,dx\\ &= -\int_\Omega(\nabla K\cdot\nabla d)fg\epsilon + Kf'g - 2HKfg\epsilon\,dx\\ &= -\int_\Omega Kf'g\,dx. \qquad\square\end{aligned}$$

LEMMA 2.6. *For two smooth functions f and g, assume $f(d(x)/\epsilon, x|_\Gamma)|_{\partial\Omega} = g(d(x)/\epsilon, x|_\Gamma)|_{\partial\Omega} = 0$. We have*

$$\int_\Omega f\Delta_\Gamma g\,dx = \int_\Omega \Delta_\Gamma fg\,dx. \tag{2.17}$$

Proof. First we have

$$\nabla g = \frac{1}{\epsilon}g'\nabla d + \nabla_\Gamma g,$$

$$\Delta g = \frac{1}{\epsilon^2}g'' - \frac{2H}{\epsilon}g' + \Delta_\Gamma g.$$

From Lemmas 2.3 and 2.4, we have

$$\begin{aligned}\int_\Omega f\Delta_\Gamma g\,dx &= \int_\Omega f\left(\Delta g - \frac{1}{\epsilon^2}g'' + \frac{2H}{\epsilon}g'\right)\,dx\\ &= \int_\Omega \Delta fg - \frac{1}{\epsilon^2}(f'' - 4Hf'\epsilon + 2Kf\epsilon^2)g + \frac{2}{\epsilon}(Kf\epsilon - Hf')g\,dx\\ &= \int_\Omega\left(\Delta f - \frac{1}{\epsilon^2}f'' + \frac{2H}{\epsilon}f'\right)g\\ &= \int_\Omega \Delta_\Gamma fg\,dx. \qquad\square\end{aligned}$$

LEMMA 2.7. *Suppose $f\in C^0(\Omega)$ and $p\in L^1(\mathbf{R})$ satisfy the bound*

$$\max_{|t|>s}|p(t)t| \le \frac{C}{s^m}, \quad m>1. \tag{2.18}$$

Then

$$\lim_{\epsilon\longrightarrow 0}\frac{1}{\epsilon}\int_\Omega p(d(x)/\epsilon)f(x)\,dx = \int_{-\infty}^{\infty}p(t)\,dt\int_\Gamma f(s)dS. \tag{2.19}$$

Please refer to [10] for the proof of Lemma 2.7.

LEMMA 2.8.

$$\begin{aligned}&\frac{1}{\epsilon}\int_\Omega p(d(x)/\epsilon)f(x|_\Gamma)\,dx\\ &\quad= \int_{-\infty}^{\infty}p(t)\,dt\int_\Gamma f(s)dS + \epsilon\int_{-\infty}^{\infty}p(t)t\,dt\int_\Gamma 2f(s)H(s)dS\\ &\qquad + \epsilon^2\int_{-\infty}^{\infty}p(t)t^2\,dt\int_\Gamma f(s)K(s)dS + o(\epsilon^3).\end{aligned} \tag{2.20}$$

Proof. The volume differential dx is a product of area differential $d\Gamma_s$, and ds is the length differential of the integral curve along ∇d. As $d\Gamma_s$ is a function of s, it is known that $(d\Gamma_s)' = -2Hd\Gamma_s$. (One can get this using the Jacobian notion. Please refer to [10] for the proof of Lemma 2.7.) Then

$$(d\Gamma_s)'' = (-2Hd\Gamma_s)' = -2(\nabla H \cdot \nabla d)d\Gamma_s - 2H(-2Hd\Gamma_s) = 2Kd\Gamma_s.$$

It follows that

$$d\Gamma_s = (1 - 2Hs + Ks^2 + O(s^3))d\Gamma.$$

Now

$$\begin{aligned}
\frac{1}{\epsilon}\int_\Omega p(d(x)/\epsilon)f(x|_\Gamma)\,dx &= \frac{1}{\epsilon}\int_\Omega p(s/\epsilon)f(x|_\Gamma)\,d\Gamma_s ds \\
&= \frac{1}{\epsilon}\int_\Omega p(s/\epsilon)f(x|_\Gamma)\,(1 - 2Hs + Ks^2 + O(s^3))d\Gamma ds \\
&= \int_{-\infty}^{\infty} p(t)\,dt \int_\Gamma f(s)d\Gamma + \epsilon\int_{-\infty}^{\infty} p(t)t\,dt\int_\Gamma 2f(s)H(s)d\Gamma \\
&\quad + \epsilon^2\int_{-\infty}^{\infty} p(t)t^2\,dt\int_\Gamma f(s)K(s)d\Gamma + o(\epsilon^3).
\end{aligned}$$

Corresponding to our notation convention, we can replace $d\Gamma$ by dS, and this gives the lemma. $\square$

Furthermore, we can prove the following lemma.

LEMMA 2.9. *Suppose $f \in C^2(\Omega)$ and $p \in L^1(\mathbf{R})$ satisfy the bound*

$$\max_{|t|>s} |p(t)t^3| \le \frac{C}{s^m}, \quad m > 1. \tag{2.21}$$

Then

$$\begin{aligned}
&\frac{1}{\epsilon}\int_\Omega p(d(x)/\epsilon)f(x)\,dx \\
&\quad = \int_{-\infty}^{\infty} p(t)\,dt\int_\Gamma f(x)dS + \epsilon\int_{-\infty}^{\infty} p(t)t\,dt\int_\Gamma -2fH + \nabla f\cdot\nabla d\,dS \\
&\quad + \epsilon^2\int_{-\infty}^{\infty} p(t)t^2\,dt\int_\Gamma fK - 2(\nabla f\cdot\nabla d)H + \frac{1}{2}\nabla(\nabla f\cdot\nabla d)\cdot\nabla d\,dS \\
&\quad + O(\epsilon^3).
\end{aligned} \tag{2.22}$$

Proof. We can expand the function f along the integration curve of ∇d by

$$f(x) = f(x|_\Gamma) + d(\nabla f\cdot\nabla d)(x|_\Gamma) + \frac{1}{2}d^2(\nabla(\nabla f\cdot\nabla d)\cdot\nabla d)(x|_\Gamma) + O(d^3).$$

Applying Lemmas 2.7 and 2.8, we get

$$
\begin{aligned}
&\frac{1}{\epsilon}\int_\Omega p(d(x)/\epsilon) f(x)\,dx \\
&\quad= \frac{1}{\epsilon}\int_\Omega p(d(x)/\epsilon) f(x|_\Gamma)\,dx + \frac{1}{\epsilon}\int_\Omega p(d(x)/\epsilon) d(x)(\nabla f\cdot\nabla d)(x|_\Gamma)\,dx \\
&\qquad+ \frac{1}{\epsilon}\int_\Omega p(d(x)/\epsilon)\frac{1}{2}d^2(x)(\nabla(\nabla f\cdot\nabla d)\cdot\nabla d)(x|_\Gamma)\,dx + O(\epsilon^3) \\
&\quad= \int_{-\infty}^{\infty} p(t)\,dt\int_\Gamma f(x)dS + \epsilon\int_{-\infty}^{\infty} p(t)t\,dt\int_\Gamma -2fH dS + \epsilon^2\int_{-\infty}^{\infty} p(t)t^2\,dt\int_\Gamma fK dS \\
&\qquad+ \epsilon\int_{-\infty}^{\infty} p(t)t\,dt\int_\Gamma (\nabla f\cdot\nabla d)dS + \epsilon^2\int_{-\infty}^{\infty} p(t)t^2\,dt\int_\Gamma -2(\nabla f\cdot\nabla d)H dS \\
&\qquad+ \epsilon^2\int_{-\infty}^{\infty} p(t)t^2\,dt\int_\Gamma \frac{1}{2}(\nabla(\nabla f\cdot\nabla d)\cdot\nabla d)dS + O(\epsilon^3) \\
&\quad= \int_{-\infty}^{\infty} p(t)\,dt\int_\Gamma f(x)dS + \epsilon\int_{-\infty}^{\infty} p(t)t\,dt\int_\Gamma -2fH + \nabla f\cdot\nabla d\,dS \\
&\qquad+ \epsilon^2\int_{-\infty}^{\infty} p(t)t^2\,dt\int_\Gamma fK - 2(\nabla f\cdot\nabla d)H + \frac{1}{2}\nabla(\nabla f\cdot\nabla d)\cdot\nabla d\,dS + O(\epsilon^3). \qquad \square
\end{aligned}
$$

2.3. Asymptotic approximation. With assumptions (A1)–(A4), now we can get the exact form of q_i term by term.

THEOREM 2.10. *With assumptions* (A1)–(A4), *we have*

$$q_0(t) = \tanh\left(\frac{t}{\sqrt{2}}\right).$$

Remark 2.11. Because $q_0 \in C^\infty(\bar{\mathbf{R}}\times\Gamma)$, it actually means $q_0(t,x) = \tanh(\frac{t}{\sqrt{2}})$, a function depending only on the first parameter.

Proof. As $\varphi = q_0 + \epsilon e_1$, we have

$$
\begin{aligned}
f(\varphi) &= \Delta\varphi - \frac{1}{\epsilon^2}(\varphi^2-1)\varphi \\
&= \frac{1}{\epsilon^2}(q_0+\epsilon e_1)'' + \frac{1}{\epsilon}(q_0+\epsilon e_1)'\Delta d - \frac{1}{\epsilon^2}(\varphi^2-1)\varphi + O(1) \\
&= \frac{1}{\epsilon^2}(q_0'' - (q_0^2-1)q_0) + O(\epsilon^{-1}).
\end{aligned}
$$

To minimize $W(\varphi) = \frac{\epsilon}{2}\int_\Omega f^2\,dx$, we have that q_0 minimizes

$$\int_\Omega (q_0'' - (q_0^2-1)q_0)^2\,dx,$$

which results in

$$q_0'' - (q_0^2-1)q_0 = 0.$$

From assumption (A4), $q_0(\infty) = 1$, $q_0(-\infty) = -1$. It follows that

$$q_0(t) = \tanh\left(\frac{t}{\sqrt{2}}\right). \qquad \square$$

With Theorem 2.10, we have

$$\begin{aligned} f(q_0) &= \Delta q_0 - \frac{1}{\epsilon^2}(q_0^2 - 1)q_0 \\ &= \frac{1}{\epsilon} q_0' \Delta d + \frac{1}{\epsilon^2} q_0'' - \frac{1}{\epsilon^2}(q_0^2 - 1)q_0 = -\frac{2H}{\epsilon} q_0', \end{aligned} \tag{2.23}$$

and we can expand f by

$$\begin{aligned} f(\varphi) &= \Delta\varphi - \frac{1}{\epsilon^2}(\varphi^2 - 1)\varphi \\ &= \Delta(q_0 + \epsilon e_1) - \frac{1}{\epsilon^2}(q_0^2 - 1)q_0 - \frac{1}{\epsilon}(3q_0^2 - 1)e_1 - 3q_0 e_1^2 - \epsilon e_1^3 \\ &= f(q_0) + \epsilon\Delta e_1 - \frac{1}{\epsilon}(3q_0^2 - 1)e_1 - 3q_0 e_1^2 - \epsilon e_1^3 \\ &= -\frac{2H}{\epsilon} q_0' + \epsilon\Delta e_1 - \frac{1}{\epsilon}(3q_0^2 - 1)e_1 - 3q_0 e_1^2 - \epsilon e_1^3. \end{aligned} \tag{2.24}$$

THEOREM 2.12. *With assumptions* (A1)–(A4), *we have*

$$q_1 = 0.$$

Proof. From assumption (A3), $e_1 = O(\epsilon)$. From (2.24), we have

$$\begin{aligned} f(\varphi) &= -\frac{2H}{\epsilon} q_0' + \epsilon\Delta e_1 - \frac{1}{\epsilon}(3q_0^2 - 1)e_1 + O(1) \\ &= -\frac{2H}{\epsilon} q_0' + \epsilon\Delta q_1 - \frac{1}{\epsilon}(3q_0^2 - 1)q_1 + O(1) \\ &= -\frac{2H}{\epsilon} q_0' + \epsilon\left(-2H\frac{1}{\epsilon} q_1' + \frac{1}{\epsilon^2} q_1''\right) - \frac{1}{\epsilon}(3q_0^2 - 1)q_1 + O(1) \\ &= -\frac{2H}{\epsilon} q_0' + \frac{1}{\epsilon} q_1'' - \frac{1}{\epsilon}(3q_0^2 - 1)q_1 + O(1). \end{aligned}$$

To minimize $W(\varphi) = \frac{\epsilon}{2}\int_\Omega f^2\,dx$, we have that q_1 minimizes

$$\begin{aligned} W(\varphi) &= \frac{\epsilon}{2}\int_\Omega \left(-\frac{2H}{\epsilon} q_0' + \frac{1}{\epsilon} q_1'' - \frac{1}{\epsilon}(3q_0^2 - 1)q_1 + O(1)\right)^2 dx \\ &= \frac{\epsilon}{2}\int_\Omega \left(-\frac{2H}{\epsilon} q_0'\right)^2 + \left(\frac{1}{\epsilon} q_1'' \frac{1}{\epsilon}(3q_0^2 - 1)q_1\right)^2 dx \\ &\quad - \frac{1}{\epsilon}\int_\Omega H q_0'(q_1'' - (3q_0^2 - 1)q_1)\,dx + O(\epsilon). \end{aligned}$$

From Lemma 2.4

$$\int_\Omega H q_0' q_1'' \, dx = \int_\Omega H q_0''' q_1 \, dx + O(\epsilon).$$

Because $q_0'' = (q_0^2 - 1)q_0$, $q_0''' = (3q_0^2 - 1)q_0'$, we have

$$W(\varphi) = \frac{1}{2\epsilon} \int_\Omega 4H^2 (q_0')^2 + (q_1'' - (3q_0^2 - 1)q_1)^2 \, dx + O(\epsilon).$$

Therefore to minimize $W(\varphi)$, we get

$$q_1'' - (3q_0^2 - 1)q_1 = 0.$$

Since $q_0(0) = 0$, $\varphi(0) = 0$, assumption (A4) gives $q_1(0) = 0$, and since $q_0(\infty) = \varphi(\infty) = 1$, we have $q_1(\infty) = 0$. Similarly, $q_1(-\infty) = 0$. Then the only solution is that $q_1 = 0$. $\square$

THEOREM 2.13. *With assumptions* (A1)–(A4), *we have*

$$q_2 = 4(H^2 - K)p_2, \qquad i.e., \; q_2\left(\frac{d(x)}{\epsilon}, x|_\Gamma\right) = 4(H^2(x|_\Gamma) - K(x|_\Gamma))p\left(\frac{d(x)}{\epsilon}\right),$$

where $x|_\Gamma$ is the projection of point x on Γ along the integral curve of ∇d, and p_2 is an independent function satisfying the one-dimensional ODE

$$p_2''(t) - (3q_0^2(t) - 1)p_2(t) = \frac{1}{2} q_0'(t) t \tag{2.25}$$

with $p_2(\pm\infty) = p_2(0) = 0$.

Proof. Since $q_1 = 0$, $e_1 = \epsilon e_2$. From (2.24), we have

$$\begin{aligned} f(\varphi) &= -\frac{2H}{\epsilon} q_0' + \epsilon^2 \Delta e_2 - (3q_0^2 - 1)e_2 + O(\epsilon^2) \\ &= -\frac{2H}{\epsilon} q_0' - 2H\epsilon e_2' + e_2'' - (3q_0^2 - 1)e_2 + O(\epsilon^2). \end{aligned} \tag{2.26}$$

Now, $W(q_0)$ is already fixed due to Theorem 2.10, and

$$\begin{aligned} W(\varphi) - W(q_0) &= \frac{\epsilon}{2} \int_\Omega (f(\varphi) + f(q_0))(f(\varphi) - f(q_0)) \, dx \\ &= \frac{\epsilon}{2} \int_\Omega (2f(q_0) - 2H\epsilon e_2' + e_2'' - (3q_0^2 - 1)e_2 + O(\epsilon^2)) \\ &\quad (-2H\epsilon e_2' + e_2'' - (3q_0^2 - 1)e_2 + O(\epsilon^2)) \, dx \\ &= \frac{\epsilon}{2} \int_\Omega 2f(q_0)(-2H\epsilon e_2' + e_2'' - (3q_0^2 - 1)e_2 + O(\epsilon^2)) \\ &\quad + (-2H\epsilon e_2' + e_2'' - (3q_0^2 - 1)e_2 + O(\epsilon^2))^2 \, dx \\ &= R_1 + R_2. \end{aligned}$$

First from Lemma 2.4, we have

$$
\begin{aligned}
R_1 &= \frac{\epsilon}{2}\int_\Omega 2f(q_0)(-2H\epsilon e_2' + e_2'' - (3q_0^2-1)e_2 + O(\epsilon^2))\,dx \\
&= -\int_\Omega 2Hq_0'(-2H\epsilon e_2' + e_2'' - (3q_0^2-1)e_2 + O(\epsilon^2))\,dx \\
&= \int_\Omega 4H^2\epsilon q_0' e_2'\,dx - \int_\Omega 2Hq_0'(e_2'' - (3q_0^2-1)e_2) + O(\epsilon^2)\,dx \\
&= \int_\Omega -4H^2\epsilon q_0'' e_2 + O(\epsilon^2)\,dx - \int_\Omega 2(Hq_0''' - 2Kq_0''\epsilon)e_2 \\
&\quad - 2Hq_0'(3q_0^2-1)e_2 + O(\epsilon^2)\,dx \\
&= \int_\Omega -4(H^2-K)\epsilon q_0'' e_2 + O(\epsilon^2)\,dx
\end{aligned}
$$

and

$$
\begin{aligned}
R_2 &= \frac{\epsilon}{2}\int_\Omega (-2H\epsilon e_2' + e_2'' - (3q_0^2-1)e_2 + O(\epsilon^2))^2\,dx \\
&= \frac{\epsilon}{2}\int_\Omega (e_2'' - (3q_0^2-1)e_2)^2 + O(\epsilon)\,dx \\
&= \frac{\epsilon}{2}\int_\Omega (q_2'' - (3q_0^2-1)q_2)^2 + O(\epsilon)\,dx.
\end{aligned}
$$

Altogether, we have

$$
W(\varphi) - W(q_0) = \frac{\epsilon}{2}\int_\Omega -8(H^2-K)q_0''q_2 + (q_2'' - (3q_0^2-1)q_2)^2\,dx + O(\epsilon^2).
$$

Also, the curvature functions $H(x)$ and $K(x)$ are smooth functions around Γ, and from Lemma 2.9 we have

$$
\int_\Omega -8(H^2-K)q_0''q_2\,dx = \int_\Omega -8(H^2-K)|_\Gamma q_0''q_2\,dx + O(\epsilon^2).
$$

Now, taking $q_2 = 4(H^2-K)p_2$, that is, $q_2(d(x)/\epsilon, x|_\Gamma) = 4(H^2(x|_\Gamma) - K(x|_\Gamma))$ $p_2(d(x)/\epsilon, x|_\Gamma)$, we have $q_2'' = 4(H^2-K)p_2''$, and

$$
\begin{aligned}
&W(\varphi) - W(q_0) \\
&\quad = \epsilon\int_\Omega 16(H^2-K)|_\Gamma^2\left(-q_0''p_2 + \frac{1}{2}(p_2'' - (3q_0^2-1)q_2)^2\right) + O(\epsilon)\,dx \\
&\quad = \epsilon\int_\Gamma 16(H^2-K)^2\int_{\gamma_s} -q_0''p_2 + \frac{1}{2}(p_2'' - (3q_0^2-1)q_2)^2\,ds\,dS + O(\epsilon^3). \qquad (2.27)
\end{aligned}
$$

To minimize $W(\varphi)$, p_2 is independent of ϵ and $x|_\Gamma$, minimizing

$$
\int_{\gamma_s} -q_0''p_2 + \frac{1}{2}(p_2'' - (3q_0^2-1)p_2)^2\,ds.
$$

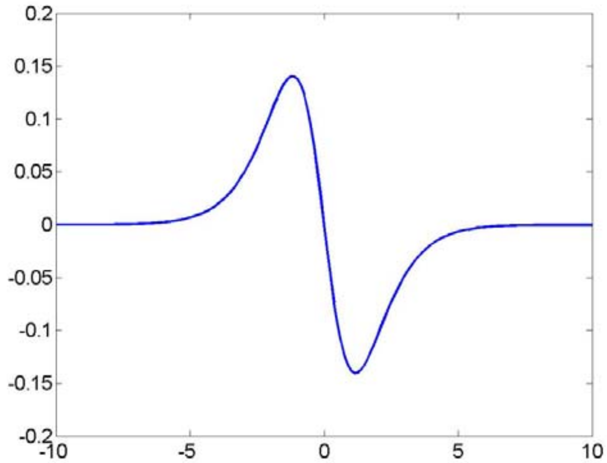


FIG. 2.1. *The second order term function $p_2(x)$.*

Changing the parameter d/ϵ to t, we have p_2 minimizing

$$E(p) = \int_{-\infty}^{\infty} -q_0(q_0^2-1)p_2 + \frac{1}{2}(p_2'' - (3q_0^2-1)p_2)^2\, dt.$$

We denote

$$s_2 = p_2'' - (3q_0^2-1)p_2. \tag{2.28}$$

Taking a variation of $E(p)$, we have

$$\frac{\delta E}{\delta p_2} = -(q_0^2-1)q_0 + s_2'' - (3q_0^2-1)s_2 = 0. \tag{2.29}$$

From assumption (A4) and $q_0(t) = \tanh(t/\sqrt{2})$ and $q_1 = 0$, we know that $p_2(\pm\infty) = p_2''(\pm\infty) = 0$. Then $s_2(\pm\infty) = 0$. Solving (2.29), we get

$$s_2 = \frac{1}{2}q_0' t.$$

Substituting it into (2.28) gives the result. □

Remark 2.14. Function p_2 can be numerically calculated by solving (2.25). A plot of p_2 is given in Figure 2.1. Theorem 2.13 states that φ depends not only on the distance function $d(x)$ but also on the local curvature properties.

For a sphere we always have $H^2(x) - K(x) = 0$ for every point $x \in \Omega$. Following Theorem 2.13 we directly have the following corollary.

COROLLARY 2.15. *If surface Γ is a sphere, we have $q_2 = 0$.*

Remark 2.16. Corollary 2.15 states that the phase field function is at least third order convergent in ϵ to the tanh profile for a sphere. Actually, we can easily verify that the phase field function is an exactly tanh function for a sphere.

THEOREM 2.17. *With assumptions* (A1)–(A4), *for given $x|_\Gamma$, we have $q_3(t, x|_\Gamma)$ minimizing*

$$\int_{-\infty}^{\infty} -Qq_3 + \frac{1}{2}(q_3'' - (3q_0^2-1)q_3)^2\, dt,$$

where $Q(t,x|_\Gamma) = 8H(H^2-K)(x|_\Gamma)[2q_0'(t) + 3q_0''(t)t] + 2\Delta_\Gamma H(x|_\Gamma)q_0'(t)$.

Proof. For computational purpose, we denote $G = 4(H^2-K)|_\Gamma$. Then $e_2 = q_2 + \epsilon e_3 = Gp_2 + \epsilon e_3$, and

$$\begin{aligned}\Delta q_2 = \nabla\cdot(\nabla q_2) &= \nabla\cdot\left(\nabla_\Gamma G p_2 + G\frac{p_2'}{\epsilon}\nabla d\right)\\ &= G\frac{p_2''}{\epsilon^2} - 2HG\frac{p_2'}{\epsilon} + \Delta_\Gamma G p_2.\end{aligned}$$

From (2.24) and $e_1 = \epsilon e_2$, we have

$$\begin{aligned} f(\varphi) &= -\frac{2H}{\epsilon}q_0' + \epsilon^2\Delta e_2 - (3q_0^2-1)e_2 - 3\epsilon^2 q_0 e_2^2 + O(\epsilon^4)\\ &= -\frac{2H}{\epsilon}q_0' + Gp_2'' - 2HGp_2'\epsilon + \Delta_\Gamma G p_2\epsilon^2 + \epsilon e_3'' - 2H\epsilon^2 e_3' + \Delta_\Gamma e_3\epsilon^3\\ &\quad - (3q_0^2-1)Gp_2 - \epsilon(3q_0^2-1)e_3 - 3G^2q_0p_2^2\epsilon^2 - 6Gq_0pe_3\epsilon^3 + O(\epsilon^4)\\ &= -\frac{2H}{\epsilon}q_0' + (Gp_2'' - (3q_0^2-1)Gp_2) + \epsilon(-2HGp_2' + e_3'' - (3q_0^2-1)e_3)\\ &\quad + \epsilon^2(\Delta_\Gamma G p_2 - 2He_3' - 3G^2q_0p_2^2) + \epsilon^3(\Delta_\Gamma e_3 - 6Gq_0p_2e_3) + O(\epsilon^4)\\ &= f(q_0) + a + \epsilon b + \epsilon^2 c + \epsilon^3 d + O(\epsilon^4),\end{aligned}$$

where

$$a = Gp_2'' - (3q_0^2-1)Gp_2, \tag{2.30}$$

$$b = -2HGp_2' + e_3'' - (3q_0^2-1)e_3, \tag{2.31}$$

$$c = \Delta_\Gamma G p_2 - 2He_3' - 3G^2q_0p_2^2, \tag{2.32}$$

$$d = \Delta_\Gamma e_3 - 6Gq_0p_2e_3. \tag{2.33}$$

Then

$$\begin{aligned} &W(\varphi) - W(q_0)\\ &\quad = \frac{\epsilon}{2}\int_\Omega (f(\varphi)+f(q_0))(f(\varphi)-f(q_0))\,dx\\ &\quad = \frac{\epsilon}{2}\int_\Omega 2f(q_0)(a+\epsilon b+\epsilon^2 c+\epsilon^3 d+O(\epsilon^4)) + (a+\epsilon b+\epsilon^2 c+\epsilon^3 d+O(\epsilon^4))^2\,dx\\ &\quad = \frac{\epsilon}{2}\int_\Omega 2f(q_0)a + [2f(q_0)b\epsilon + a^2] + \epsilon[2f(q_0)c\epsilon + 2ab]\\ &\qquad + \epsilon^2[2f(q_0)d\epsilon + 2ac + b^2] + O(\epsilon^3)\,dx.\end{aligned}$$

Below we focus on the terms involving q_3. We use ellipses to denote all the terms not involving q_3. First, $\int_\Omega 2f(q_0)a\,dx$ is independent of q_3. Then from Lemma 2.4

$$\begin{aligned}\int_\Omega 2f(q_0)b\epsilon\,dx &= \int_\Omega 2f(q_0)(-2HGp_2')\epsilon + 2f(q_0)(e_3'' - (3q_0^2-1)e_3)\epsilon\,dx\\ &= \int_\Omega \cdots - 4Hq_0'(e_3'' - (3q_0^2-1)e_3)\,dx\\ &= \int_\Omega \cdots - 4[Hq_0''' - 2Kq_0''\epsilon - Hq_0'(3q_0^2-1)]e_3\,dx\\ &= \int_\Omega \cdots + 8Kq_0''e_3\epsilon\,dx.\end{aligned}$$

$\int_\Omega a^2\,dx$ is also independent of q_3.

$$\begin{aligned}\epsilon\int_\Omega 2f(q_0)c\epsilon\,dx &= \epsilon\int_\Omega -4Hq_0'(\Delta_\Gamma Gp_2 - 2He_3' - 3G^2q_0p_2^2)\,dx\\ &= \epsilon\int_\Omega \cdots + 8(-2H(H^2-K)q_0'\epsilon - H^2q_0'')e_3 + \cdots\,dx\\ &= \int_\Omega \cdots - 16H(H^2-K)q_0'e_3\epsilon^2 - 8H^2q_0''e_3\epsilon\,dx.\\ \epsilon\int_\Omega 2ab\,dx &= \epsilon\int_\Omega \cdots + 2a(e_3'' - (3q_0^2-1)e_3)\,dx\\ &= \int_\Omega \cdots + 2(a'' - 4Ha'\epsilon + 2Ka\epsilon^2)e_3\epsilon - 2(3q_0^2-1)ae_3\epsilon\,dx\\ &= \int_\Omega \cdots + 2a''e_3\epsilon - 8Ha'e_3\epsilon^2 - 2(3q_0^2-1)ae_3\epsilon + O(\epsilon^3)\,dx.\\ \epsilon^2\int_\Omega \epsilon 2f(q_0)d\,dx &= \int_\Omega -4Hq_0'\Delta_\Gamma e_3\epsilon^2 + 24HGq_0'q_0p_2e_3\epsilon^2\,dx.\\ \epsilon^2\int_\Omega 2ac\,dx &= \epsilon^2\int_\Omega 2a(\Delta_\Gamma Gp_2 - 2He_3' - 3G^2q_0p_2^2)\,dx\\ &= \epsilon^2\int_\Omega \cdots - 4[Ka\epsilon - Ha']e_3\,dx\\ &= \int_\Omega \cdots + O(\epsilon^3) + 4Ha'e_3\epsilon^2\,dx.\\ \epsilon^2\int_\Omega b^2\,dx &= \epsilon^2\int_\Omega \cdots - 4HGp_2'(e_3'' - (3q_0^2-1)e_3) + (e_3'' - (3q_0^2-1)e_3)^2\,dx\\ &= \int_\Omega \cdots - 4HG(p_2''' - (3q_0^2-1)p_2')e_3\epsilon^2 + (e_3'' - (3q_0^2-1)e_3)^2\epsilon^2\\ &\quad + O(\epsilon^3)\,dx.\end{aligned}$$

Combining all the items involving q_3 above, we need to minimize

$$\begin{aligned}E(e_3) = &\int_\Omega 8Kq_0''e_3\epsilon - 16H(H^2-K)q_0'e_3\epsilon^2 - 8H^2q_0''e_3\epsilon + 2a''e_3\epsilon - 8Ha'e_3\epsilon^2\\&- 2(3q_0^2-1)ae_3\epsilon - 4Hq_0'\Delta_\Gamma e_3\epsilon^2 + 24HGq_0'q_0p_2e_3\epsilon^2 + 4Ha'e_3\epsilon^2\\&- 4HG(p_2'''-(3q_0^2-1)p_2')e_3\epsilon^2 + (e_3''-(3q_0^2-1)e_3)^2\epsilon^2 + O(\epsilon^3)\,dx\\=&\int_\Omega [-8(H^2-K)q_0'' + 2a'' - 2(3q_0^2-1)a]e_3\epsilon\\&+ [-16H(H^2-K)q_0'e_3 - 4Ha'e_3 - 4Hq_0'\Delta_\Gamma e_3 + 24HGq_0'q_0p_2e_3\\&- 4HG(p_2'''-(3q_0^2-1)p_2')e_3]\epsilon^2 + (e_3''-(3q_0^2-1)e_3)^2\epsilon^2 + O(\epsilon^3)\,dx.\end{aligned}$$

Now we can simplify above $E(e_3)$. From (2.28), we have $a = Gs_2 = Gq_0't/2$, $a'' - (3q_0^2-1)a = G(q_0^2-1)q_0 = Gq_0''$, $s_2' = p_2''' - (3q_0^2-1)p_2' - 6q_0q_0'p_2$, and from Lemmas 2.9 and 2.6,

$$\int_\Omega -4Hq_0'\Delta_\Gamma e_3\,dx = \int_\Omega -4H_\Gamma q_0'\Delta_\Gamma e_3\,dx + O(\epsilon) = \int_\Omega -4\Delta_\Gamma Hq_0'e_3\,dx + O(\epsilon).$$

From Lemma 2.9, since $\nabla(H^2-K)\cdot\nabla d = 4H(H^2-K)$,

$$\begin{aligned}\int_\Omega -8(H^2-K)q_0''e_3\epsilon\,dx &= \int_\Omega -8[(H^2-K)|_\Gamma + 4H(H^2-K)|_\Gamma d]q_0''e_3\epsilon + O(\epsilon^3)\,dx\\&= \int_\Omega -2[G + 4H_\Gamma Gd]q_0''e_3\epsilon + O(\epsilon^3)\,dx.\end{aligned}$$

Altogether we have

$$\begin{aligned}E(e_3) = &\epsilon^2\int_\Omega \left[-16H(H^2-K)q_0' - 8HGs_2' - 4\Delta_\Gamma Hq_0' - 8H_\Gamma G\frac{d}{\epsilon}q_0''\right]e_3\\&+ (e_3''-(3q_0^2-1)e_3)^2 + O(\epsilon)\,dx\\=&\epsilon^2\int_\Omega 4H(H^2-K)(x|_\Gamma)\left[-4q_0' - 8s_2' - 8\frac{d}{\epsilon}q_0''\right]q_3 - 4\Delta_\Gamma H(x|_\Gamma)q_0'q_3\\&+ (q_3''-(3q_0^2-1)q_3)^2 + O(\epsilon)\,dx.\end{aligned}$$

Since $s_2 = q_0't/2$, denoting $Q = 8H(H^2-K)(x|_\Gamma)[2q_0'(t) + 3q_0''(t)t] + 2\Delta_\Gamma H(x|_\Gamma)q_0'(t)$, we have q_3 minimizing

$$E(e_3) = 2\epsilon^2\int_\Omega -Qq_3 + \frac{1}{2}(q_3''-(3q_0^2-1)q_3)^2\,dx.$$

From Lemma 2.7,

$$E(e_3) = 2\epsilon^3\int_\Gamma\int_{-\infty}^{\infty} -Qq_3 + \frac{1}{2}(q_3''-(3q_0^2-1)q_3)^2\,dt\,dS + o(\epsilon^3).$$

Therefore, for any given $x|_\Gamma$, $q_3(t, x|_\Gamma)$ minimizes

$$\int_{-\infty}^{\infty} -Qq_3 + \frac{1}{2}(q_3''-(3q_0^2-1)q_3)^2\,dt. \qquad \square$$

A direct corollary from Theorem 2.17 is that the equilibrium surface Γ must satisfy

$$\Delta_\Gamma H + 2H(H^2 - K) = 0.$$

Because this result is very important, we discuss it in section 3.

3. Equilibrium state and Willmore force of surface Γ. Based on Theorems 2.10, 2.12, 2.13, and 2.17, we have the following theorem about the equilibrium surface Γ in the phase field model.

THEOREM 3.1. *With assumptions* (A1)–(A4), *if φ minimizes energy $W(\varphi)$ defined in* (1.4), Γ, *the zero level set of φ asymptotically satisfies that*

$$\Delta_\Gamma H + 2H(H^2 - K) = 0. \tag{3.1}$$

Proof. Since φ minimizes energy $W(\varphi)$, we have that

$$\varphi = q_0 + \epsilon^2 q_2 + \epsilon^3 q_3 + O(\epsilon^4),$$

where q_3 minimizes

$$\int_{-\infty}^{\infty} -Qq_3 + \frac{1}{2}(q_3'' - (3q_0^2 - 1)q_3)^2 \, dt, \tag{3.2}$$

where $Q = 8H(H^2 - K)[2q_0' + 3q_0''t] + 2\Delta_\Gamma H q_0'$.

Now, denote $s_3 = q_3'' - (3q_0^2 - 1)q_3$. Taking the variation of (3.2), we have that

$$s_3'' - (3q_0^2 - 1)s_3 = Q. \tag{3.3}$$

From assumption (A4), we have that $q_3(\pm\infty) = q_3'(\pm\infty) = 0$. Then $s_3(\pm\infty) = s_3'(\pm\infty) = 0$. Now noticing that

$$q_0'' = q_0(q_0^2 - 1), \quad (q_0')'' - (3q_0^2 - 1)q_0' = 0 \,,$$

from (3.3) we have that

$$\begin{aligned} \int_{-\infty}^{\infty} Qq_0' \, dt &= \int_{-\infty}^{\infty} [s_3'' - (3q_0^2 - 1)s_3]q_0' \, dt \\ &= \int_{-\infty}^{\infty} s_3[(q_0')'' - (3q_0^2 - 1)q_0'] \, dt = 0. \end{aligned} \tag{3.4}$$

On the other hand,

$$\begin{aligned} \int_{-\infty}^{\infty} Qq_0' \, dt &= \int_{-\infty}^{\infty} [8H(H^2 - K)[2q_0' + 3q_0''t] + 2\Delta_\Gamma H q_0']q_0' \, dt \\ &= 8H(H^2 - K) \int_{-\infty}^{\infty} [2(q_0')^2 + 3q_0''q_0't] \, dt + 2\Delta_\Gamma H \int_{-\infty}^{\infty} (q_0')^2 \, dt \\ &= (\Delta_\Gamma H + 2H(H^2 - K)) \int_{-\infty}^{\infty} 2(q_0')^2 \, dt. \end{aligned}$$

Comparing this with (3.4), we have

$$\Delta_\Gamma H + 2H(H^2 - K) = 0. \qquad \square$$

Equation (3.1) is exactly the Willmore identity for Willmore's problem. Theorem 3.1 verifies that the equilibrium state of the phase field model is consistent with that of the sharp interface model. Equation (3.1) is a necessary condition for the existence of function q_3. If surface Γ does not satisfy (3.1), then to minimize the phase field energy $W(\varphi)$, we can still choose optimal q_0, q_1, and q_2 but not q_3. This is the main reason why the gradient flow in [7, 8] works, which will be explained below.

In our phase field model in finding the equilibrium shape of Γ, we first start from an initial phase field function φ^0, whose zero level set is Γ^0. Then we search a lower energy state along the gradient flow of $W(\varphi)$. Denote the gradient by

$$g(\varphi) = \frac{\partial W(\varphi)}{\partial \varphi} = \epsilon \Delta f(\varphi) - \frac{1}{\epsilon}(3\varphi^2 - 1) f(\varphi), \tag{3.5}$$

where f is defined in (2.1). The gradient flow is as flowing

$$\varphi_t = -g(\varphi) \tag{3.6}$$

with initial condition $\varphi|_{t=0} = \varphi^0$. In [7, 8] we developed several numerical schemes for the gradient flow (3.6), including a fully implicit scheme solving by Newton's method and some semi-implicit or explicit schemes. With the asymptotic analysis in section 2.3, we can now explain why the surface Γ moves along the gradient flow by using a kind of explicit scheme.

Suppose at step n that we have phase field φ^n with zero level set Γ^n. With the explicit surface Γ^n, for the next step along the gradient we get the fastest descent of the energy. With the asymptotic expansion, obviously we can select

$$\varphi^{n+1}(x) = \tanh\left(\frac{d(x, \Gamma^n)}{\sqrt{2}\epsilon}\right) + 4\epsilon^2 (H^2(x|_{\Gamma^n}) - K(x|_{\Gamma^n})) p\left(\frac{d(x, \Gamma^n)}{\epsilon}\right) + \epsilon^3 q_3^{n+1} + \cdots .$$

Now, the gradient flow of φ is actually the gradient flow of q_3 while q_0, q_1, and q_2 are fixed to be optimal. Based on Theorem 2.17, the gradient flow of q_3 is

$$q_{3,t} = -Q + s_3'' - (3q_0^2 - 1)s_3.$$

Theorem 3.1 states that if Γ^n is not an equilibrium shape along the gradient flow of q_3, we cannot get an optimal q_3. Actually along the gradient flow of q_3, we get $q_3(\Gamma^n) \neq 0$, i.e., Γ^n is not the zero level set anymore! With a small time step, along the gradient flow of q_3, Γ^n moves to be Γ^{n+1} with a decreasing energy $W(\varphi^{n+1}) < W(\varphi^n)$. This process will stop until Γ reaches an equilibrium state.

We can also describe how the zero level set Γ moves along the gradient flow. With the movement of level set Γ_d, we have $\varphi(x(t), t) = d$. Then $\varphi_t + \nabla\varphi \cdot \dot{x}(t) = 0$. Denote the induced elastic force in the phase field model by ξ_ϵ. Based on the relation between momentum and energy, we have

$$\xi_\epsilon \cdot \dot{x} = -W_t = -\frac{\partial W(\varphi)}{\partial \varphi}\varphi_t = -\frac{\partial W(\varphi)}{\partial \varphi}(-\nabla\varphi \cdot \dot{x}),$$

which results in the formula for the elastic force in the phase field model:

$$\xi_\epsilon = \frac{\partial W(\varphi)}{\partial \varphi}\nabla\varphi. \tag{3.7}$$

One can refer to [12, 10] for more discussions about it.

On the other hand, it is known [26, 17] that the Willmore flow force, the gradient of the sharp interface Willmore energy (1.1), is given by

$$(\Delta_\Gamma H + 2H(H^2 - K))\vec{n}(x), \tag{3.8}$$

where $\vec{n}(x)$ is the normal vector on Γ. A basic question of phase field models minimizing (1.4) is to verify the consistency of these two forces as $\epsilon \to 0$. In another words, we need to verify

$$\lim_{\epsilon\to 0} \xi_\epsilon \sim c(\Delta_\Gamma H + 2H(H^2 - K))\vec{n}(x), \tag{3.9}$$

where a constant c may apply. Unlike the sharp interface model, in the phase field model the elastic force exists everywhere but concentrates around the transition layer. Therefore we can only do a weak form of (3.9). Some preliminary analyses of formula (3.9) are given in [12, 10].

Now in the case where Γ is not an equilibrium shape, although there is no optimal selection of q_3, we can still have the optimal selection of q_0, q_1, and q_2. Further, we have the following theorem for the weak form of (3.9).

THEOREM 3.2. *With assumptions* (A1)–(A4), *for zero level set* Γ *if* φ *minimizes* $W(\varphi)$ *to the second order term, we have*

$$\int_\Omega -\xi_\epsilon \cdot v\, dx = \frac{4\sqrt{2}}{3}\int_\Gamma (\Delta_\Gamma H + 2H(H^2 - K))\vec{n}\cdot v\, dx + O(\epsilon)$$

for any continuous vector field v.

Remark 3.3. A major improvement over the method in [12, 10] is that we do not make any special assumptions on v, such as a divergence-free condition.

In general, the normal direction is defined as pointing towards the outside of Γ. But the normal direction $\vec{n}(x)$ defined by ∇d is towards the inside of Γ because we set the distance function d positive inside and negative outside. This difference results in the negative sign before ξ_ϵ.

Proof. In this proof, we need to expand the phase field function with the third order term, i.e.,

$$\varphi(x) = q_0\left(\frac{d(x)}{\epsilon}, x|_\Gamma\right) + \epsilon^2 q_2\left(\frac{d(x)}{\epsilon}, x|_\Gamma\right) + \epsilon^3 q_3\left(\frac{d(x)}{\epsilon}, x|_\Gamma\right) + O(\epsilon^3).$$

From (2.26) we have that

$$f(\varphi) = -\frac{2H}{\epsilon}q_0' + (q_2'' - (3q_0^2 - 1)q_2) + \epsilon[-2Hq_2' + q_3'' - (3q_0^2 - 1)q_3] + O(\epsilon^2). \tag{3.10}$$

Denote that

$$a = q_2'' - (3q_0^2 - 1)q_2; \tag{3.11}$$

$$b = q_3'' - (3q_0^2 - 1)q_3. \tag{3.12}$$

We have

$$f = -\frac{2H}{\epsilon}q_0' + a + \epsilon[-2Hq_2' + b] + O(\epsilon^2).$$

To calculate Δf, first we obtain

$$\begin{aligned}\Delta(Hq_0') &= \Delta H q_0' + 2(\nabla H\cdot\nabla d)\frac{1}{\epsilon}q_0'' + H(-2H)\frac{1}{\epsilon}q_0'' + H\frac{1}{\epsilon^2}q_0''' \\ &= \Delta H q_0' + 2(H^2-K)\frac{1}{\epsilon}q_0'' + H\frac{1}{\epsilon^2}q_0'''\end{aligned}$$

and

$$\Delta(Hq_2') = H\frac{1}{\epsilon^2}q_2''' + O\left(\frac{1}{\epsilon}\right).$$

Second, from (2.7) we have that

$$\begin{aligned}\Delta a &= \frac{1}{\epsilon^2}a'' - \frac{2H}{\epsilon}a' + O(1),\\ \Delta b &= \frac{1}{\epsilon^2}b'' + O\left(\frac{1}{\epsilon}\right).\end{aligned}$$

Altogether we have that

$$\epsilon\Delta f = -\frac{2H}{\epsilon^2}q_0''' - \frac{4}{\epsilon}(H^2-K)q_0'' - 2\Delta H q_0' + \frac{1}{\epsilon}a'' - 2Ha' - 2Hq_2''' + b'' + O(\epsilon).$$

And because

$$\begin{aligned}\frac{1}{\epsilon}(3\varphi^2-1)f &= \frac{1}{\epsilon}(3q_0^2 - 1 + 6q_0q_2\epsilon^2 + O(\epsilon^3))f\\ &= -\frac{2H}{\epsilon^2}(3q_0^2-1)q_0' + \frac{1}{\epsilon}(3q_0^2-1)a\\ &\quad + (3q_0^2-1)[-2Hq_2' + b] - 12Hq_0q_0'q_2 + O(\epsilon),\end{aligned}$$

we have

$$\begin{aligned}g(\varphi) &= \epsilon\Delta f(\varphi) - \frac{1}{\epsilon}(3\varphi^2-1)f(\varphi)\\ &= -\frac{4}{\epsilon}(H^2-K)q_0'' - 2\Delta H q_0' + \frac{1}{\epsilon}[a'' - (3q_0^2-1)a] + [b'' - (3q_0^2-1)b]\\ &\quad - 2Ha' - 2Hq_2''' + 2Hq_2'(3q_0^2-1) + 12Hq_0q_0'q_2 + O(\epsilon)\\ &= -\frac{4}{\epsilon}(H^2-K)q_0'' - 2\Delta H q_0' + \frac{1}{\epsilon}[a'' - (3q_0^2-1)a]\\ &\quad + [b'' - (3q_0^2-1)b] - 4Ha' + O(\epsilon).\end{aligned}$$

From Theorem 2.13, we have

$$a'' - (3q_0^2-1)a = 4(H^2-K)|_\Gamma(q_0^2-1)q_0.$$

Then

$$\begin{aligned}g(\varphi) &= -\frac{4}{\epsilon}(H^2-K)q_0'' - 2\Delta H q_0' + \frac{4}{\epsilon}(H^2-K)|_\Gamma q_0''\\ &\quad + [b'' - (3q_0^2-1)b] - 4Ha' + O(\epsilon),\end{aligned}$$

because

$$-\frac{4}{\epsilon}(H^2-K)q_0''+\frac{4}{\epsilon}(H^2-K)|_\Gamma q_0''=-\frac{4}{\epsilon}(\nabla(H^2-K)\cdot\nabla d)dq_0''+O(\epsilon)$$
$$=-16H(H^2-K)(q_0''t)+O(\epsilon)$$

and

$$4Ha'=16H(H^2-K)(q_0't/2)'=8H(H^2-K)(q_0''t+q_0');$$

also from Lemma 2.3, $\Delta H=\Delta_\Gamma H+4H(H^2-K)$; altogether we have that

$$\begin{aligned}g(\varphi)&=-16H(H^2-K)(q_0''t)-2\Delta_\Gamma Hq_0'-8H(H^2-K)q_0'\\&\quad-8H(H^2-K)(q_0''t+q_0')+[b''-(3q_0^2-1)b]+O(\epsilon)\\&=-Q+[b''-(3q_0^2-1)b]+O(\epsilon),\end{aligned}\tag{3.13}$$

where $Q=8H(H^2-K)[2q_0'+3q_0''t]+2\Delta_\Gamma Hq_0'$.

For any continuous vector field $v(x)$, from Lemma 2.8, $\int_{-\infty}^{\infty}[b''-(3q_0^2-1)b]q_0'\,dt=\int_{-\infty}^{\infty}[q_0'''-(3q_0^2-1)q_0']b\,dt=0$, and $\int_{-\infty}^{\infty}q_0''q_0'tdt=-\frac{1}{2}\int_{-\infty}^{\infty}(q_0')^2\,dt$, we have

$$\begin{aligned}\int_\Omega-\xi_\epsilon\cdot v\,dx&=\int_\Omega-g(\varphi)\nabla\varphi\cdot v\,dx\\&=\int_\Omega(Q-[b''-(3q_0^2-1)b]+O(\epsilon))\left(\frac{1}{\epsilon}q_0'\nabla d+O(1)\right)\cdot v\,dx\\&=\int_\Omega\frac{1}{\epsilon}(Q-[b''-(3q_0^2-1)b])q_0'\nabla d\cdot v\,dx+O(\epsilon)\\&=\int_{-\infty}^{\infty}[2q_0'+3q_0''t]q_0'\,dt\int_\Gamma 8H(H^2-K)\vec{n}\cdot v\,dx\\&\quad+\int_{-\infty}^{\infty}(q_0')^2\,dt\int_\Gamma 2\Delta_\Gamma H\vec{n}\cdot v\,dx\\&\quad-\int_{-\infty}^{\infty}[b''-(3q_0^2-1)b]q_0'\,dt\int_\Gamma\vec{n}\cdot v\,dx+O(\epsilon)\\&=\int_{-\infty}^{\infty}2(q_0')^2\,dt\int_\Gamma(\Delta_\Gamma H+2H(H^2-K))\vec{n}\cdot v\,dx+O(\epsilon)\\&=\frac{4\sqrt{2}}{3}\int_\Gamma(\Delta_\Gamma H+2H(H^2-K))\vec{n}\cdot v\,dx+O(\epsilon).\qquad\square\end{aligned}$$

If Γ is the equilibrium state, from Theorem 2.17 we have

$$b''-(3q_0^2-1)b=Q.$$

Then from (3.13) we have

$$g(\varphi)=O(\epsilon).$$

This result is trivial, because in the equilibrium state, eventually we have $g(\varphi)=0$. We can even reverse the above analysis to get another proof of Theorems 2.10, 2.12, 2.13, and 2.17.

4. Effect of spontaneous curvature. With the presence of spontaneous curvature c, we denote

$$f_c(\varphi) = \Delta\varphi - \frac{1}{\epsilon^2}(\varphi^2 - 1)(\varphi + \sqrt{2}c\epsilon), \tag{4.1}$$

and the elastic bending energy with spontaneous curvature is given by [11]

$$W_c(\varphi) = \frac{\epsilon}{2}\int_\Omega f_c(\varphi)^2\,dx. \tag{4.2}$$

The variation of $W_c(\varphi)$ is denoted by

$$g_c(\varphi) = \epsilon\Delta f_c(\varphi) - \frac{1}{\epsilon}(3\varphi^2 - 1 + 2\sqrt{2}c\epsilon\varphi)f_c(\varphi). \tag{4.3}$$

For simplicity, here we consider only the case with constant spontaneous curvature.

First, we have the following theorem about the asymptotic expansion of the phase field function minimizing energy $W_c(\varphi)$.

THEOREM 4.1. *With assumptions* (A1)–(A4), *in the spontaneous curvature case we have*

$$q_0(t) = \tanh\left(\frac{t}{\sqrt{2}}\right), \tag{4.4}$$

$$q_1 = 0, \tag{4.5}$$

$$q_2 = 4(H^2 - K + 2cH - c^2)p_2, \tag{4.6}$$

where p_2 is defined in Theorem 2.13.

Proof. Expanding

$$\varphi = q_0 + \epsilon q_1 + \epsilon^2 q_2 + \epsilon^3 q_3 + O(\epsilon^4)$$

and denoting

$$a_0 = q_0'' - (q_0^2 - 1)q_0,$$
$$a_i = q_i'' - (3q_0^2 - 1)q_i, \quad i = 1, 2, 3,$$

we have

$$\begin{aligned}\Delta\varphi = &\frac{1}{\epsilon^2}(q_0'' + \epsilon q_1'' + \epsilon^2 q_2'' + \epsilon^3 q_3'')\\ &-\frac{2H}{\epsilon}(q_0' + \epsilon q_1' + \epsilon^2 q_2') + \Delta_\Gamma(q_0' + \epsilon q_1') + O(\epsilon^2)\end{aligned}$$

and

$$\begin{aligned}\frac{1}{\epsilon^2}(\varphi^2 - 1)\varphi &= \frac{1}{\epsilon^2}[q_0^2 - 1 + \epsilon(2q_0q_1) + \epsilon^2(q_1^2 + 2q_0q_2) + \epsilon^3(2q_0q_3 + 2q_1q_2)\\ &\quad + O(\epsilon^4)](q_0 + \epsilon q_1 + \epsilon^2 q_2 + \epsilon^3 q_3 + O(\epsilon^4))\\ &= \frac{1}{\epsilon^2}[(q_0^2 - 1)q_0 + \epsilon(3q_0^2 - 1)q_1 + \epsilon^2((3q_0^2 - 1)q_2 + 3q_0q_1^2)\\ &\quad + \epsilon^3((3q_0^2 - 1)q_3 + 6q_0q_1q_2 + q_1^3) + O(\epsilon^4)].\end{aligned}$$

Therefore

$$f(\varphi) = \frac{1}{\epsilon^2}a_0 + \frac{1}{\epsilon}[a_1 - 2Hq_0'] + [a_2 - 2Hq_1' + \Delta_\Gamma q_0' - 3q_0q_1^2]$$
$$+ \epsilon[a_3 - 2Hq_2' + \epsilon\Delta_\Gamma q_1' - 6q_0q_1q_2 - q_1^3] + O(\epsilon^2).$$

Since

$$\frac{\sqrt{2}c}{\epsilon}(\varphi^2 - 1) = \frac{\sqrt{2}c}{\epsilon}(q_0^2 - 1) + 2\sqrt{2}cq_0q_1 + \sqrt{2}c\epsilon(q_1^2 + 2q_0q_2) + O(\epsilon^2),$$

we have

$$\begin{aligned}
f_c(\varphi) &= f(\varphi) - \frac{\sqrt{2}c}{\epsilon}(\varphi^2 - 1)\\
&= \frac{1}{\epsilon^2}a_0 + \frac{1}{\epsilon}[a_1 - 2Hq_0' - \sqrt{2}c(q_0^2 - 1)]\\
&\quad + [a_2 - 2Hq_1' + \Delta_\Gamma q_0' - 3q_0q_1^2 - 2\sqrt{2}cq_0q_1]\\
&\quad + \epsilon[a_3 - 2Hq_2' + \epsilon\Delta_\Gamma q_1' - 6q_0q_1q_2 - q_1^3 - \sqrt{2}c(q_1^2 + 2q_0q_2)] + O(\epsilon^2).
\end{aligned}$$

(1) To minimize $W_c(\varphi) = \frac{\epsilon}{2}\int_\Omega f_c(\varphi)^2\,dx$, we need to minimize $\int_\Omega a_0^2\,dx$. Then

$$a_0 = 0 \quad \Rightarrow \quad q_0(t) = \tanh\left(\frac{t}{\sqrt{2}}\right).$$

(2) Now we need to minimize

$$\int_\Omega (a_1 - 2Hq_0' - \sqrt{2}c(q_0^2 - 1))^2\,dx = \int_\Omega (a_1 - 2(H - c)q_0')^2\,dx.$$

Since

$$\begin{aligned}
\int_\Omega (H - c)q_0'a_1\,dx &= \int_\Omega (H - c)q_0'(q_1'' - (3q_0^2 - 1)q_1)\,dx\\
&= \int_\Omega (H - c)(q_0''' - (3q_0^2 - 1)q_0')q_1\,dx + O(\epsilon) = O(\epsilon),
\end{aligned}$$

we have

$$a_1 = 0 \quad \Rightarrow \quad q_1 = 0.$$

(3) Now we have that

$$f_c(\varphi) = -\frac{2}{\epsilon}(H - c)q_0' + a_2 + \epsilon(a_3 - 2Hq_2' - 2\sqrt{2}cq_0q_2) + O(\epsilon^2).$$

Then we need to minimize

$$\int_\Omega -\frac{4}{\epsilon}(H - c)q_0'a_2 + a_2^2 - 4(H - c)q_0'(a_3 - 2Hq_2' - 2\sqrt{2}cq_0q_2)\,dx.$$

Since

$$
\begin{aligned}
&-\frac{4}{\epsilon}\int_\Omega (H-c)q_0' a_2\,dx\\
&\quad = -\frac{4}{\epsilon}\int_\Omega (H-c)q_0'(q_2''-(3q_0^2-1)q_2)\,dx - \frac{4}{\epsilon}\int_\Omega (H-c)(q_0'''-(3q_0^2-1)q_0')q_2\\
&\qquad -2Kq_0''\epsilon q_2 + 4Hcq_0''\epsilon q_2 + O(\epsilon^2)\,dx\,dx = \int_\Omega 8(K-2cH)q_0''q_2 + O(\epsilon)\,dx
\end{aligned}
$$

and similarly $\int_\Omega -4(H-c)q_0' a_3\,dx = \int_\Omega O(\epsilon)\,dx$, we need to minimize

$$
\int_\Omega 8(K-2cH)q_0''q_2 - 8H(H-c)q_0''q_2 - 8c(H-c)q_0''q_2 + a_2^2\,dx,
$$

i.e.,

$$
\int_\Omega -8(H^2-K+2cH-c^2)q_0''q_2 + a_2^2\,dx.
$$

Then we have

$$
q_2 = 4(H^2-K+2cH-c^2)p_2,
$$

where p_2 is defined in Theorem 2.13. ∎

We also have the following theorem about the Willmore flow force and the equilibrium state of the shape with spontaneous curvature.

THEOREM 4.2. *With assumptions* (A1)–(A4), *if* φ *minimizes the energy* $W_c(\varphi)$, *the equilibrium state of* Γ *in the phase field model asymptotically satisfies*

$$
\Delta_\Gamma H + 2H(H^2-K) + 2Kc - 2Hc^2 = 0. \tag{4.7}
$$

And if φ *minimizes the* $W_c(\varphi)$ *to the second order term with the zero level set* Γ, *the elastic force satisfies*

$$
\int_\Omega -\xi_\epsilon \cdot v\,dx = \frac{4\sqrt{2}}{3}\int_\Gamma (\Delta_\Gamma H + 2H(H^2-K) + 2Kc - 2Hc^2)\vec{n}\cdot v\,dx + O(\epsilon)
$$

for any continuous vector field v.

Proof. We can continue with the proof of Theorem 4.1 to calculate $g_c(\varphi)$. From

$$
f_c(\varphi) = -\frac{2}{\epsilon}(H-c)q_0' + a_2 + \epsilon(a_3 - 2Hq_2' - 2\sqrt{2}cq_0q_2) + O(\epsilon^2),
$$

we have

$$
\begin{aligned}
\epsilon\Delta f_c = &-\frac{2H}{\epsilon^2}q_0''' - \frac{4}{\epsilon}(H^2-K)q_0'' - 2\Delta H q_0' + \frac{1}{\epsilon}a_2'' - 2Ha_2' - 2Hq_2''' + a_3''\\
&+ 2c\left(\frac{1}{\epsilon^2}q_0''' - \frac{2H}{\epsilon}q_0''\right) - 2\sqrt{2}c(q_0q_2)'' + O(\epsilon)
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{1}{\epsilon}(3\varphi^2-1+2\sqrt{2}c\epsilon\varphi)f_c\\
&\quad=\frac{1}{\epsilon}(3q_0^2-1+6q_0q_2\epsilon^2+2\sqrt{2}c\epsilon q_0+O(\epsilon^3))f_c\\
&\quad=-\frac{2(H-c)}{\epsilon^2}(3q_0^2-1)q_0'+\frac{1}{\epsilon}[(3q_0^2-1)a_2-4\sqrt{2}c(H-c)q_0q_0']\\
&\qquad+(3q_0^2-1)(a_3-2Hq_2'-2\sqrt{2}cq_0q_2)+2\sqrt{2}cq_0a_2-12(H-c)q_0q_0'q_2+O(\epsilon).
\end{aligned}
$$

Then

$$
\begin{aligned}
g_c(\varphi)&=\epsilon\Delta f_c-\frac{1}{\epsilon}(3\varphi^2-1+2\sqrt{2}c\epsilon\varphi)f_c\\
&=-\frac{4}{\epsilon}(H^2-K)q_0''-2\Delta Hq_0'+\frac{1}{\epsilon}(a_2''-(3q_0^2-1)a_2)+(a_3''-(3q_0^2-1)a_3)\\
&\quad-2Ha_2'-2Hq_2'''-\frac{4}{\epsilon}cHq_0''-2\sqrt{2}c(q_0q_2)''-\frac{4}{\epsilon}c(H-c)q_0''\\
&\quad+2Hq_2'(3q_0^2-1)+2\sqrt{2}cq_0q_2(3q_0^2-1)\\
&\quad-2\sqrt{2}cq_0a_2+12(H-c)q_0q_0'q_2+O(\epsilon)\\
&=-\frac{4}{\epsilon}(H^2-K+2cH-c^2)q_0''+\frac{1}{\epsilon}(a_2''-(3q_0^2-1)a_2)-2\Delta Hq_0'\\
&\quad+(a_3''-(3q_0^2-1)a_3)-4Ha_2'-2\sqrt{2}c(q_0q_2)''+2\sqrt{2}cq_0q_2(3q_0^2-1)\\
&\quad-2\sqrt{2}cq_0(q_2''-(3q_0^2-1)q_2)-12cq_0q_0'q_2+O(\epsilon)\\
&=-Q+(a_3''-(3q_0^2-1)a_3)+O(\epsilon).
\end{aligned}
$$

Now we derive Q. Because $a_2''-(3q_0^2-1)a_2=4(H^2-K+2cH-c^2)|_\Gamma q_0''$, we have

$$
\begin{aligned}
&-\frac{4}{\epsilon}(H^2-K+2cH-c^2)q_0''+\frac{1}{\epsilon}(a_2''-(3q_0^2-1)a_2)\\
&\quad=-\frac{4}{\epsilon}(\nabla(H^2-K+2cH-c^2)\cdot\nabla d)dq_0''+O(\epsilon)\\
&\quad=[-16H(H^2-K)-8c(2H^2-K)](q_0''t)+O(\epsilon).
\end{aligned}
$$

Also, $\Delta H=\Delta_\Gamma H+4H(H^2-K)$ and

$$
\begin{aligned}
-4Ha_2'&=-16H(H^2-K+2cH-c^2)s_2'\\
&=-8H(H^2-K+2cH-c^2)(q_0''t+q_0').
\end{aligned}
$$

Altogether,

$$\begin{aligned}
Q &= [16H(H^2-K)+8c(2H^2-K)](q_0''t)+2\Delta_\Gamma H q_0'+8H(H^2-K)q_0'\\
&\quad +8H(H^2-K+2cH-c^2)(q_0''t+q_0')\\
&\quad -4\sqrt{2}c[q_0''q_2-q_0'q_2'-q_0q_2''+q_0q_2(3q_0^2-1)]\\
&= 8H(H^2-K)(3q_0''t+2q_0')+2\Delta_\Gamma H q_0'+8c(2H^2-K)(q_0''t)\\
&\quad +8Hc(2H-c)(q_0''t+q_0')-4\sqrt{2}c[q_0''q_2-q_0'q_2'-q_0q_2''+q_0q_2(3q_0^2-1)].
\end{aligned}$$

When φ minimizes $W_c(\varphi)$, $g_c(\varphi)=0$. It follows that

$$a_3''-(3q_0^2-1)a_3=Q.$$

Therefore

$$\int_{-\infty}^{\infty} Qq_0'\,dt=0,$$

since

$$\int_{-\infty}^{\infty} q_0''tq_0'\,dt=-\frac{1}{2}\int_{-\infty}^{\infty}(q_0')^2\,dt$$

and

$$\begin{aligned}
&\int_{-\infty}^{\infty}[q_0''q_2-q_0'q_2'-q_0q_2''+q_0q_2(3q_0^2-1)]q_0'\,dx\\
&\quad=\int_{-\infty}^{\infty}[q_0''q_0'+2q_0'q_0''-(q_0q_0')''+q_0q_0'(3q_0^2-1)]q_2\,dx\\
&\quad=\int_{-\infty}^{\infty}[q_0q_0'''+q_0q_0'(3q_0^2-1)]q_2\,dx=0.
\end{aligned}$$

Finally we have

$$\begin{aligned}
\int_{-\infty}^{\infty} Qq_0'\,dt &= [4H(H^2-K)+2\Delta_\Gamma H-4c(2H^2-K)+4Hc(2H-c)]\int_{-\infty}^{\infty}(q_0')^2\,dt\\
&= [\Delta_\Gamma H+2H(H^2-K)+2Kc-2Hc^2]\int_{-\infty}^{\infty}2(q_0')^2\,dt=0.
\end{aligned}$$

It follows that

$$\Delta_\Gamma H+2H(H^2-K)+2Kc-2Hc^2=0,$$

which is the equilibrium equation for Γ.

On the other hand, if we have optimal q_0, q_1, and q_2, regardless of q_3 we have a nonequilibrium Γ, since $g_c(\varphi)=-Q+(a_3''-(3q_0^2-1)a_3)+O(\epsilon)$. The elastic force satisfies

$$\begin{aligned}
\int_\Omega -\xi_\epsilon \cdot v\,dx &= \int_\Omega -g_c(\varphi)\nabla\varphi\cdot v\,dx \\
&= \int_\Omega (Q-(a_3''-(3q_0^2-1)a_3)+O(\epsilon))\left(\frac{1}{\epsilon}q_0'\nabla d+O(1)\right)\cdot v\,dx \\
&= \int_\Omega \frac{1}{\epsilon}Qq_0'\nabla d\cdot v\,dx - \int_\Omega \frac{1}{\epsilon}(a_3''-(3q_0^2-1)a_3)q_0'\nabla d\cdot v\,dx + O(\epsilon) \\
&= \int_{-\infty}^{\infty} 2(q_0')^2\,dt \int_\Gamma [\Delta_\Gamma H + 2H(H^2-K)+2Kc-2Hc^2]\vec{n}\cdot v\,dx \\
&\quad - \int_{-\infty}^{\infty} 2(q_0'''-(3q_0^2-1)q_0')a_3\,dt\int_\Gamma \vec{n}\cdot v\,dx + O(\epsilon) \\
&= \frac{4\sqrt{2}}{3}\int_\Gamma [\Delta_\Gamma H + 2H(H^2-K)+2Kc-2Hc^2]\vec{n}\cdot v\,dx + O(\epsilon)
\end{aligned}$$

for any continuous vector field v. ∎

Remark 4.3. Theorem 4.2 also proved that the phase field model is consistent with the sharp interface case, in which the elastic bending energy is defined by

$$W_c(\Gamma) = \int_\Gamma (H-c)^2\,dS = \int_\Gamma H^2\,dS - 2c\int_\Gamma H\,dS + c^2|\Gamma|.$$

One can refer to [26] for the variations for the three items:

$$\begin{aligned}
\int_\Gamma H^2\,dS &\quad\rightarrow\quad \Delta_\Gamma H + 2H(H^2-K), \\
\int_\Gamma H\,dS &\quad\rightarrow\quad -K, \\
|\Gamma| = \int_\Gamma dS &\quad\rightarrow\quad -2H.
\end{aligned}$$

Actually, this theorem gives us the confidence to state that if we use a phase field function φ with a special kind of profile q to formulate a kind of energy $W(\varphi)$, with φ preserving the profile in the process minimizing $W(\varphi)$, the results of our phase field model should be consistent with those of the sharp interface model.

5. Error estimates. From the discussion of section 3 we know that if Γ is an equilibrium shape, we can expand the phase field function φ up to the third order term due to Theorems 2.10, 2.12, 2.13, and 2.17; if Γ is not an equilibrium shape, in the process of gradient flow to minimize phase field energy $W(\varphi)$, we can still expand φ to the second order term. In this section, we use the approximation of φ in the second order term to do some error estimates. The error estimate of the elastic bending energy further verifies the consistency of the phase field model with the sharp interface model. Besides the elastic bending energy formula, we can also estimate the error of some Euler number formulae derived from the tanh profile, and these error estimates can give us some new identities.

LEMMA 5.1.

$$W(q_0(d(x)/\epsilon)) = \frac{4\sqrt{2}}{3}\int_\Gamma H^2\,dS + 2C_1\epsilon^2\int_\Gamma (4H^4 - 5H^2K + K^2)\,ds + O(\epsilon^3),$$

where $C_1 = \int_{-\infty}^{\infty}(q_0')^2t^2\,dt \approx 0.608050.$

Proof. From (2.23), we have

$$W(q_0) = \frac{\epsilon}{2}\int_\Omega f(q_0)^2\,dx = \frac{1}{2\epsilon}\int_\Omega 4H^2(q_0')^2\,dx.$$

From Lemma 2.9, we have

$$\begin{aligned} W(q_0) = {} & 2\int_\Gamma H^2\,dS\int_{-\infty}^{\infty}(q_0')^2\,dt + 2\epsilon\int_\Gamma -2H^3 + 2H(2H^2-K)\,dS\int_{-\infty}^{\infty}(q')^2t\,dt \\ & + 2\epsilon^2\int_\Gamma H^2K - 2H(2H(2H^2-K)) \\ & + (12H^4 - 10H^2K + K^2)\,dS\int_{-\infty}^{\infty}(q_0')^2t^2\,dt + O(\epsilon^3) \\ = {} & \frac{4\sqrt{2}}{3}\int_\Gamma H^2\,dS + 2\epsilon^2\int_\Gamma (4H^4 - 5H^2K + K^2)\,ds\int_{-\infty}^{\infty}(q_0')^2t^2\,dt + O(\epsilon^3). \qquad \square \end{aligned}$$

THEOREM 5.2 (error estimate of energy). *If* φ *minimizes* $W(\varphi)$, *then*

$$W(\varphi) - \frac{4\sqrt{2}}{3}\int_\Gamma H^2\,dS = 6C_1\epsilon^2\int_\Gamma H^2(H^2-K)\,dS + O(\epsilon^3), \tag{5.1}$$

where $C_1 = \int_{-\infty}^{\infty}(q_0')^2t^2\,dt \approx 0.608050.$

Proof. From Theorem 2.13, equation (2.27), we have

$$\begin{aligned} W(\varphi) - W(q_0(d/\epsilon)) = 16\epsilon^2\left(\int_\Gamma (H^2-K)^2\,dS\right)\int_{-\infty}^{\infty} & -q_0''p_2 \\ & + \frac{1}{2}(p_2'' - (3q_0^2-1)p_2)^2 + O(\epsilon)\,dt. \end{aligned}$$

Because $p_2'' - (3q_0^2-1)p_2 = s_2 = q_0't/2$ and $s_2'' - (3q_0^2-1)s_2 = q_0''$, we have that

$$\begin{aligned} & \int_{-\infty}^{\infty} -q_0''p_2 + \frac{1}{2}(p_2'' - (3q_0^2-1)p_2)^2 + O(\epsilon)\,dt \\ & = \int_{-\infty}^{\infty} -(s_2'' - (3q_0^2-1)s_2)p_2 + \frac{1}{2}s_2^2 + O(\epsilon)\,dt \\ & = \int_{-\infty}^{\infty} -(p_2'' - (3q_0^2-1)p_2)s_2 + \frac{1}{2}s_2^2 + O(\epsilon)\,dt \\ & = \int_{-\infty}^{\infty} -\frac{1}{2}s_2^2 + O(\epsilon)\,dt = -\frac{1}{8}C_1 + O(\epsilon), \end{aligned}$$

where $C_1 = \int_{-\infty}^{\infty} (q_0')^2 t^2 \, dt$. Then

$$W(\varphi) - W(q_0) = -2C_1\epsilon^2 \int_\Gamma (H^2 - K)^2 \, dS + O(\epsilon^3).$$

From Lemma 5.1, we have that

$$W(\varphi) - \frac{4\sqrt{2}}{3} \int_\Gamma H^2 \, dS = 6C_1\epsilon^2 \int_\Gamma H^2(H^2 - K) \, dS + O(\epsilon^3). \tag{5.2}$$ ∎

Now, denote

$$\chi(\varphi) = -\frac{3\sqrt{2}\epsilon}{16\pi} \int_\Omega \left(\Delta\varphi - \frac{1}{\epsilon^2}\varphi(\varphi^2 - 1)\right) \frac{1}{\epsilon^2}(\varphi^2 - 1)\varphi \, dx. \tag{5.3}$$

The following lemma states that $\chi(q_0)$ is an approximation of the Euler number of Γ.

LEMMA 5.3.

$$\chi(q_0) = \chi(\Gamma) + O(\epsilon).$$

Proof. Lemma 2.3 gives $K = 2H^2 - \nabla H \nabla d$, with Lemma 2.9, and the Euler number

$$\begin{aligned}
\chi(\Gamma) &= \frac{1}{4\pi} \int_\Gamma K \, dS = \frac{3}{16\sqrt{2}\epsilon\pi} \int_\Omega K(q_0^2 - 1)^2 \, dx + O(\epsilon) \\
&= \frac{3}{16\sqrt{2}\epsilon\pi} \int_\Omega 2H^2(q_0^2 - 1)^2 - \nabla H \nabla d (q_0^2 - 1)^2 \, dx + O(\epsilon) \\
&= \frac{3}{16\sqrt{2}\epsilon\pi} \int_\Omega 2H^2(q_0^2 - 1)^2 + H\Delta d(q_0^2 - 1)^2 + 4H(q_0^2 - 1)q_0 \nabla d \nabla q_0 \, dx + O(\epsilon) \\
&= \frac{3}{4\sqrt{2}\epsilon\pi} \int_\Omega H(q_0^2 - 1)q_0 \nabla d \nabla q_0 \, dx + O(\epsilon) \\
&= -\frac{3}{8\epsilon^2\pi} \int_\Omega H(q_0^2 - 1)^2 q_0 \, dx + O(\epsilon) \\
&= -\frac{3\sqrt{2}\epsilon}{16\pi} \int_\Omega \left(\Delta q_0 - \frac{1}{\epsilon^2} q_0(q_0^2 - 1)\right) \frac{1}{\epsilon^2}(q_0^2 - 1)q_0 \, dx + O(\epsilon).
\end{aligned}$$ ∎

We have the following error estimate. One can refer to [13] for the first proof of Lemma 5.3 and more discussions of formula (5.3).

THEOREM 5.4 (error estimate of the Euler number). *If φ minimizes $W(\varphi)$, we have*

$$\chi(\varphi) = \frac{3}{16\pi\sqrt{2}} W(\varphi) + O(\epsilon).$$

Proof.

$$\begin{aligned}
\chi(\varphi)-\chi(q_0) &= -\frac{3\sqrt{2}\epsilon}{16\pi}\int_\Omega (f(\varphi)-f(q_0))\frac{1}{\epsilon^2}(\varphi^2-1)\varphi \\
&\quad + f(q_0)\frac{1}{\epsilon^2}((\varphi^2-1)\varphi-(q_0^2-1)q_0)\,dx \\
&= -\frac{3\sqrt{2}\epsilon}{16\pi}\int_\Omega (\epsilon^2\Delta e_2-(3q_0^2-1)e_2+O(\epsilon^2))\frac{1}{\epsilon^2}(\varphi^2-1)\varphi \\
&\quad - \frac{2H}{\epsilon}q_0'((3q_0^2-1)e_2+O(\epsilon^2))\,dx \\
&= -\frac{3\sqrt{2}\epsilon}{16\pi}\int_\Omega (\epsilon^2\Delta e_2-(3q_0^2-1)e_2+O(\epsilon^2)) \\
&\quad \times\left(\frac{1}{\epsilon^2}(q_0^2-1)q_0+O(1)\right)+O\left(\frac{1}{\epsilon}\right)\,dx \\
&= -\frac{3\sqrt{2}\epsilon}{16\pi}\int_\Omega (\epsilon^2\Delta e_2-(3q_0^2-1)e_2)\frac{1}{\epsilon^2}(q_0^2-1)q_0+O\left(\frac{1}{\epsilon}\right)\,dx \\
&= -\frac{3\sqrt{2}\epsilon}{16\pi}\int_\Omega (-2H\epsilon e_2'+e_2''-(3q_0^2-1)e_2)\frac{1}{\epsilon^2}(q_0^2-1)q_0+O\left(\frac{1}{\epsilon}\right)\,dx \\
&= -\frac{3\sqrt{2}\epsilon}{16\pi}\int_\Omega (e_2''-(3q_0^2-1)e_2)\frac{1}{\epsilon^2}(q_0^2-1)q_0+O\left(\frac{1}{\epsilon}\right)\,dx.
\end{aligned}$$

From Theorem 2.13, we have

$$e_2''-(3q_0^2-1)e_2 = q_2''-(3q_0^2-1)q_2+O(\epsilon) = 2(H^2-K)q_0't+O(\epsilon).$$

Therefore

$$\begin{aligned}
\chi(\varphi)-\chi(q_0) &= -\frac{3\sqrt{2}\epsilon}{16\pi}\int_\Omega \frac{2}{\epsilon^2}(H^2-K)|_\Gamma q_0(q_0^2-1)q_0't+O\left(\frac{1}{\epsilon}\right)\,dx \\
&= -\frac{6\sqrt{2}}{16\pi}\left(\int_\Gamma H^2-K\,dS\right)\left(\int_{-\infty}^{\infty} q_0''q_0't\,dt\right)+O(\epsilon).
\end{aligned}$$

Because

$$\int_{-\infty}^{\infty} q_0''q_0't\,dt = -\frac{1}{2}\int_{-\infty}^{\infty}(q_0')^2\,dt = -\frac{\sqrt{2}}{3}\,,$$

we have

$$\chi(\varphi)-\chi(q_0) = \frac{1}{4\pi}\int_\Gamma H^2-K\,dS+O(\epsilon). \tag{5.4}$$

Since $\chi(q_0)=\frac{1}{4\pi}\int_\Gamma K\,dS+O(\epsilon)$, we get

$$\chi(\varphi) = \frac{1}{4\pi}\int_\Gamma H^2\,dS+O(\epsilon).$$

From Theorem 5.2, we have

$$W(\varphi) = \frac{4\sqrt{2}}{3} \int_\Gamma H^2 \, dS + O(\epsilon^2).$$

Then we have

$$\chi(\varphi) = \frac{3}{16\pi\sqrt{2}} W(\varphi) + O(\epsilon). \qquad \square$$

Remark 5.5. Theorem 5.4 states that unlike $\chi(q_0)$, we cannot use $\chi(\varphi)$ to calculate the Euler number of Γ. It is unfortunate that the Euler number $\chi(q_0)$ in (5.4) is canceled out. Otherwise we could get a modified Euler number formula.

For any $n \geq 1$, denote

$$\chi_n(\varphi) = -\frac{n\sqrt{2}\epsilon}{8\pi\beta} \int \left(\Delta\varphi - \frac{1}{\epsilon^2}\varphi(\varphi^2 - 1) \right) \frac{1}{\epsilon^2} (\varphi^2 - 1)^{n-1} \varphi \, dx, \tag{5.5}$$

where $\beta = \int_{-1}^{1} (1 - x^2)^{n-1} \, dx$. We can prove that $\chi_n(q_0)$ is also an approximation of the Euler number $\chi(\Gamma)$. Actually formula (5.3) is a special case of χ_n with $n = 2$. And similar to the proof of Theorem 5.4, we have

$$\chi_n(\varphi) = \frac{3}{16\pi\sqrt{2}} W(\varphi) + O(\epsilon)$$

for every $n \geq 1$.

6. Numerical experiments. In this section, numerical experiments are performed to verify the results of this paper. We do not verify the expansion of φ directly. Instead, we try to verify Theorem 5.4. Our experiments start from some initial shapes represented by φ^0. Following the gradient flow of $W(\varphi)$, finally we reach some equilibrium shapes of Willmore's problem. In this process, we verify the relation of $\chi(\varphi)$ to $W(\varphi)$ as stated by Theorem 5.4.

Our experiments here are performed in the three-dimensional axis-symmetrical case. We do not add any volume and surface area constraints so that we can focus on the Willmore flow of the surface. Here we use the Fourier spectral method in order to get high accuracy results. One may refer to [7, 8, 11] for some detailed descriptions of the numerical schemes of the gradient flow (3.6), and [5] gives the convergence analysis of those schemes.

The numerical simulations presented in this paper are concerned with the z-axial symmetrical cases. The computational domain is taken to be $[-\pi, \pi]^2$ of the $x - z$ plane, which is divided into a 128×128 mesh. Our first experiment starts from a torus with the initial phase field function selected to be q_0:

$$\begin{aligned} \varphi^0(x) &= q_0(d(x)/\epsilon) \\ &= \tanh\left(\left(0.25\pi - \sqrt{(x_1 - 0.5\pi\cos\theta)^2 + (x_2 - 0.5\pi\sin\theta)^2 + x_3^2} \right) / (\sqrt{2}\epsilon) \right), \end{aligned}$$

where $\theta = \cos^{-1}(x_1/\|x\|)$. The parameter $\epsilon = 3h = 0.1473$. A cross section of the initial shape is given in the left panel of Figure 6.1. It is not an equilibrium shape and it gradually changes to an equilibrium shape (right panel of Figure 6.1) along the gradient flow.

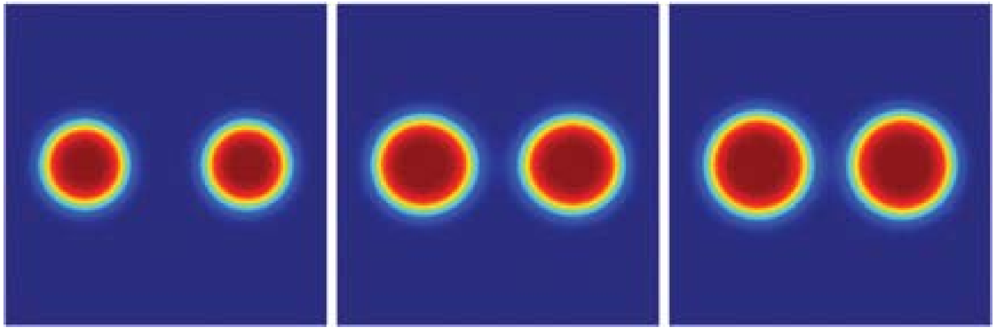


FIG. 6.1. *Cross-section views of the deformation of a torus into an equilibrium shape (time* $t = 0.00, 1.07, 4.45$*).*

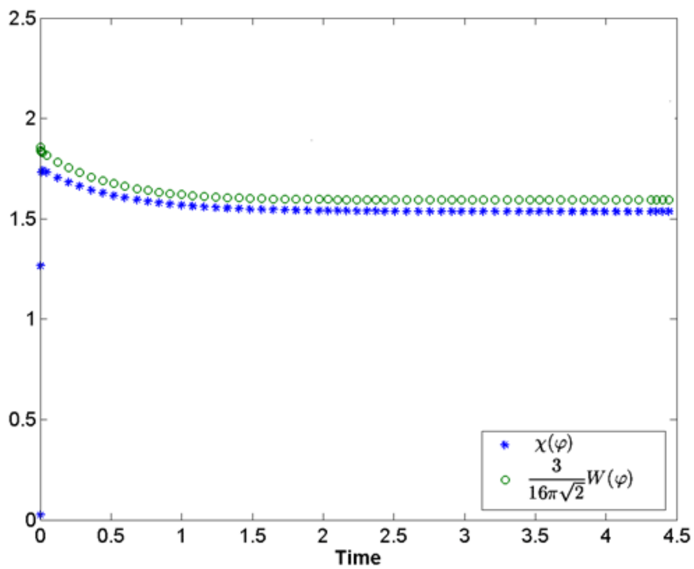


FIG. 6.2. *The plot of* $\chi(\varphi)$ *and* $3W(\varphi)/(16\pi\sqrt{2})$.

Figure 6.2 gives the plot of $\chi(\varphi)$ and $3W(\varphi)/(16\pi\sqrt{2})$. In the beginning, because $\varphi = q_0$, $\chi(\varphi) = 0.02$, which is close to the Euler number of a torus 0. Along the gradient flow, the value of $\chi(\varphi)$ approaches $3W(\varphi)/(16\pi\sqrt{2})$ very quickly. And during most of the gradient flow, they keep very close to each other. The difference is about 0.0576, which is relatively small compared to their value ($< 3.75\%$). And this difference decreases as we decrease the value of ϵ. This experiment is redone for $\epsilon = 2h$, and the difference changes to 0.0253, which is about 1.63% of their value.

Our second experiment is performed to simulate the merging of two spheres on a 128 grid mesh size with $\epsilon = 2h = 0.0982$. The cross-section views are given in Figure 6.3. Figure 6.4 gives the plot of $\chi(\varphi)$ and $3W(\varphi)/(16\pi\sqrt{2})$. Most of the time ($t > 1.0$), the value of φ coincides perfectly with the value of $3W(\varphi)/(16\pi\sqrt{2})$. There is a jump of φ before the merging of the two spheres. To analyze this jump, we show in Figure 6.5 the plot of the cross-section views around their merging time. From this graph, we can clearly see a singular point when their surfaces come to a

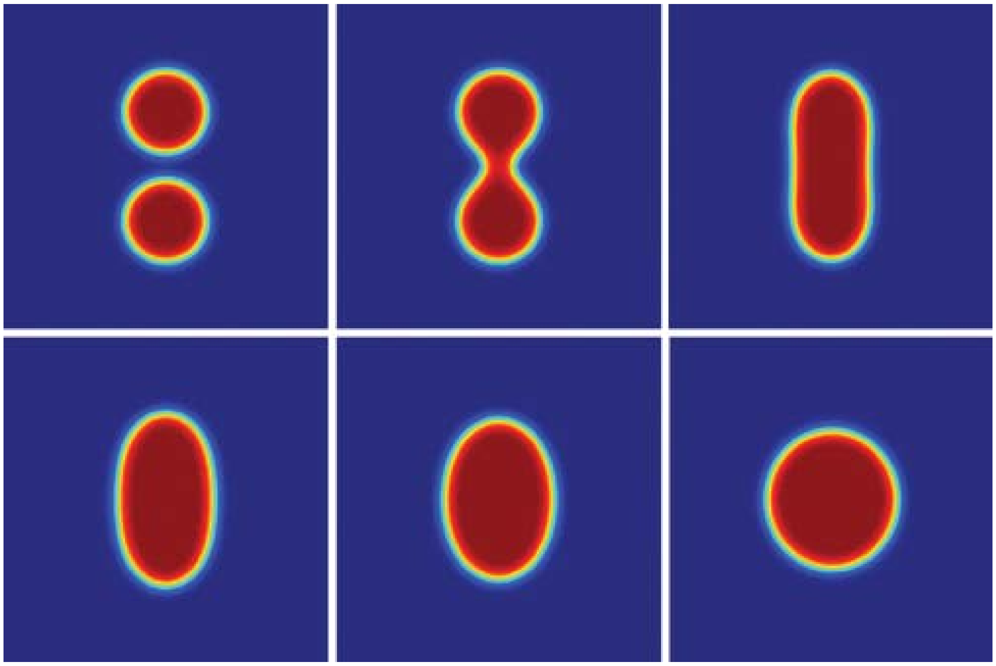


FIG. 6.3. *Cross-section views of the merging of two spheres (time* $t = 0.00, 1.08, 2.06, 3.04, 4.03,$ 7.89*).*

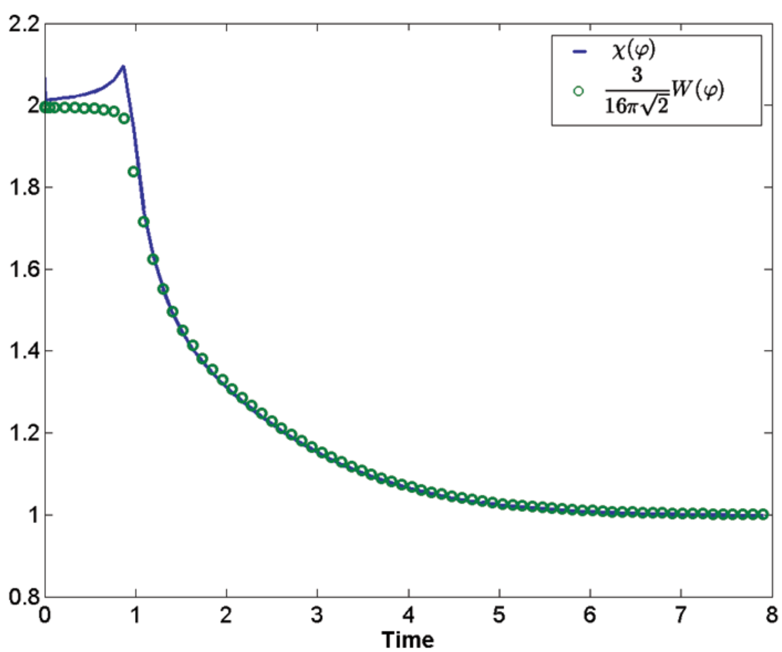


FIG. 6.4. *The plot of* $\chi(\varphi)$ *and* $3W(\varphi)/(16\pi\sqrt{2})$.

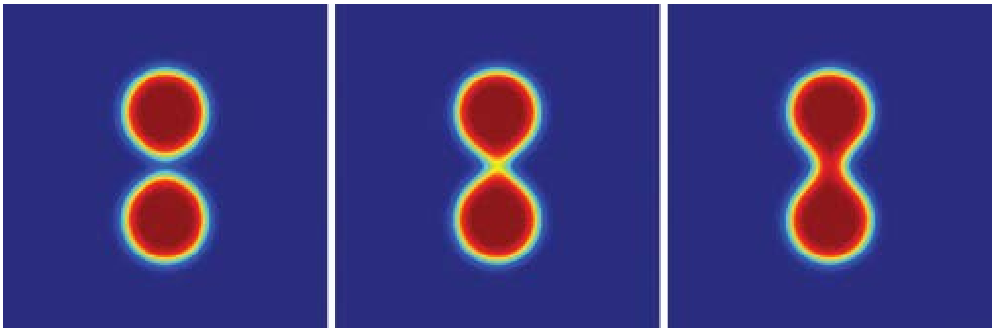


FIG. 6.5. *Cross-section views of the merging of two spheres (time* $t = 0.87, 0.97, 1.08$*).*

self-intersection. This singular point results in the nonsmoothness of the distance function and further results in the jump before their merging together. One may refer to [9] for more detailed analysis of the effect of singular points in retrieving the topological information within the phase field framework.

These experiments give numerical verification of Theorem 5.4 that $\varphi = 3W(\varphi)/(16\pi\sqrt{2}) + O(\epsilon)$, which is derived by the asymptotic analysis of section 2.3 under a general ansatz.

7. Conclusion. In this paper, we derived an explicit asymptotic expansion of the phase field functions minimizing the Willmore energy and elastic bending energy based on some relaxed assumptions. Those assumptions and approximations are verified by our numerical experiments. The asymptotic analysis results of the phase field functions are used to proved the consistency of the phase field model with the sharp interface model. They are further used to derive some error estimates of the energy and some formulae for the Euler number. The detailed analysis in this paper presents a clear structure of the phase field function and provides us a better understanding of why the phase phase models for Willmore's problem or the equilibrium lipid vesicle membrane problem have been so successful in modeling those problems.

Some future work can be pursued. First, we may study how some constraints, such as the volume/surface area constraints, change the profile of the phase field functions. Second, we can conduct some three-dimensional experiments to verify the expansion of the phase field functions. Finally, we can do more analysis on other formulae for the elastic bending energy and derive highly accurate formulae to detect the Euler number and other geometric and topological information of the surface within the phase field framework.

Acknowledgments. The author wishes to thank Prof. Qiang Du and Prof. Chun Liu for insightful discussions and collaborations on this subject.

REFERENCES

[1] T. Biben, K. Kassner, and C. Misbah, *Phase-field approach to 3D vesicle dynamics*, Phys. Rev. E., 72 (2005), article 041921.
[2] M. Bloor and M. Wilson, *Method for efficient shape parameterization of fluid membranes and vesicles*, Phys. Rev. E, 61 (2000), pp. 4218–4229.
[3] Y. C. Chang, T. Y. Hou, B. Merriman, and S. Osher, *A level set formulation of Eulerian interface capturing methods for incompressible fluid flows*, J. Comput. Phys., 124 (1996), pp. 449–464.

[4] H. Döbereiner, E. Evans, M. Kraus, U. Seifert, and M. Wortis, *Mapping vesicle shapes into the phase diagram: A comparison of experiment and theory*, Phys. Rev. E, 55 (1997), pp. 4458–4474.

[5] Q. Du and X. Wang, *Convergence of numerical approximations to a phase field bending elasticity model of membrane deformations*, Internat. J. Numer. Anal. Modeling, 4 (2007), pp. 441–459.

[6] Q. Du and J. Zhang, *Adaptive finite element method for a phase field bending elasticity model of vesicle membrane deformations*, SIAM J. Sci. Comput., to appear.

[7] Q. Du, C. Liu, and X. Wang, *A phase field approach in the numerical study of the elastic bending energy for vesicle membranes*, J. Comput. Phys., 198 (2004), pp. 450–468.

[8] Q. Du, C. Liu, and X. Wang, *Simulating the deformation of vesicle membranes under elastic bending energy in three dimensions*, J. Comput. Phys., 212 (2005), pp. 757–777.

[9] Q. Du, C. Liu, and X. Wang, *Retrieving topological information for phase field models*, SIAM J. Appl. Math., 65 (2005), pp. 1913–1932.

[10] Q. Du, C. Liu, R. Ryham, and X. Wang, *A phase field formulation of the Willmore problem*, Nonlinearity, 18 (2005), pp. 1249–1267.

[11] Q. Du, C. Liu, R. Ryham, and X. Wang, *Phase field modeling of the spontaneous curvature effect in cell membranes*, Commun. Pure Appl. Anal., 4 (2005), pp. 537–548.

[12] Q. Du, C. Liu, R. Ryham, and X. Wang, *Modeling vesicle deformations in flow fields via energetic variational approaches*, Nonlinearity, submitted.

[13] Q. Du, C. Liu, R. Ryham, and X. Wang, *Diffuse interface energies capturing the Euler number: Relaxation and renomalization*, Commun. Math. Sci., 5 (2007), pp. 233–242.

[14] J. Glimm, D. Marchesin, and O. McBryan, *A numerical method for two phase flow with unstable interface*, J. Comput. Phys., 39 (1981), pp. 179–200.

[15] W. Helfrich, *Elastic properties of lipid bilayers: Theory and possible experiments*, Z. Naturforsch. C, 28 (1973), pp. 693–703.

[16] Y. Jie, Q. Liu, J. Liu, and Z. Ou-Yang, *Numerical observation of nonaxisymmetric vesicles in fluid membranes*, Phys. Rev. E, 58 (1998), pp. 4730–4736.

[17] E. Kuwert and R. Schätzle, *The Willmore flow with small initial energy*, J. Differential Geom., 57 (2001), pp. 409–441.

[18] J. Li and Y. Renardy, *Numerical study of flows of two immiscible liquids at low Reynolds number*, SIAM Rev., 42 (2000), pp. 417–439.

[19] Z. Ou-Yang, J. Liu, and Y. Xie, *Geometric Methods in the Elastic Theory of Membranes in Liquid Crystal Phases*, World Scientific, Singapore, 1999.

[20] J. Quinn and R. Cherry, *Structural and Dynamic Properties of Lipids and Membranes*, Portland Press, London, 1992.

[21] U. Seifert, *Configurations of fluid membranes and vesicles*, Adv. in Phys., 46 (1997), pp. 13–137.

[22] Z. Tu and Z. Ou-Yang, *A geometric theory on the elasticity of bio-membranes*, J. Phys. A, 37 (2004), pp. 11407–11429.

[23] S. O. Unverdi and G. Tryggvason, *A front-tracking method for viscous, incompressible multi-fluid flows*, J. Comput. Phys., 100 (1992), pp. 25–37.

[24] X. Wang, *Phase Field Models and Simulations of Vesicle Bio-membranes*, Ph.D. thesis, Department of Mathematics, Penn State University, 2005; available online from http://etda.libraries.psu.edu/theses/approved/WorldWideFiles/ETD-961/bubble.pdf.

[25] X. Wang and Q. Du, *Diffusive interface models for multi-component lipid membranes and open membranes*, J. Math. Biol., to appear.

[26] T. J. Willmore, *Riemannian Geometry*, Oxford Science Publications, Clarendon Press, Oxford University Press, New York, 1993.

[27] Y. Yin, J. Yin, and D. Ni, *General mathematical frame for open or closed biomembranes* I: *Equilibrium theory and geometrically constraint equation*, J. Math. Biol., 51 (2005), pp. 403–413.

SIAM J. MATH. ANAL.
Vol. 39, No. 5, pp. 1402–1427

SPHERICALLY SYMMETRIC ISENTROPIC COMPRESSIBLE FLOWS WITH DENSITY-DEPENDENT VISCOSITY COEFFICIENTS*

ZHENHUA GUO†, QUANSEN JIU‡, AND ZHOUPING XIN§

Abstract. We prove the existence of global weak solutions to the compressible Navier–Stokes equations with density-dependent viscosity coefficients when the initial data are large and spherically symmetric by constructing suitable aproximate solutions. We focus on the case where those coefficients vanish on vacuum. The solutions are obtained as limits of solutions in annular regions between two balls, and the equations hold in the sense of distribution in the entire space-time domain. In particular, we prove the existence of spherically symmetric solutions to the Saint–Venant model for shallow water.

Key words. compressible Navier–Stokes, density-dependent viscosity, weak solutions

AMS subject classifications. 35Q30, 76N10

DOI. 10.1137/070680333

1. Introduction.

The compressible Navier–Stokes equations with density-dependent viscosity coefficients can be written as

$$\rho_t + \operatorname{div}(\rho \mathbf{U}) = 0, \tag{1.1}$$

$$(\rho \mathbf{U})_t + \operatorname{div}(\rho \mathbf{U} \otimes \mathbf{U}) - \operatorname{div}(h(\rho) D(\mathbf{U})) - \nabla(g(\rho)\operatorname{div}\mathbf{U}) + \nabla P(\rho) = 0, \tag{1.2}$$

where $t \in (0, +\infty)$ and $\mathbf{x} \in \mathbb{R}^N, N = 2, 3$, $\rho(\mathbf{x}, t), \mathbf{U}(\mathbf{x}, t)$, and $P(\rho) = \rho^\gamma (\gamma > 1)$ stand for the fluid density, velocity, and pressure, respectively,

$$D(\mathbf{U}) = \frac{\nabla \mathbf{U} +^t \nabla \mathbf{U}}{2}$$

is the strain tensor, and $h(\rho)$ and $g(\rho)$ are the Lamé viscosity coefficients satisfying

$$h(\rho) > 0, \quad h(\rho) + N g(\rho) \geq 0. \tag{1.3}$$

In the past several decades, significant progress on the system (1.1)–(1.2) with positive constant viscosity coefficients has been achieved by many authors. Concerning the global existence and the large-time behavior of solutions for sufficiently small data, the system (1.1)–(1.2) (as well as the full compressible Navier–Stokes equations) is well-understood in the sense that if the data are small perturbations of an uniform nonvacuum state, then there exists a (smooth or weak) solution which is time-asymptotically stable (see [21]). The situation, however, becomes more complex

*Received by the editors January 17, 2007; accepted for publication (in revised form) September 6, 2007; published electronically January 9, 2008.
http://www.siam.org/journals/sima/39-5/68033.html

†Center for Nonlinear Studies, Department of Mathematics, Northwest University, Xi'an, 710069, People's Republic of China (zhenhua.guo.math@gmail.com). Supported in part by the NSFC (grant 10401012) and grants from RGC of HKSAR CUHK4028/04P, CUHK4040/06P.

‡School of Mathematical Sciences, Capital Normal University, Beijing 100037, People's Republic of China (jiuqs@mail.cnu.edu.cn). Supported in part by the NSFC (grant 10431060) and grants from RGC of HKSAR CUHK4028/04P, CUHK4040/06P.

§The Institute of Mathematical Sciences, The Chinese University of HongKong, Shatin, N.T., HongKong (zpxin@ims.cuhk.edu.hk). Supported in part by Zheng Ge Ru Founds, grants from RGC of HKSAR CUHK4028/04P, CUHK4040/06P, and RGC Central Allocation grant CA05/06.SC01.

when the data are large, and a number of important questions, for example, the existence of global solutions in the case of heat-conducting gases and the uniqueness of weak solutions, still remain open. The first general result on weak solutions was obtained by Lions in [19], in which he used the method of weak convergence to obtain global weak solutions provided that the specific heat ratio γ is appropriately large, for example, $\gamma \geq 3N/(N+2)$, $N = 2, 3$. There have been many generalizations of this result: see [5, 6, 14, 7, 10, 15, 23], and the references therein.

It is noted that, in dealing with large amplitude solutions, one has to face the possible appearance of a vacuum state in general. However, as observed in [9, 26, 20], the compressible Navier–Stokes equations with constant viscosity coefficients behave singularly in the presence of vacuum. By some physical considerations, Liu, Xin, and Yang in [20] introduced the modified compressible Navier–Stokes equations with density-dependent viscosity coefficients for isentropic fluids. As presented in [20], in deriving the compressible Navier–Stokes equations from the Boltzmann equations by the Chapman–Enskog expansions, the viscosity depends on the temperature and correspondingly depends on the density for isentropic cases. Meanwhile, in geophysical flows, many mathematical models correspond to (1.1)–(1.2) (see [1, 3, 19]). In particular, the viscous Saint–Venant system for shallow water is expressed exactly as (1.1)–(1.2) with $N = 2, h(\rho) = \rho, g(\rho) = 0$, and $\gamma = 2$. Shallow water equations are to describe vertically averaged flows in three-dimensional shallow domains in term of the mean velocity $\mathbf{U}$ and the variation of the depth ρ due to the free surface (see [19, 3]), which is widely used in geophysical flows. Global smooth solutions for data close to equilibrium were established in [24], and related topics have been extensively studied in [1, 3], and the references therein. Nevertheless, the global existence of weak solutions for large data to the shallow water equations or more generally to the multidimensional compressible Navier–Stokes equations (1.1)–(1.2) ($N = 2, 3$) is still open. This is mainly due to the facts, that for these models, new mathematical challenges are encountered. First, we cannot preclude spontaneous cavitation (vacuum appearing) for the solutions of (1.1) and (1.2) even if the initial data are far from the vacuum. Second, when dealing with vanishing viscosity coefficients on vacuum, the velocity cannot even be defined when the density vanishes, and hence we will have no uniform estimates for the velocity. Finally, the system (1.1)–(1.2) is highly degenerate at vacuum.

For one-dimensional compressible Navier–Stokes equations (1.1) and (1.2) with $h(\rho) = \rho^\alpha, g(\rho) = 0 (\alpha \in (0, 1))$, there is much literature on the well-posedness theory of the solutions (see [11, 13, 17, 20, 25, 27, 28], and the references therein). In particular, initial-boundary-value problems for one-dimensional (1.1)–(1.2) with $h(\rho) = \rho^\alpha (\alpha > 1/2)$ and $P = \rho^\gamma (\gamma \geq 1)$ were studied by Li, Li, and Xin recently in [18], and interesting phenomena of vacuum vanishing and blowup of solutions were found there. However, few results are available for multidimensional problems. The first multidimensional result is due to Bresch, Desjardins, and Lin [3], where they showed the L^1 stability of weak solutions for the Korteweg system (with the Korteweg stress tensor $k\rho\nabla\triangle\rho$), and their result was later improved in [1] to include the case of vanishing capillarity ($k = 0$) but with an additional quadratic friction term $r\rho|\mathbf{U}|\mathbf{U}$. An interesting new entropy estimate is established in [3] in an priori way, which provides some high regularity for the density. Recently, Mellet and Vasseur [22] proved the L^1 stability of weak solutions of the system of (1.1)–(1.2) based on the new entropy estimate, extending the corresponding L^1 stability results of [3, 1] to the case $r = k = 0$. Meanwhile, although L^1 stability is considered as one of the main steps to prove the existence of weak solutions, the global existence of weak solutions of the compressible Navier–Stokes equations with density-dependent viscosity (1.1)–(1.2) is still open in the multidimensional cases. The key issue now is how to construct approximate solutions satisfying the a priori estimates required in the L^1 stability analysis. It seems highly nontrivial to do so due to the degeneracy of the viscosities near vacuum and the additional entropy inequality to be held in the construction of approximate solutions. Very recently, Bresch and Desjardins constructed approximate solutions for the two-dimensional viscous shallow water systems with drag terms or a capillarity term and for the compressible Navier–Stokes equations with cold pressure (see [2]) and proved the global existence of weak solutions to these systems (see [2, 4]). However, this construction of the approximate solutions in [2] seems not applicable to building approximate solutions for the standard compressible Navier–Stokes equations with density-dependent viscosity coefficients.

In our paper, we will construct a class of approximate solutions and furthermore prove the global existence of weak solutions for spherically symmetric solutions of the compressible Navier–Stokes equations with the viscosity coefficients depending on the density. For simplicity of the presentation, in this paper we will give only the proof of the global existence of the three-dimensional spherically symmetric solutions of (1.1)–(1.2) with $h(\rho) = \rho, g(\rho) = 0$. Our result holds true for general $h(\rho) = \rho^\alpha, g(\rho) = (\alpha - 1)\rho^\alpha$ for some $\alpha > \frac{N-1}{N}(N = 2, 3)$. More general $h(\rho)$ and $g(\rho)$ satisfying $g(\rho) = \rho h'(\rho) - h(\rho)$ and other restrictions given in [22] can be handled in a similar way. It should be noted that the shallow water equations corresponding to the case of $N = 2, \alpha = 1, \gamma = 2$ in (1.1)–(1.2) are covered, and therefore we obtain the global spherically symmetric solutions of the shallow water equations.

It seems to be difficult to adapt the analysis in [6, 19] due to the degeneracy of the viscosities near vacuum which may appear. Thus we construct the approximate solutions by solving the approximate systems of (1.1)–(1.2) with $h^\varepsilon(\rho) = h(\rho) + \varepsilon\rho^\beta, g^\varepsilon(\rho) = g(\rho)+\varepsilon(\beta-1)\rho^\beta$ for some fixed $0 < \beta < 1$ ($\beta = 3/4$, for example) instead of $h(\rho), g(\rho)$ in (1.1)–(1.2). This is motivated by the approach of Jiang, Xin, and Zhang [13], in which the one-dimensional case is considered and $h(\rho)$ can be regarded as ρ^α, and $g(\rho) = (\alpha - 1)\rho^\alpha$ for $0 < \alpha < 1$. However, compared with the one-dimensional equations, there are some new difficulties encountered for radial symmetric three-dimensional Navier–Stokes systems. In particular, the three-dimensional spherically symmetric equations become singular at $r = 0$, and more new source terms appear in both Eulerian and Lagrangian radial symmetric equations (see (2.6)–(2.7) in section 2 and (3.11) in section 3), which lead to some difficulties in obtaining the lower bound of the density. Therefore we will use the radial symmetric system only on the annular domain $\Omega_\varepsilon = \Omega \setminus \bar{B}_\varepsilon(0)$, where Ω is a ball of radius R centered at the origin in $\mathbb{R}^3$ and $B_\varepsilon(0)$ is a ball with radius ε and center 0, to exclude the singularity at the origin when we construct approximate solutions, and rewrite the Lagrangian equation as a new form (see (3.21) in section 3) which makes it possible to obtain the lower bounds of the approximate solutions.

By the approach mentioned above, we can obtain a class of approximate solutions with the required a priori uniform estimates such as energy estimates and entropy estimates. However, it should be noted that such approximate solutions are defined and estimated on the annular domain $\Omega_\varepsilon = \Omega \setminus \bar{B}_\varepsilon(0)$, and the L^1-stability analysis as in [22] can provide the convergence of the terms in (1.1)–(1.2) for the approximate solutions away from $r = 0$. Thus, to take the limit of the approximate solutions to obtain weak solutions which are defined on the entire domain Ω, we need to define the approximate solutions on $B_\varepsilon(0)$. Note that the usual zero extensions as in [8, 10] are not suitable here, since such extension would yield that $\nabla\sqrt{\rho}$ belongs to $L^\infty(0,T;L^2_{loc}(\Omega\setminus\{0\}))$ only so that the nonlinear diffusion terms in the definition of weak solutions (see (2.15)) will not make sense. An appropriate extension is presented in this paper, one of whose advantages is that it preserves the uniform $L^\infty(0,T;H^1(\Omega))$ estimate of $\sqrt{\rho^\varepsilon}$ such that we can obtain the convergence of the pressure term $(\rho^\varepsilon)^\gamma$ and the diffusion terms which are difficult to be dealt with due to the density-dependent viscosity coefficients. Also, though it seems difficult to obtain some uniform estimates for $\mathbf{U}^j$ separately because of the possible appearance of the vacuum, an extra estimate for $\mathrm{esssup}_{0\le t\le T}\int_\Omega \rho^j|\mathbf{U}^j|^{2+\eta}dx$, with some small $\eta\in(0,1)$, guarantees the convergence of the nonlinear convection terms.

The subsequent contents of the paper are organized as follows. In section 2 we will present the main result of this paper. In section 3 we will show various a priori estimates of the solutions. In this section, the estimates depending on ε, especially the lower bound of the density, are building blocks for constructing the approximate solutions and entropy estimates. On the other hand, the estimates independent of ε are the starting point for the convergence of the approximate solutions. Based on these, in section 4, we will first construct approximate solutions and then take the limits to obtain the global existence of weak solutions of the original system.

2. Main result. Set $h(\rho)=\rho$ and $g(\rho)=0$ in (1.1)–(1.2). The isentropic compressible Navier–Stokes equations become

$$\rho_t+\mathrm{div}(\rho\mathbf{U})=0, \tag{2.1}$$

$$(\rho\mathbf{U})_t+\mathrm{div}(\rho\mathbf{U}\otimes\mathbf{U})-\mathrm{div}(\rho D(\mathbf{U}))+\nabla P(\rho)=0 \tag{2.2}$$

for $t\in(0,+\infty)$ and $\mathbf{x}\in\mathbb{R}^3$.

We are concerned with spherically symmetric solutions to (2.1)–(2.2) in a ball Ω of radius R centered at the origin in $\mathbb{R}^3$. To this end, we denote

$$|\mathbf{x}|=r,\quad \rho(\mathbf{x},t)=\rho(r,t),\quad \mathbf{U}(\mathbf{x},t)=u(r,t)\frac{\mathbf{x}}{r}. \tag{2.3}$$

The initial and boundary conditions of (2.1)–(2.2) are

$$(\rho,\rho\mathbf{U})|_{t=0}=(\rho_0,\mathbf{m}_0), \tag{2.4}$$

$$\mathbf{m}=\rho\mathbf{U}=0 \text{ on } \partial\Omega. \tag{2.5}$$

For simplicity, we will take $D(\mathbf{U})=\nabla\mathbf{U}$ in (2.2), though the full strain tensor could be considered in a similar way. This leads to the following system of equations for $r>0$:

$$\rho_t+(\rho u)_r+\frac{2\rho u}{r}=0, \tag{2.6}$$

$$(\rho u)_t+(\rho u^2+\rho^\gamma)_r+\frac{2\rho u^2}{r}-(\rho u_r)_r-\rho\left(\frac{2u}{r}\right)_r=0, \tag{2.7}$$

with the initial condition

$$(\rho,\rho u)|_{t=0}=(\rho_0,m_0) \tag{2.8}$$

and the boundary conditions

$$\rho u(0,t)=0,\quad \rho u(R,t)=0. \tag{2.9}$$

It is easy to get the following usual a priori energy estimate for smooth solutions to (2.6), (2.7), and (2.9):

$$\frac{d}{dt}\int_0^R \left(\frac{1}{2}\rho u^2 + \frac{1}{\gamma-1}\rho^\gamma\right) r^2 dr + \int_0^R \rho(u_r^2 r^2 + 2u^2) dr \le 0. \tag{2.10}$$

Morover, the system (2.1)–(2.2) admits an additional a priori estimate, as observed by Bresch, Desjardins, and Lin [3], which reads as follows.

LEMMA 2.1 (see [22]). *Assume that $h(\rho)$ and $g(\rho)$ are two C^2 functions such that*

$$g(\rho) = \rho h'(\rho) - h(\rho)$$

holds true. Then the following inequality holds for smooth solutions of (1.1)–(1.2) *with $\rho > 0$:*

$$\frac{d}{dt}\int_\Omega \left(\frac{1}{2}\rho|\mathbf{U} + \nabla\varphi(\rho)|^2 + \frac{1}{\gamma-1}\rho^\gamma\right) dx + \int_\Omega \nabla\varphi(\rho)\cdot\nabla\rho^\gamma dx \le 0, \tag{2.11}$$

with φ such that

$$\varphi'(\rho) = \frac{h'(\rho)}{\rho}.$$

In particular, for three-dimensional spherically symmetric equations (2.6)–(2.7), one has the following.

LEMMA 2.2. *If (ρ, u) is a smooth solution to* (2.6)–(2.9)*, with $\rho > 0$, then the following inequality holds:*

$$\frac{d}{dt}\int_0^R \left\{\frac{1}{2}\rho u^2 + \rho_r u + |(\sqrt{\rho})_r|^2\right\} r^2 dr + \int_0^R \frac{4}{\gamma}((\rho^{\gamma/2})_r r)^2 dr \le 0. \tag{2.12}$$

Now we give a definition of weak solutions to (2.1)–(2.5).

DEFINITION 2.1. *A pair $(\rho, \mathbf{U})$ is said to be a weak solution to* (2.1)–(2.2) *provided that*

(1) $\rho \ge 0$ *a.e., and*

$$\rho \in L^\infty(0,T; L^1(\Omega)\cap L^\gamma(\Omega)) \cap C([0,\infty); W^{1,\infty}(\Omega)^*),$$
$$\sqrt{\rho} \in L^\infty(0,T; H^1(\Omega)), \sqrt{\rho}\mathbf{U} \in L^\infty(0,T; L^2(\Omega)),$$

where $W^{1,\infty}(\Omega)^*$ is the dual space of $W^{1,\infty}(\Omega)$;

(2) *for any $t_2 \ge t_1 \ge 0$ and any $\zeta \in C^1(\bar\Omega \times [t_1, t_2])$, the mass equation* (2.1) *holds in the following sense:*

$$\int_\Omega \rho\zeta dx|_{t_1}^{t_2} = \int_{t_1}^{t_2}\int_\Omega (\rho\zeta_t + \rho\mathbf{U}\cdot\nabla\zeta) dxdt; \tag{2.13}$$

(3) *for any $\psi = (\psi^1, \psi^2, \psi^3) \in C^2(\bar\Omega \times [0,T])$ satisfying $\psi(\mathbf{x}, t) = 0$ on $\partial\Omega$ and $\psi(\mathbf{x}, T) = 0$, it holds that*

$$\begin{aligned}&\int_\Omega \mathbf{m_0}\cdot\psi(0,\cdot)dx + \int_0^T\int_\Omega [\sqrt{\rho}(\sqrt{\rho}\mathbf{U})\cdot\partial_t\psi + \sqrt{\rho}\mathbf{U}\otimes\sqrt{\rho}\mathbf{U} : \nabla\psi] dxdt\\ &+ \int_0^T\int_\Omega \rho^\gamma \mathrm{div}\psi dxdt + \langle\rho\nabla\mathbf{U}, \nabla\psi\rangle = 0,\end{aligned} \tag{2.14}$$

where the diffusion term makes sense when written as

$$\langle \rho\nabla\mathbf{U}, \nabla\psi\rangle = -\int_0^T\int_\Omega \sqrt{\rho}(\sqrt{\rho}\mathbf{U})\cdot\Delta\psi dxdt$$
$$-2\int_0^T\int_\Omega(\sqrt{\rho}\mathbf{U})\cdot(\nabla\sqrt{\rho}\cdot\nabla)\psi dxdt. \tag{2.15}$$

We remark that, in the definition of the weak solution, (2.15) implies $\rho\nabla U \in L^2(0,T;W^{-1,1}(\Omega))$, which follows from the fact that $\sqrt{\rho}\in L^\infty(0,T;H^1(\Omega))$, $\sqrt{\rho}\mathbf{U}\in L^\infty(0,T;L^2(\Omega))$. This will be shown in section 4.

In this paper, we will construct global three-dimensional spherically symmetric weak solutions to (2.1)–(2.2) with the initial-boundary conditions (2.4)–(2.5). The initial data are assumed to satisfy

$$\rho_0 \geq 0 \text{ a.e. in } \Omega; \quad \mathbf{m}_0 = 0 \text{ a.e. on } \{x\in\Omega|\rho_0(x)=0\}; \tag{2.16}$$

$$\rho_0\in W^{1,2}(\Omega); \quad \frac{\mathbf{m}_0^2}{\rho_0}\in L^1(\Omega); \quad \frac{\mathbf{m}_0^{2+\eta}}{\rho_0^{1+\eta}}\in L^1(\Omega), \tag{2.17}$$

where $\eta\in(0,1)$ is some small constant. It follows from (2.17) that

$$\rho_0\mathbf{U}_0^{2+\eta}\in L^1(\Omega); \quad \rho_0\mathbf{U}_0^2\in L^1(\Omega). \tag{2.18}$$

The main result of this paper can be stated as follows.

THEOREM 2.1. *For $N = 3$ and $1 < \gamma < 3$, if the initial data have the form*

$$\rho_0 = \rho_0(|\mathbf{x}|), \quad \mathbf{U}_0 = u_0(|\mathbf{x}|)\frac{\mathbf{x}}{r}$$

and satisfy (2.16)–(2.17)*, then the initial-boundary-value problem* (2.1)–(2.5) *has a global spherically symmetric weak solution*

$$\rho = \rho(|\mathbf{x}|,t), \quad \mathbf{U} = u(|\mathbf{x}|,t)\frac{\mathbf{x}}{r}$$

satisfying, for all $T > 0$,

$$\rho(\mathbf{x},t)\in C([0,T];L^{\frac{3}{2}}(\Omega)), \quad \sqrt{\rho}\mathbf{U}\in L^\infty(0,T;L^2(\Omega)), \tag{2.19}$$

$$\int_\Omega \rho(\mathbf{x},t)dx = \int_\Omega \rho_0(\mathbf{x})dx. \tag{2.20}$$

Moreover, it holds that

$$\sup_{t\in[0,T]}\int_\Omega\left(\frac{1}{2}\rho|\mathbf{U}|^2 + \frac{1}{\gamma-1}\rho^\gamma + |\nabla\sqrt{\rho}|^2 + \rho|U|^{2+\eta}\right)dx \leq C, \tag{2.21}$$

where C is a constant.

Remark 2.1. In fact, our analysis applies to slightly more general viscosity coefficients $h(\rho)$ and $g(\rho)$. For instance, our results hold true for the following situations:

(1) $h(\rho) = \rho^\alpha$ and $g(\rho) = (\alpha-1)\rho^\alpha$, with $\alpha > \frac{N-1}{N}$, where the restriction of α results from the Lamé viscosity coefficients relation (1.3) and the usual energy estimates.

(2) $h(\rho)$ and $g(\rho)$ satisfy the relation

$$g(\rho) = \rho h'(\rho) - h(\rho)$$

and some additional restrictions given in [22].

Remark 2.2. It can be checked easily that, for $N = 2$, the conclusions in Theorem 2.1 hold true for any $\gamma > 1$. Consequently, we obtain the existence of a global spherically symmetric solution to the Saint–Venant model for shallow water, which is a particular case of (2.1)–(2.2) with $N = 2, h(\rho) = \rho, g(\rho) = 0$, and $\gamma = 2$ (see [3, 19]).

Remark 2.3. It should be noted that the boundary condition (2.5) is appropriate from the physical point of view, since, if the vacuum appears on the boundary, the velocity itself is meaningless and the momentum can be controllable. On the other hand, if no vacuum appears on the boundary, the boundary condition (2.5) is equivalent to $\mathbf{U}(R, t) = 0$.

To make sense of the boundary condition (2.5) for weak solutions in Theorem 2.1, we note that $\mathbf{U} = u(r)\frac{\mathbf{x}}{r}$ and ρu satisfy

$$\int_0^R \rho\varphi r^2 dr|_{t_1}^{t_2} = \int_{t_1}^{t_2} \int_0^R (\rho\varphi_t + \rho u\varphi_r) r^2 dr dt \tag{2.22}$$

for functions $\varphi \in C^1([0, R] \times [t_1, t_2])$; see (4.25) in Proposition 4.5 in section 4. In fact, (2.22) holds for any φ which is Lipschitz continuous. In particular, set $\varphi(r, t) = \varphi_1(t)\varphi_2(r)$, where $\varphi_1(t)$ and $\varphi_2(r)$ are Lipschitz continuous functions satisfying $\varphi_1(t) \equiv 1$ in $[t_1, t_2]$ and

$$\varphi_2(r) = \begin{cases} 1, & r \in [0, R - \delta], \\ 1 - \frac{1}{\delta}(r - (R - \delta)), & r \in [R - \delta, R]. \end{cases}$$

Substituting $\varphi_1(t)$ and $\varphi_2(r)$ into (2.22) and using (2.20), one gets

$$\begin{aligned} \frac{1}{\delta}\left|\int_{t_1}^{t_2}\int_{R-\delta}^R \rho u r^2 dr dt\right| \leq \Bigg| &\int_{R-\delta}^R \rho(r, t_2)(\varphi_2(r) - 1) r^2 dr \\ &- \int_{R-\delta}^R \rho(r, t_1)(\varphi_2(r) - 1) r^2 dr \Bigg| \to 0 \end{aligned}$$

as $\delta \to 0$. This implies that $(\rho u)(R, t) = 0$ in the sense of trace.

3. Approximate solutions and their estimates. The key point of the proof of Theorem 2.1 is to construct smooth approximate solutions satisfying the a priori estimates required in the L^1-stability analysis. The crucial issue is to obtain lower and upper bounds of the density, as mentioned in the introduction. To this end, we study the following system as an approximate system of (2.1)–(2.2):

$$\rho_t + \operatorname{div}(\rho\mathbf{U}) = 0, \tag{3.1}$$

$$\begin{aligned} &(\rho\mathbf{U})_t + \operatorname{div}(\rho\mathbf{U} \otimes \mathbf{U}) - \operatorname{div}((\rho + \varepsilon\rho^{\frac{3}{4}})\nabla\mathbf{U}) + \nabla\left(\frac{\varepsilon}{4}\rho^{\frac{3}{4}}\operatorname{div}\mathbf{U}\right) \\ &\quad + \nabla P(\rho) = 0, \end{aligned} \tag{3.2}$$

where $\varepsilon > 0$ is a constant.

When $\rho(\mathbf{x},t)=\rho(r,t), \mathbf{U}(\mathbf{x},t)=u(r,t)\frac{\mathbf{x}}{r}$, (3.1)–(3.2) becomes

$$\rho_t+(\rho u)_r+\frac{2\rho u}{r}=0, \tag{3.3}$$

$$(\rho u)_t+(\rho u^2+\rho^\gamma)_r+\frac{2\rho u^2}{r}+(\rho+\varepsilon\rho^{\frac{3}{4}})_r\frac{2u}{r}=\left(\left(\rho+\frac{3\varepsilon}{4}\rho^{\frac{3}{4}}\right)\left(u_r+\frac{2u}{r}\right)\right)_r \tag{3.4}$$

for $r>0$. We will first construct the smooth solution of (3.3)–(3.4) in the truncated region $0<\varepsilon<r<R$ with the initial condition

$$(\rho,\rho u)(r,0)=(\rho_0+\varepsilon,m_0),$$

and boundary conditions

$$u(r,t)|_{r=\varepsilon}=0,\quad u(r,t)|_{r=R}=0. \tag{3.5}$$

For approximate solutions with a positive lower bound for the density, (3.5) is equivalent to $\rho u(r,t)|_{r=\varepsilon}=0$ and $\rho u(r,t)|_{r=R}=0$.

We assume that the initial data are smooth and satisfy the bounds (2.16)–(2.17) with constants independent of ε. As discussed in the introduction, we shall eventually take a sequence of inner radii ε_j tending to 0. From now on, the dependence on j will be suppressed if there would be no confusion.

In the following, we will state the energy and entropy estimates which have been presented in the preceding section for these approximate solutions.

LEMMA 3.1. *Let $(\rho^\varepsilon,u^\varepsilon)$ be smooth solutions to* (3.3)–(3.4) *defined on* $[\varepsilon,R]\times[0,T]$ *with boundary conditions* (3.5) *such that* $\rho^\varepsilon>0$. *Then there exists a constant* C *independent of* ε *such that*

$$\int_\varepsilon^R \rho^\varepsilon(r,t)r^2dr\le C, \tag{3.6}$$

$$\begin{aligned}&\int_\varepsilon^R\left(\frac{1}{2}\rho^\varepsilon(u^\varepsilon)^2+\frac{1}{\gamma-1}(\rho^\varepsilon)^\gamma\right)r^2dr\\&+\int_0^T\int_\varepsilon^R\left(\rho^\varepsilon+\frac{\varepsilon}{4}(\rho^\varepsilon)^{\frac{3}{4}}\right)((u_r^\varepsilon)^2r^2+(u^\varepsilon)^2)drdt\le C,\end{aligned} \tag{3.7}$$

$$\begin{aligned}&\int_\varepsilon^R\frac{1}{2}\rho^\varepsilon\left|u^\varepsilon+(\log\rho^\varepsilon)_r+\frac{3\varepsilon}{4}(\rho^\varepsilon)^{-\frac{5}{4}}\rho_r^\varepsilon\right|^2r^2dr\\&+\int_0^T\int_\varepsilon^R\left(\gamma(\rho^\varepsilon)^{\gamma-2}+\frac{3\varepsilon}{4}\gamma(\rho^\varepsilon)^{\gamma-\frac{9}{4}}\right)|\rho_r^\varepsilon|^2r^2drdt\le C.\end{aligned} \tag{3.8}$$

Remark 3.1. Note that $h_\varepsilon(\rho)=\rho+\varepsilon\rho^{\frac{3}{4}}$ and $g_\varepsilon(\rho)=-\frac{\varepsilon}{4}\rho^{\frac{3}{4}}$ satisfy the relation $g_\varepsilon(\rho)=\rho h_\varepsilon'(\rho)-h_\varepsilon(\rho)$. In general, one can choose to approximate the system (2.1)–(2.2) by taking

$$h_\varepsilon(\rho)=\rho+\varepsilon\rho^\alpha,\quad g_\varepsilon(h)=\varepsilon(\alpha-1)\rho^\alpha,$$

which satisfy

$$g_\varepsilon(\rho)=\rho h_\varepsilon'(\rho)-h_\varepsilon(\rho),$$

where $\frac{N-1}{N}<\alpha<1, N=2,3$. We take $\alpha=\frac{3}{4}$ for the three-dimensional case here.

To guarantee the global existence of these approximate solutions, we need to give some detailed estimates on the density. We start with the following pointwise bounds for ρ^ε.

LEMMA 3.2. *Given $\varepsilon > 0$, there is an absolute constant C independent of ε such that*

$$0 \le \rho^\varepsilon(r,t) \le \frac{C}{\varepsilon^2} \tag{3.9}$$

for $\varepsilon \le r \le R$ and $t \ge 0$.

Proof. To simplify the presentation, we drop the superscript ε. Let $r(t)$ denote a particle path, i.e.,

$$\frac{dr(t)}{dt} = u(r(t),t).$$

Then along the particle path, (3.3) can be solved to get

$$\rho(r(t),t)r^2 = \rho_0(r(0))r(0)^2 e^{-\int_0^t u_r(r(s),s)ds},$$

which implies that $\rho \ge 0$ provided that $\rho_0 \ge 0$.

It follows from (3.7) and (3.8) that

$$\int_\varepsilon^R \frac{\rho_r^2}{\rho} r^2 dr \le C \tag{3.10}$$

for some absolute constant C independent of ε.

Then the second inequality in (3.9) follows from (3.6) and (3.10). The proof of the lemma is finished. ∎

To derive the a priori estimates about the velocity of the approximate solutions, the crucial step is to obtain the lower bound of the density. To this end, we introduce Lagrangian coordinates for the radial system (3.3)–(3.4) as follows. Let $\varepsilon > 0$ be fixed, and define

$$x(r,t) = \int_\varepsilon^r \rho r^2 dr, \tau = t.$$

Set $\int_\varepsilon^1 \rho r^2 dr = 1$ for any fixed $\varepsilon > 0$ without loss of generality. Then

$$\frac{\partial x}{\partial r} = \rho r^2, \quad \frac{\partial x}{\partial t} = -\rho u r^2, \quad \frac{\partial \tau}{\partial r} = 0, \quad \frac{\partial \tau}{\partial t} = 1.$$

Then the system (3.3)–(3.4) becomes

$$\begin{cases} \rho_\tau + \rho^2(r^2u)_x = 0, \\ r^{-2}u_\tau + (\rho^\gamma)_x = [(\rho^2 + \frac{3\varepsilon}{4}\rho^{\frac{7}{4}})(r^2u)_x]_x - (\rho + \varepsilon\rho^{\frac{3}{4}})_x \frac{2u}{r} \end{cases} \tag{3.11}$$

for $\tau > 0$ and $0 \le x \le 1$.

The corresponding initial data are

$$(\rho, \rho u)(\cdot, 0) = (\rho_0 + \varepsilon, m_0),$$

and the boundary conditions are

$$u(0,\tau) = 0, \quad u(1,\tau) = 0. \tag{3.12}$$

For this system, the following a priori estimates hold.

LEMMA 3.3. *For all $\tau \in [0,T]$, it holds that*

$$\int_0^1 \left(\frac{u^2(x,\tau)}{2} + \frac{\rho^{\gamma-1}(x,\tau)}{\gamma-1}\right) dx + \int_0^\tau \int_0^1 \left(\frac{2u^2}{r^2} + \rho^2 u_x^2 r^4\right) dxds$$
$$+ \left(1-\frac{\lambda}{2}\right) \int_0^\tau \int_0^1 \varepsilon \frac{u^2}{\rho^{\frac14} r^2} dxds + \left(\frac34 - \frac{1}{2\lambda}\right) \int_0^\tau \int_0^1 \varepsilon \rho^{\frac74} u_x^2 r^4 dxds$$
$$\leq \int_0^1 \left(\frac{u_0^2}{2} + \frac{\rho_0^{\gamma-1}}{\gamma-1}\right) dx, \tag{3.13}$$

where $\lambda \in (\frac23, 2)$, and

$$0 \leq \rho(x,\tau) \leq C(\varepsilon, T), \tag{3.14}$$

$$\varepsilon \leq r(x,\tau) \leq R, \tag{3.15}$$

$$\int_0^1 u^4 dx + \int_0^\tau \int_0^1 \left(\frac{4u^4}{r^2} + 6\rho^2 u^2 u_x^2 r^4 + \frac{2\varepsilon u^4}{\rho^{\frac14} r^2} + \varepsilon \rho^{\frac74} u^2 u_x^2 r^4\right) dxds$$
$$\leq \int_0^1 u_0^4 dx + C(\varepsilon, T). \tag{3.16}$$

Proof. By multiplying $(3.11)_2$ by $r^2 u$ and using $(3.11)_1$ and some standard manipulations, one gets

$$\frac{d}{d\tau} \int_0^1 \left(\frac{u^2}{2} + \frac{\rho^{\gamma-1}}{\gamma-1}\right) dx + \int_0^1 \left(\frac{2u^2}{r^2} + \rho^2 u_x^2 r^4\right) dx + \int_0^1 \left\{\varepsilon \frac{u^2}{\rho^{\frac14} r^2} + \frac34 \varepsilon \rho^{\frac74} u_x^2 r^4\right\} dx$$
$$= \varepsilon \int_0^1 \rho^{\frac34} u u_x r dx \leq \frac{\lambda}{2} \int_0^1 \frac{\varepsilon u^2}{\rho^{\frac14} r^2} dx + \frac{1}{2\lambda} \int_0^1 \varepsilon \rho^{\frac74} u_x^2 r^4 dx$$

for any $\lambda \in (\frac23, 2)$. Thus (3.13) holds.

Next, (3.14) follows from Lemma 3.2, and (3.15) holds trivially.

Now we prove (3.16). By multiplying $(3.11)_2$ by $r^2 u^3$ and using $(3.11)_1$ and integration by parts, we have by direct computation that

$$\frac14 \frac{d}{d\tau} \int_0^1 u^4 dx + \int_0^1 \left(\frac{2u^4}{r^2} + 3\rho^2 u^2 u_x^2 r^4\right) dx + \int_0^1 \left(\frac{\varepsilon u^4}{\rho^{\frac14} r^2} + \frac94 \varepsilon \rho^{\frac74} u^2 u_x^2 r^4\right) dx$$
$$= 2\int_0^1 \varepsilon \rho^{\frac34} u^3 u_x r dx + \int_0^1 \left(\rho^{\gamma-1} \frac{2u^3}{r} + 3\rho^\gamma u^2 u_x r^2\right) dx. \tag{3.17}$$

Using Hölder and Young's inequality and Lemma 3.2, one can bound each term of the right-hand side of (3.17) as follows:

$$2\varepsilon \int_0^1 \rho^{\frac34} u^3 u_x r dx \leq \frac12 \int_0^1 \frac{\varepsilon u^4}{\rho^{\frac14} r^2} dx + 2\int_0^1 \varepsilon \rho^{\frac74} u^2 u_x^2 r^4 dx;$$
$$2\int_0^1 \rho^{\gamma-1} \frac{u^3}{r} dx \leq 2\left(\int_0^1 \rho^{4(\gamma-1)} r^2 dx\right)^{\frac14} \left(\int_0^1 \frac{u^4}{r^2} dx\right)^{\frac34} \leq \frac12 \int_0^1 \frac{u^4}{r^2} dx + C;$$

and

$$
\begin{aligned}
3\int_0^1 \rho^\gamma u^2 u_x r^2 dx &\le \frac{3}{2}\int_0^1 \rho^2 u^2 u_x^2 r^4 dx + \frac{3}{2}\left(\int_0^1 \rho^{4\gamma-4} r^2 dx\right)^{\frac{1}{2}} \left(\int_0^1 \frac{u^4}{r^2} dx\right)^{\frac{1}{2}} \\
&\le \frac{3}{2}\int_0^1 \rho^2 u^2 u_x^2 r^4 dx + \frac{1}{2}\int_0^1 \frac{u^4}{r^2} dx + C.
\end{aligned}
$$

Putting the above three estimates into (3.17) yields

$$
\frac{1}{4}\frac{d}{d\tau}\int_0^1 u^4 dx + \int_0^1 \left(\frac{u^4}{r^2} + \frac{3}{2}\rho^2 u^2 u_x^2 r^4\right) dx + \int_0^1 \left(\frac{\varepsilon u^4}{2\rho^{\frac{1}{4}} r^2} + \frac{1}{4}\varepsilon \rho^{\frac{7}{4}} u^2 u_x^2 r^4\right) dx \le C,
$$

which implies (3.16) directly. ∎

The following estimate can be obtained by modifying the analysis in [13].

LEMMA 3.4. *There is a positive constant* $C = C(\varepsilon, T, \|u_0\|_{L^4}, \|(\rho_0^{\frac{3}{4}})_x\|_{L^4})$ *such that*

$$
\int_0^1 ((\rho^{\frac{3}{4}})_x)^4 (x,\tau) dx \le C \tag{3.18}
$$

for any $\tau \in [0,T]$.

Proof. $(3.11)_1$ can be rewritten as

$$
(\rho + \varepsilon \rho^{\frac{3}{4}})_{x\tau} = -\left[\left(\rho^2 + \frac{3\varepsilon}{4}\rho^{\frac{7}{4}}\right)(r^2 u)_x\right]_x . \tag{3.19}
$$

Thus, substituting (3.19) into $(3.11)_2$ yields

$$
r^2(\rho + \varepsilon \rho^{\frac{3}{4}})_{x\tau} + (\rho + \varepsilon\rho^{\frac{3}{4}})_x 2ur = -u_\tau - (\rho^\gamma)_x r^2. \tag{3.20}
$$

Direct computation shows that

$$
\frac{\partial r}{\partial \tau} = u.
$$

Consequently, (3.20) becomes

$$
(r^2(\rho + \varepsilon\rho^{\frac{3}{4}})_x)_\tau = -u_\tau - (\rho^\gamma)_x r^2. \tag{3.21}
$$

Integrating over $[0,t]$ shows that

$$
\begin{aligned}
&u(x,t) - u_0(x) + \int_0^t (\rho^\gamma)_x r^2(x,s) ds \\
&= r_0^2 \left(\frac{4}{3}\rho_0^{\frac{1}{4}} + \varepsilon\right)\partial_x\left(\rho_0^{\frac{3}{4}}\right) - r^2\left(\frac{4}{3}\rho^{\frac{1}{4}} + \varepsilon\right)\partial_x\left(\rho^{\frac{3}{4}}\right).
\end{aligned} \tag{3.22}
$$

By multiplying (3.22) by $(\partial_x(\rho^{\frac{3}{4}})r^2)^3$ and integrating over $[0,1]$ with respect to x, one gets

$$
\begin{aligned}
\int_0^1 \left(\frac{4}{3}\rho^{\frac{1}{4}} + \varepsilon\right)\left(\partial_x(\rho^{\frac{3}{4}})r^2\right)^4 dx \le C\left(\int_0^1 \left(\partial_x\left(\rho^{\frac{3}{4}}\right)r^2\right)^4 dx\right)^{\frac{3}{4}} \Bigg\{ &\|u-u_0\|_{L^4} \\
+ \left\|\partial_x\left(\rho_0^{\frac{3}{4}}\right)\right\|_{L^4} + \left(\int_0^t \|\partial_x \rho^\gamma\|_{L^4}^4 ds\right)^{\frac{1}{4}} &\Bigg\}.
\end{aligned} \tag{3.23}
$$

Using Lemma 3.3, $\varepsilon \le r, r_0 \le R$, and Young's inequality, one gets from (3.23) that there is a positive constant C depending on $\varepsilon, T, \|u_0\|_{L^4}$, and $\|(\rho_0^{\frac{3}{4}})_x\|_{L^4}$ such that

$$\varepsilon \int_0^1 (\partial_x(\rho^{\frac{3}{4}})r^2)^4 dx \le \frac{\varepsilon}{2} \int_0^1 (\partial_x(\rho^{\frac{3}{4}})r^2)^4 dx + C\int_0^t \int_0^1 (\partial_x \rho^\gamma)^4 dxds + C, \tag{3.24}$$

whence

$$\int_0^1 (\partial_x(\rho^{\frac{3}{4}}))^4 dx \le C + C\int_0^t \max_{[0,1]}(\rho^{4\gamma-3}) \int_0^1 (\partial_x(\rho^{\frac{3}{4}}))^4 dxds. \tag{3.25}$$

By applying Gronwall's inequality to (3.25) and making use of Lemma 3.2, we obtain

$$\int_0^1 (\partial_x(\rho^{\frac{3}{4}}))^4 dx \le C. \tag{3.26}$$

This completes the proof. □

Now we can obtain a positive lower bound for the density.

LEMMA 3.5. *There is a positive constant* $C = C(\varepsilon, T, \|u_0\|_{L^4}, \|(\rho_0^{\frac{3}{4}})_x\|_{L^4})$ *such that*

$$\rho(x,\tau) \ge C \tag{3.27}$$

for all $x \in [0,1]$ *and* $\tau \in [0,T]$.

Proof. Set $v(x,\tau) = \frac{1}{\rho(x,\tau)}$, and $V(\tau) = \max_{[0,1]\times[0,\tau]} v(x,s)$. Then $(3.11)_1$ becomes $v_\tau = (ru)_x$, which implies that $\int_0^1 v(x,\tau)dx = \int_0^1 v(x,0)dx \le C_0$, due to (3.12). Then it follows from Sobolev's embedding $W^{1,1}([0,1]) \hookrightarrow L^\infty([0,1])$ that, for any $0 < \beta < 1$,

$$\begin{aligned} v^\beta(x,\tau) &\le \int_0^1 v^\beta(x,\tau)dx + \int_0^x |\partial_x v^\beta| dx \\ &\le \left(\int_0^1 v dx\right)^\beta + \beta\int_0^x v^{\beta+\frac{3}{4}} \left|\frac{\rho_x}{\rho^{\frac{1}{4}}}\right| dx \\ &\le C + C\beta V^\beta \left(\int_0^1 v dx\right)^{\frac{3}{4}} \left(\int_0^1 ((\rho^{\frac{3}{4}})_x)^4 dx\right)^{\frac{1}{4}} \le C + C\beta V^\beta. \end{aligned} \tag{3.28}$$

Thus by choosing $\beta > 0$ small enough, which may depend on ε and T, we obtain

$$V(T) \le C,$$

where C is a positive constant depending on $\varepsilon, T, \|u_0\|_{L^4}$, and $\|(\rho_0^{\frac{3}{4}})_x\|_{L^4}$. The proof of the lemma is completed. □

Remark 3.2. When we construct approximate solutions in section 4.1, the initial data ρ_0, u_0 will be replaced by the mollified ones $\rho_0^{\varepsilon,\delta}, u_0^{\varepsilon,\delta}$ (see section 4.1 for more details), and hence the positive constant in Lemmas 3.4 and 3.5 will be replaced by a positive constant which depends on $\varepsilon, T, \|u_0^{\varepsilon,\delta}\|_{L^4}$, and $\|[(\rho_0^{\varepsilon,\delta})^{\frac{3}{4}}]_x\|_{L^4}$.

4. Proof of Theorem 2.1. In this section, we will prove Theorem 2.1 by completing the constructions of approximate solutions, applying the a priori bounds of sections 2 and 3 to take appropriate limits.

4.1. The existence of the approximate solutions. Consider the following approximate system in Lagrangian coordinates:

$$\begin{cases} \rho_\tau + \rho^2 (r^2 u)_x = 0, \\ r^{-2} u_\tau + (\rho^\gamma)_x = [(\rho^2 + \frac{3\varepsilon}{4}\rho^{\frac{7}{4}})(r^2 u)_x]_x - (\rho + \varepsilon \rho^{\frac{3}{4}})_x \frac{2u}{r} \end{cases} \tag{4.1}$$

for $\tau > 0, 0 \le x \le 1$, with

$$(\rho, \rho u)(\cdot, 0) = (\rho_0 + \varepsilon, m_0)$$

and

$$u(0,\tau) = 0, \ u(1,\tau) = 0.$$

First we regularize the initial data as follows. Let J_δ be a standard mollifier (in r) of width δ. Let $(\rho_0 + \varepsilon, u_0)$ be the initial data in Eulerian coordinates, where $u_0 = \frac{m_0}{\rho_0 + \varepsilon}$.

(1) Extend $\rho_0 + \varepsilon$ continuously outside $[\varepsilon, R]$ by taking $\rho_0(\varepsilon) + \varepsilon$ on $[0, \varepsilon]$ and $\rho_0(R) + \varepsilon$ on $[R, \infty)$, mollify with J_δ, restrict it to $[\varepsilon, R]$, and then multiply by a constant to normalize the total mass to be

$$M_0 = \int_0^R (\rho_0 + \varepsilon) r^2 dr.$$

The resulting density function is denoted by $\rho_0^{\varepsilon,\delta}(r)$.

(2) Redefine u_0 to be zero on $[0, \varepsilon + 2\delta]$ and $[R - 2\delta, R]$, and then mollify it with J_δ to get the smooth approximate initial velocity denoted by $u_0^{\varepsilon,\delta}(r)$. Note that $u_0^{\varepsilon,\delta}(r)$ is identically zero on a neighborhood of $r = \varepsilon$ and $r = R$.

The resulting data $(\rho_0^{\varepsilon,\delta}, u_0^{\varepsilon,\delta})$ then satisfy the hypotheses (2.16)–(2.17) (uniformly bounded on ε and δ). For any fixed $\varepsilon > 0$, we denote the corresponding initial data in Lagrangian coordinates by $(\rho_0^\delta, u_0^\delta)$. Then $\rho_0^\delta \in C^{1+\beta}[0,1]$ and $u_0^\delta \in C^{2+\beta}[0,1]$ for any $0 < \beta < 1$. Moreover,

$$\rho_0^\delta \to \rho_0 + \varepsilon \text{ in } W^{1,2}([0,1]), \quad u_0^\delta \to u_0 \text{ in } L^2([0,1]) \tag{4.2}$$

as $\delta \to 0$ and

$$u_0^\delta(0,\tau) = u_0^\delta(1,\tau) = 0.$$

Now consider the initial-boundary-value problem (4.1) with the initial data $(\rho_0 + \varepsilon, u_0)$ replaced by $(\rho_0^\delta, u_0^\delta)$. Note, however, that ε is fixed and positive at this stage of the argument, so that there are no singularities in the equations, and the construction of these approximate solutions is essentially an one-dimensional problem. For this problem one can apply the standard argument (see [12, 16], for instance) to obtain the existence of a unique local solution (ρ^δ, u^δ), with $\rho^\delta, \rho^\delta_x, \rho^\delta_{\tau x}, u^\delta, u^\delta_x, u^\delta_\tau, u^\delta_{xx} \in C^{\beta,\beta/2}([0,1] \times [0,T^*])$ for some $T^* > 0$. It follows from Lemmas 3.2–3.5 and (4.2) that ρ^δ is bounded from below and above, $(u^\delta)^2$ and ρ^δ_x are bounded in $L^\infty([0,T]; L^2)$, and u^δ_x is bounded in $L^2([0,T]; L^2)$ for any $T > 0$ because of $\varepsilon < r < R$. Furthermore, one can differentiate the equations (4.1) and apply the energy method to derive bounds of high-order derivatives of (ρ^δ, u^δ). Then we can apply the Schauder theory for linear parabolic equations to conclude that the $C^{\beta,\beta/2}([0,1] \times [0,T])$-norms of ρ^δ, ρ^δ_x, $\rho^\delta_{\tau x}$, u^δ, u^δ_x, u^δ_τ and, u^δ_{xx} are bounded a priorily. Therefore, we can continue the local solution globally in time and obtain that there exists a unique global solution (ρ^δ, u^δ) of (4.1) with the initial data (ρ_0, u_0) replaced by $(\rho_0^\delta, u_0^\delta)$ such that, for any $T > 0$,

$$\rho^\delta, \rho_x^\delta, \rho_{\tau x}^\delta, u^\delta, u_x^\delta, u_\tau^\delta, u_{xx}^\delta \in C^{\beta,\beta/2}([0,1]\times[0,T])$$

for some $0<\beta<1$, and $\rho^\delta > 0$ on $[0,1]\times[0,T]$. This can be done in a similar way as in [14]. Thus the solution, which can be denoted as $(\rho^{\varepsilon,\delta}, u^{\varepsilon,\delta})$, satisfies (4.1). By transforming it into Euler coordinates again by

$$x = \int_\varepsilon^r \rho(r,\tau) r^2 dr, \tau = t,$$

we can obtain the solutions $(\rho^{\varepsilon,\delta}(r,t), u^{\varepsilon,\delta}(r,t))$ to the approximate system (3.3)–(3.4), and consequently Lemma 3.1 holds for these approximate solutions.

4.2. The passage to limit. So far, $(\rho^{\varepsilon,\delta}, u^{\varepsilon,\delta})$ are defined on $\varepsilon \le r \le R$. To take the limit $\{\varepsilon_j, \delta_j\} \to 0$, we extend $\rho^{\varepsilon_j,\delta_j}(r,t), u^{\varepsilon_j,\delta_j}(r,t)$ to the whole domain Ω in the following way:

$$\tilde\rho^{\varepsilon_j,\delta_j} = \begin{cases} \rho^{\varepsilon_j,\delta_j}(r,t),\ r\in[\varepsilon_j, R], \\ \rho^{\varepsilon_j,\delta_j}(\varepsilon_j,t),\ r\in[0,\varepsilon_j], \end{cases} \tag{4.3}$$

$$\tilde u^{\varepsilon_j,\delta_j} = \begin{cases} u^{\varepsilon_j,\delta_j}(r,t),\ r\in[\varepsilon_j, R], \\ 0,\ r\in[0,\varepsilon_j]. \end{cases} \tag{4.4}$$

For simplicity, we denote the obtained approximate solutions $\{\tilde\rho^{\varepsilon_j,\delta_j}, \tilde u^{\varepsilon_j,\delta_j}\}$ by $\{\rho^j, u^j\}$. Let $\rho^j(\mathbf{x},t) = \rho^j(r,t), \mathbf{U}^j(\mathbf{x},t) = u^j(r,t)\frac{\mathbf{x}}{r}$, and denote $\Omega_\varepsilon = \Omega\backslash B_\varepsilon(0)$ for $\epsilon > 0$ and $\Omega_{\frac1n} = \Omega\backslash B_{\frac1n}(0)$ for $n = 1,2,\ldots$.

The following then follows from Lemma 3.1.

Lemma 4.1. *There exists a constant C independent of ε and δ such that*

$$\sup_{t\in[0,T]} \int_{\Omega_{\varepsilon_j}} \rho^j(x,t)dx \le C, \tag{4.5}$$

$$\begin{aligned}&\sup_{t\in[0,T]} \int_{\Omega_{\varepsilon_j}} \left(\frac12 \rho^j|\mathbf{U}^j|^2 + \frac{1}{\gamma-1}(\rho^j)^\gamma\right)(x,t)dx + \int_0^T\int_{\Omega_{\varepsilon_j}} \rho^j|\nabla\mathbf{U}^j|^2(x,t)dxdt \\ &\quad + \frac14\int_0^T\int_{\Omega_{\varepsilon_j}} \varepsilon(\rho^j)^{\frac34}|\nabla\mathbf{U}^j|^2(x,t)dxdt \le C,\end{aligned} \tag{4.6}$$

$$\begin{aligned}&\sup_{t\in[0,T]} \int_{\Omega_{\varepsilon_j}} \frac12\rho^j\left|\mathbf{U}^j + \nabla\log\rho^j + \frac34\varepsilon(\rho^j)^{-\frac54}\nabla\rho^j\right|^2(x,t)dx \\ &\quad + \int_0^T\int_{\Omega_{\varepsilon_j}} \left(\frac4\gamma|\nabla(\rho^j)^{\frac\gamma2}|^2 + \frac{48\varepsilon\gamma}{(4\gamma-1)^2}|\nabla(\rho^j)^{\frac{4\gamma-1}{8}}|^2\right)(x,t)dxdt \le C.\end{aligned} \tag{4.7}$$

Moreover, the following uniform estimate holds:

$$\sup_{t\in[0,T]}\left(\|\sqrt{\rho^j}\|_{H^1(\Omega)} + \int_\Omega \rho^j|\mathbf{U}^j|^2\right) \le C. \tag{4.8}$$

Proof. (4.5)–(4.7) follow directly from Lemma 3.1. It suffices to prove (4.8).

First, it holds that

$$\sup_{t\in[0,T]} \|\nabla\sqrt{\rho^j}\|_{L^2(\Omega)} \le C, \tag{4.9}$$

where C is a constant independent of ε and δ. Indeed, in view of the extension (4.3), one has

$$\partial_i\sqrt{\rho^j}(x,t) = \begin{cases} \partial_i\sqrt{\rho^j}, & x\in\Omega_{\varepsilon_j}, t\in[0,T],\\ 0, & x\in\overline{B}_{\varepsilon_j}, t\in[0,T]\end{cases}$$

for $i=1,2,3$. Consequently, (4.9) follows from (4.6) and (4.7).

Next, we verify that

$$\sup_{t\in[0,T]} \|\sqrt{\rho^j}\|_{L^2(\Omega)} \le C, \tag{4.10}$$

where C is a constant independent of ε and δ.

Thanks to the upper bound estimates of the density (3.9) and (4.5), one has

$$\begin{aligned}&\sup_{t\in[0,T]}\int_0^R \rho^j r^2 dr \le \sup_{t\in[0,T]}\int_0^{\varepsilon_j} \rho^j r^2 dr + \sup_{t\in[0,T]}\int_{\varepsilon_j}^R \rho^j r^2 dr\\ &\le \frac{C}{\varepsilon_j^2}\int_0^{\varepsilon_j} r^2 dr + C \le \frac{C\varepsilon_j}{3} + C \le C\end{aligned}$$

for all $0<\varepsilon_j<R$, which yields (4.10). The first part of (4.8) follows from (4.9) and (4.10). The other part of (4.8) can be checked easily, and the proof of the lemma is finished. ∎

Remark 4.1. Compared with the usual zero extensions in [10, 12], the extensions (4.3) and (4.4) keep the $L^\infty(0,T;H^1(\Omega))$-norm of $\sqrt{\rho^j}$, which will be used later.

PROPOSITION 4.1. *There exists a subsequence of $\{\rho^j\}$, still denoted by itself, such that, as $j\to\infty$,*

$$\rho^j(\mathbf{x},t)\to\rho(\mathbf{x},t) \tag{4.11}$$

strongly in $C([0,T],L^{3/2}(\Omega))$. Moreover, $\rho(\mathbf{x},t)=\rho(r,t)$ is a spherically symmetric function.

Proof. It follows from (4.8) that $\sqrt{\rho^j}$ is bounded in $L^\infty(0,T;L^q(\Omega))$ for $q\in[2,6]$. Thus ρ^j is bounded in $L^\infty(0,T;L^3(\Omega))$, and $\rho^j\mathbf{U}^j=\sqrt{\rho^j}\sqrt{\rho^j}\mathbf{U}^j$ is bounded in $L^\infty(0,T;L^{3/2}(\Omega))$ due to (4.8). The continuity equation yields that $\partial_t\rho^j$ is bounded in $L^\infty(0,T;W^{-1,3/2}(\Omega))$. Moreover, since $\nabla\rho^j=2\sqrt{\rho^j}\nabla\sqrt{\rho^j}$, we have that $\nabla\rho^j$ is bounded in $L^\infty(0,T;L^{3/2}(\Omega))$. Hence (4.11) is obtained thanks to the Aubin–Lions lemma. Clearly, $\rho(\mathbf{x},t)=\rho(r,t)$ is spherically symmetric. ∎

PROPOSITION 4.2. *Suppose that $1<\gamma<3$. Then $(\rho^j)^\gamma$ converges to ρ^γ strongly in $L^1((0,T);L^1(\Omega))$.*

Proof. This follows directly from the fact that ρ^j is bounded in $L^\infty(0,T;L^3(\Omega))$ and (4.12). ∎

The following proposition will enable us to take the limit in the nonlinear convection term.

PROPOSITION 4.3. *If $1<\gamma<3$ and*

$$\int_0^R \rho_0|u_0|^{2+\eta} r^2 dr \le C, \tag{4.12}$$

then the following estimate is true:

$$\frac{d}{dt}\int_{\varepsilon_j}^R \rho^j \frac{|u^j|^{2+\eta}}{2+\eta} r^2 dr + \int_{\varepsilon_j}^R \left(\frac{3}{4}\rho^j + \frac{\varepsilon}{8}(\rho^j)^{\frac{3}{4}}\right)|u^j|^\eta (u_r^j)^2 r^2 dr$$
$$+\int_{\varepsilon_j}^R \left(\frac{7}{4}\rho^j + \frac{3\varepsilon}{8}(\rho^j)^{\frac{3}{4}}\right)|u^j|^{\eta+2} dr \le C$$

for some small $\eta \in (0,1)$. *In particular,*

$$\int_{\Omega_{\varepsilon_j}} \rho^j \frac{|\mathbf{U}^j|^{2+\eta}}{2+\eta} dx \le C, \tag{4.13}$$

where $\Omega_{\varepsilon_j} = \Omega \setminus B_{\varepsilon_j}(0)$ *and* C *is a constant independent of* ε *and* δ.

To prove Proposition 4.3, we need the following lemma.

LEMMA 4.2. *The pressure* $(\rho^j)^\gamma$ *is bounded in* $L^{\frac{5}{3}}((0,T); L^{\frac{5}{3}}(\Omega_{\varepsilon_j}))$.

Proof. It follows from Lemma 4.1 that $(\rho^j)^{\gamma/2} \in L^2(0,T;H^1(\Omega_{\varepsilon_j}))$, and hence $(\rho^j)^\gamma \in L^1(0,T;L^3(\Omega_{\varepsilon_j}))$. Since $(\rho^j)^\gamma$ is bounded in $L^\infty(0,T;L^1(\Omega_{\varepsilon_j}))$ by (4.6), the Hölder inequality gives

$$\|(\rho^j)^\gamma\|_{L^{5/3}((0,T)\times\Omega_{\varepsilon_j})} \le \|(\rho^j)^\gamma\|^{2/5}_{L^\infty(0,T;L^1(\Omega_{\varepsilon_j}))} \|(\rho^j)^\gamma\|^{3/5}_{L^1(0,T;L^3(\Omega_{\varepsilon_j}))} \le C,$$

where C is independent of ε and δ. This finishes the proof of the lemma. ∎

Now we can prove Proposition 4.3.

Proof of Proposition 4.3. Let $\eta \in (0, \frac{1}{2})$. Multiplying (3.4) by $r^2 u^j |u^j|^\eta$ and integrating the resulting equation yield

$$\int_{\varepsilon_j}^R \rho^j \partial_t \frac{|u^j|^{2+\eta}}{2+\eta} r^2 dr + \int_{\varepsilon_j}^R \rho^j u^j \left(\frac{|u^j|^{2+\eta}}{2+\eta}\right)_r r^2 dr$$
$$+(1+\eta)\int_{\varepsilon_j}^R \left(\rho^j + \frac{3\varepsilon_j}{4}(\rho^j)^{\frac{3}{4}}\right)|u^j|^\eta (u_r^j)^2 r^2 dr$$
$$+\int_{\varepsilon_j}^R (2\rho^j + \varepsilon_j(\rho^j)^{\frac{3}{4}})|u^j|^{\eta+2} dr + \int_{\varepsilon_j}^R |u^j|^\eta\, u^j\, |((\rho^j)^\gamma)_r|\, r^2 dr$$
$$\le \left(\varepsilon_j + \frac{\eta\varepsilon_j}{2}\right)\int_{\varepsilon_j}^R (\rho^j)^{\frac{3}{4}} |u^j|^{\eta+1}|u_r^j| r dr$$
$$\le \left(\frac{\varepsilon_j}{2} + \frac{\eta\varepsilon_j}{4}\right)\left[\int_{\varepsilon_j}^R (\rho^j)^{\frac{3}{4}}|u^j|^\eta (u_r^j)^2 r^2 dr + \int_{\varepsilon_j}^R (\rho^j)^{\frac{3}{4}}|u^j|^{\eta+2} dr\right].$$

It follows that

$$\int_{\varepsilon_j}^R \rho^j \partial_t \frac{|u^j|^{2+\eta}}{2+\eta} r^2 dr + \int_{\varepsilon_j}^R \rho^j u^j \left(\frac{|u^j|^{2+\eta}}{2+\eta}\right)_r r^2 dr$$
$$+\int_{\varepsilon_j}^R \left(\rho^j + \frac{\varepsilon_j}{8}(\rho^j)^{\frac{3}{4}}\right)|u^j|^\eta (u_r^j)^2 r^2 dr + \int_{\varepsilon_j}^R \left(2\rho^j + \frac{3\varepsilon_j}{8}(\rho^j)^{\frac{3}{4}}\right)|u^j|^{\eta+2} dr$$
$$+\int_{\varepsilon_j}^R |u^j|^\eta u^j ((\rho^j)^\gamma)_r r^2 dr \le 0. \tag{4.14}$$

Moreover, multiplying (3.3) by $\frac{r^2|u^j|^{\eta+2}}{2+\eta}$ and integrating by parts show that

$$\int_{\varepsilon_j}^R \frac{|u^j|^{2+\eta}}{2+\eta}\partial_t\rho^j r^2 dr - \int_{\varepsilon_j}^R \rho^j u^j \left(\frac{|u^j|^{2+\eta}}{2+\eta}\right)_r r^2 dr = 0. \tag{4.15}$$

Summing over (4.14) and (4.15) leads to

$$\begin{aligned} &\frac{d}{dt}\int_{\varepsilon_j}^R \rho^j \frac{|u^j|^{2+\eta}}{2+\eta} r^2 dr + \int_{\varepsilon_j}^R \left(\rho^j + \frac{\varepsilon_j}{8}(\rho^j)^{\frac{3}{4}}\right)|u^j|^\eta (u_r^j)^2 r^2 dr \\ &+ \int_{\varepsilon_j}^R \left(2\rho^j + \frac{3\varepsilon_j}{8}(\rho^j)^{\frac{3}{4}}\right)|u^j|^{\eta+2} dr \le \left|\int_{\varepsilon_j}^R |u^j|^\eta\, u^j\, |((\rho^j)^\gamma)_r|\, r^2 dr\right|. \end{aligned} \tag{4.16}$$

Noting that

$$\begin{aligned} &\left|\int_{\varepsilon_j}^R |u^j|^\eta\, u^j\, |((\rho^j)^\gamma)_r|\, r^2 dr\right| \\ &\le (1+\eta)\left|\int_{\varepsilon_j}^R |u^j|^\eta |u_r^j| (\rho^j)^\gamma r^2 dr\right| + 2\int_{\varepsilon_j}^R |u^j|^{\eta+1}(\rho^j)^\gamma r dr \\ &\le \frac{1}{4}\int_{\varepsilon_j}^R \rho^j |u^j|^\eta |u_r^j|^2 r^2 dr + C\int_{\varepsilon_j}^R (\rho^j)^{2\gamma-1}|u^j|^\eta r^2 dr + 2\int_{\varepsilon_j}^R |u^j|^{\eta+1}(\rho^j)^\gamma r dr \\ &\le \frac{1}{4}\int_{\varepsilon_j}^R \rho^j |u^j|^\eta |u_r^j|^2 r^2 dr + C\int_{\varepsilon_j}^R ((\rho^j)^{2\gamma-1-\frac{\eta}{2}})^{\frac{2}{2-\eta}} r^2 dr + C \\ &\quad + C(R)\int_{\varepsilon_j}^R (\rho^j)^{(\gamma-\frac{\eta+1}{\eta+2})(2+\eta)} r^2 dr + \frac{1}{4}\int_{\varepsilon_j}^R \rho^j |u^j|^{2+\eta} dr, \end{aligned}$$

we obtain from (4.16) that

$$\begin{aligned} &\frac{d}{dt}\int_{\varepsilon_j}^R \rho^j \frac{|u^j|^{2+\eta}}{2+\eta} r^2 dr + \int_{\varepsilon_j}^R \left(\frac{3}{4}\rho^j + \frac{\varepsilon_j}{8}(\rho^j)^{\frac{3}{4}}\right)|u^j|^\eta (u_r^j)^2 r^2 dr \\ &+ \int_{\varepsilon_j}^R \left(\frac{7}{4}\rho^j + \frac{3\varepsilon_j}{8}(\rho^j)^{\frac{3}{4}}\right)|u^j|^{\eta+2} dr \\ &\le C\int_{\varepsilon_j}^R (\rho^j)^{(2\gamma-1-\frac{\eta}{2})\frac{2}{2-\eta}} r^2 dr + C(R)\int_{\varepsilon_j}^R (\rho^j)^{(\gamma-\frac{\eta+1}{\eta+2})(2+\eta)}\, r^2 dr + C. \end{aligned} \tag{4.17}$$

Using Lemma 4.2, one can check easily that the right-hand side of (4.17) is bounded for small η under the condition

$$2\gamma - 1 < \frac{5}{3}\gamma,$$

which is satisfied if $1 < \gamma < 3$. The proof of the proposition is finished. □

It is noted that (4.12) is satisfied due to (2.17). Moreover, it follows from (4.4) and (4.13) that

$$\int_\Omega \rho^j \frac{|\mathbf{U}^j|^{2+\eta}}{2+\eta} dx \le C. \tag{4.18}$$

Consequently, since

$$\int_\Omega (\rho^j|\mathbf{U}^j|^2)^{1+\zeta}dx \le \left(\int_\Omega \rho^j|\mathbf{U}^j|^{2+\eta}dx\right)^{\frac{2+2\zeta}{2+\eta}} \left(\int_\Omega (\rho^j)^{1+\frac{(2+\eta)\zeta}{\eta-2\zeta}}dx\right)^{\frac{\eta-2\zeta}{2+\eta}},$$

and as ζ is small enough, we deduce the following.

COROLLARY 4.1. *If $1<\gamma<3$, then $\sqrt{\rho^j}\mathbf{U}^j$ is bounded in $L^\infty(0,T;L^{2+2\zeta}(\Omega))$ for some small $\zeta>0$.*

Thanks to Propositions 4.1 and 4.3, Corollary 4.1, and Lemmas 4.4 and 4.6 in [25], we have the following.

PROPOSITION 4.4. (1) *Up to a subsequence, $\mathbf{m}^j=\rho^j\mathbf{U}^j$ converges strongly in $L^1((0,T)\times\Omega_{\frac{1}{n}})$ and $L^2(0,T;L^{1+\zeta}(\Omega_{\frac{1}{n}}))$ to some $\mathbf{m}(\mathbf{x},t)$ for any positive integer n.*

(2) *$\sqrt{\rho^j}\mathbf{U}^j$ converges strongly in $L^2((0,T)\times\Omega_{\frac{1}{n}})$ to $\frac{\mathbf{m}}{\sqrt{\rho}}$ (defined to be zero when $m=0$) for any positive integer n. In particular, $\mathbf{m}(x,t)=0$ a.e. on $\{\rho(\mathbf{x},t)=0\}$, and there exists a function $\mathbf{U}(\mathbf{x},t)$ such that*

$$\mathbf{m}(\mathbf{x},t)=\rho(\mathbf{x},t)\mathbf{U}(\mathbf{x},t).$$

This proposition can be proved exactly as in [22], and the details will be omitted.

The following then follows from Propositions 4.1 and 4.4.

COROLLARY 4.2. *Let $m^j(r,t)=\rho^ju^j(r,t)$. Then*

(1) *there exists a function $m(r,t)$ such that $\mathbf{m}(\mathbf{x},t)=m(r,t)\frac{\mathbf{x}}{r}$ and $m^j(r,t)=\rho^ju^j(r,t)$ converges to $m(r,t)$ strongly in $L^2(0,T;L^{1+\zeta}_{loc}((0,R);r^2dr))$;*

(2) *there exists a function $u(r,t)$ such that $\mathbf{U}(\mathbf{x},t)=u(r,t)\frac{\mathbf{x}}{r}$ and the quantity $\sqrt{\rho^j}u^j$ converges strongly in $L^2((0,T);L^2_{loc}((0,R);r^2dr))$ to $\frac{m}{\sqrt{\rho}}$ (defined to be zero when $m=0$).*

Proof. Since $\mathbf{m}^j(\mathbf{x},t)=m^j(r,t)\frac{\mathbf{x}}{r}$, so $m^j(r,t)=|\mathbf{m}^j(x,t)|$ converges almost everywhere to $m(r,t)=|\mathbf{m}(\mathbf{x},t)|$ due to the fact that $\mathbf{m}^j(\mathbf{x},t)$ converges almost everywhere to $\mathbf{m}(x,t)$ by the first part of Proposition 4.4. Therefore $\mathbf{m}(\mathbf{x},t)=m(r,t)\frac{\mathbf{x}}{r}$. Moreover, noting that $\rho(\mathbf{x},t)=\rho(r,t)$ by Proposition 4.1 and $\mathbf{m}(\mathbf{x},t)=\rho(\mathbf{x},t)\mathbf{U}(\mathbf{x},t)$ by Proposition 4.4, we obtain

$$m(r,t)\frac{\mathbf{x}}{r}=\rho(r,t)\mathbf{U}(\mathbf{x},t).$$

Therefore there exists a spherically function $u(r,t)$ such that $m(r,t)=\rho u(r,t)$.

The rest of the corollary follows directly from Proposition 4.4, and the proof of the corollary is finished. ◻

Now we show that $(\rho,\mathbf{U})$ obtained in Propositions 4.1–4.4 satisfy the weak form of (2.1), that is, (2.13) holds.

PROPOSITION 4.5. *Let $(\rho,\mathbf{U})$ be the limit described as in Propositions 4.1–4.4. Then (2.13) holds. Moreover, $\rho\in C([0,\infty);W^{1,\infty}(\Omega)^*)$.*

Proof. We first derive the weak form of (3.3). For any $\varphi(r,t)\in C^1([0,R]\times[t_1,t_2])$, it follows from (3.3), (4.3), and (4.4) that

$$\begin{aligned}&\int_0^R \rho^j\varphi r^2dr|_{t_1}^{t_2}-\int_{t_1}^{t_2}\int_0^R(\rho^j\varphi_t+\rho^ju^j\varphi_r)r^2drdt\\ &=\int_0^{\varepsilon_j}\rho^j\varphi r^2dr|_{t_1}^{t_2}-\int_{t_1}^{t_2}\int_0^{\varepsilon_j}\rho^j\varphi_tr^2drdt,\end{aligned}\tag{4.19}$$

Due to Proposition 4.1, it holds that

$$\int_0^R \rho^j \varphi r^2 dr \to \int_0^R \rho\varphi r^2 dr \tag{4.20}$$

and

$$\int_{t_1}^{t_2}\int_0^R \rho^j \varphi_t r^2 drdt \to \int_{t_1}^{t_2}\int_0^R \rho\varphi_t r^2 drdt \tag{4.21}$$

as $j \to \infty$.

It follows from (4.8) that $\sqrt{\rho^j}$ is bounded in $L^\infty(0,T;L^q(\Omega))$ for $q \in [2,6]$. Thus $\sqrt{\rho^j}$ (or its subsequence) converges strongly in $L^2(0,T;L^2(\Omega))$ to $\sqrt{\rho}$ due to Proposition 4.1. Moreover, Corollary 4.1 yields that $\sqrt{\rho^j}u^j$ is bounded in $L^\infty(0,T;L^{2+2\zeta}(\Omega))$ for some small $\zeta > 0$, and Corollary 4.2 yields that $\sqrt{\rho^j}u^j$ converges almost everywhere to $\sqrt{\rho}u$. Hence $\sqrt{\rho^j}u^j$ converges strongly to $\sqrt{\rho}u$ in $L^2(0,T;L^2(\Omega))$. So

$$\begin{aligned}&\int_{t_1}^{t_2}\int_0^R \rho^j u^j \varphi_r r^2 drdt = \int_{t_1}^{t_2}\int_0^R \sqrt{\rho^j}(\sqrt{\rho^j}u^j)\varphi_r r^2 drdt \to\\ &\int_{t_1}^{t_2}\int_0^R \sqrt{\rho}(\sqrt{\rho}u)\varphi_r r^2 drdt = \int_{t_1}^{t_2}\int_0^R \rho u\varphi_r r^2 drdt\end{aligned} \tag{4.22}$$

as $j \to \infty$.

Moreover, we have

$$\begin{aligned}&\left|\max_{t\in[0,T]}\int_0^{\varepsilon_j} \rho^j\varphi r^2 dr\right| \le C\max_{t\in[0,T]}\left(\int_0^R (\rho^j)^{\frac{3}{2}} r^2 dr\right)^{\frac{2}{3}}(\varepsilon_j)^{\frac{1}{3}}\\ &\le C(\varepsilon_j)^{\frac{1}{3}} \to 0\end{aligned} \tag{4.23}$$

and, similarly,

$$\begin{aligned}&\left|\int_{t_1}^{t_2}\int_0^{\varepsilon_j} \rho^j\varphi_t r^2 drdt\right|\\ &\le C(\varepsilon_j)^{\frac{1}{3}} \to 0\end{aligned} \tag{4.24}$$

as $j \to \infty$.

Therefore, in view of (4.20)–(4.24), by taking limit $j \to \infty$ in (4.19), we obtain

$$\int_0^R \rho\varphi r^2 dr|_{t_1}^{t_2} = \int_{t_1}^{t_2}\int_0^R (\rho\varphi_t + \rho u\varphi_r) r^2 drdt. \tag{4.25}$$

Now let $\zeta : \bar{\Omega} \times [t_1,t_2] \to \mathbb{R}$ be any C^1 function. Define

$$\varphi(r,t) := \int_S \zeta(ry,t)dS_y,$$

where the integral is over the unit sphere $S = S^2$ in $\mathbb{R}^3$. Then it follows from (4.25) that

$$\int_\Omega \rho\zeta(x,\cdot)dx|_{t_1}^{t_2} = \int_{t_1}^{t_2}\int_\Omega \{\rho\zeta_t + \rho(x,t)\mathbf{U}\cdot\nabla\zeta\}(x,t)dxdt.$$

This establishes the weak form of the mass equation.

Now we prove that $\rho \in C([0,\infty); W^{1,\infty}(\Omega)^*)$. If ϕ is a C^1 function of x, then by the continuity equation we have

$$\begin{aligned}
&\left|\int_\Omega \rho\phi dx|_{t_1}^{t_2}\right| = \left|\int_{t_1}^{t_2}\int_\Omega \rho(x,t)\mathbf{U}\cdot\nabla\phi dxdt\right| \\
&\le \|\nabla\phi\|_{L^\infty}\int_{t_1}^{t_2}\left(\int_\Omega \rho dx\right)^{1/2}\left(\int_\Omega \rho|\mathbf{U}|^2dx\right)^{1/2}dt \\
&\le C(T)\|\nabla\phi\|_{L^\infty}|t_2-t_1|.
\end{aligned}$$

This implies that

$$\|\rho(\cdot,t_2)-\rho(\cdot,t_1)\|_{W^{1,\infty}(\Omega)^*} \le C(T)|t_2-t_1|$$

for all $t_1,t_2\in[0,T]$. The proof of the proposition is complete. ∎

In the following, we prove that $(\rho,\mathbf{U})$ satisfies (2.14).

PROPOSITION 4.6. *Let* $(\rho,\mathbf{U})$ *be the limit described as in Propositions* 4.1–4.4. *Then* (2.14) *holds.*

Proof. Let ϕ be a C^2-function on $[0,R]\times[0,T]$ with $\phi(0,t)=\phi(R,t)=0$ for all $t\in[0,T]$. Then it follows from (3.4) that

$$\begin{aligned}
&\int_{\varepsilon_j}^R \rho_0^j u_0^j\phi(r,0)r^2dr + \int_0^T\int_{\varepsilon_j}^R\left(\rho^j u^j\phi_t + \rho^j(u^j)^2\phi_r + (\rho^j)^\gamma\left(\phi_r+\frac{2\phi}{r}\right)\right)r^2drdt \\
&-\int_0^T\int_{\varepsilon_j}^R \rho^j\left(u_r^j\phi_r+\frac{2u^j\phi}{r^2}\right)r^2drdt = \int_0^T\int_{\varepsilon_j}^R\frac{3}{4}\varepsilon_j(\rho^j)^{\frac34}\left(u_r^j+\frac{2u^j}{r}\right)\left(\phi_r+\frac{2}{r}\phi\right)r^2drdt \\
&-\int_0^T\int_{\varepsilon_j}^R\varepsilon_j(\rho^j)^{\frac34}\left(\frac{2u_r^j\phi}{r}+\frac{2u^j\phi_r}{r}+\frac{2}{r^2}u^j\phi\right)r^2drdt+\varepsilon_b^j, \qquad (4.26)
\end{aligned}$$

where

$$\varepsilon_b^j = \int_0^T\left\{\left[\rho^j+\frac34\varepsilon_j(\rho^j)^{\frac34}u_r^j\right](\varepsilon_j,t)\varepsilon_j^2\phi(\varepsilon_j,t)-\varepsilon_j^2(\rho^j)^\gamma(\varepsilon_j,t)\phi(\varepsilon_j,t)\right\}dt. \qquad (4.27)$$

We claim that

$$\lim_{\varepsilon_j\to0^+}\varepsilon_b^j = 0. \qquad (4.28)$$

To check this, we drop the superscript j and denote ε_j by ε for convenience. First, we show that

$$\lim_{\varepsilon\to0^+}\varepsilon^2\int_0^T\rho^\gamma(\varepsilon,t)\phi(\varepsilon,t)dt = 0. \qquad (4.29)$$

Indeed, note that

$$\begin{aligned}
&\left|\varepsilon^2\int_0^T\rho^\gamma(\varepsilon,t)\phi(\varepsilon,t)dt\right| \le \max_{0\le t\le T}|\phi(\varepsilon,t)|\int_0^T\varepsilon^2\rho^\gamma(\varepsilon,t)dt \\
&\le \max_{0\le t\le T}|\phi(\varepsilon,t)|\left[\int_0^T\int_\varepsilon^R\rho^\gamma(r,t)r^2drdt+\int_0^T\int_\varepsilon^R|\partial_r(\rho^\gamma)(r,t)|r^2drdt\right].
\end{aligned}$$

Since

$$\int_0^T\int_\varepsilon^R \rho^\gamma(r,t)r^2drdt \le C_0$$

and

$$\begin{aligned}\int_0^T\int_\varepsilon^R |\partial_r(\rho^\gamma)|r^2drdt &= 2\int_0^T\int_\varepsilon^R |\rho^{\frac{\gamma}{2}}||\partial_r(\rho^{\frac{\gamma}{2}})|r^2drdt\\ &\le \int_0^T\int_\varepsilon^R \rho^\gamma r^2drdt + \int_0^T\int_\varepsilon^R |\partial_r(\rho^{\frac{\gamma}{2}})|^2r^2drdt \le C_0\end{aligned}$$

due to (4.6) and (4.7), (4.29) follows from the fact that $\lim_{\varepsilon\to0^+}\max_{0\le t\le T}|\phi(\varepsilon,t)|=0$ since $\phi(0,t)\equiv 0$ and $\phi\in C^2$. Next, we show that

$$\lim_{\varepsilon\to0^+}\int_0^T(\rho u_r)(\varepsilon,t)\phi(\varepsilon,t)\varepsilon^2dt=0. \tag{4.30}$$

Thanks to (3.3) and the boundary condition that $u(\varepsilon,t)=0$, one has

$$\rho_t(\varepsilon,t)+\rho(\varepsilon,t)\partial_r u(\varepsilon,t)=0.$$

Thus,

$$\begin{aligned}&\lim_{\varepsilon\to0^+}\int_0^T(\rho u_r)(\varepsilon,t)\phi(\varepsilon,t)\varepsilon^2dt = \lim_{\varepsilon\to0^+}\left(-\varepsilon^2\int_0^T\partial_t\rho(\varepsilon,t)\phi(\varepsilon,t)dt\right)\\ &=\lim_{\varepsilon\to0^+}\left[\varepsilon^2\rho_0(\varepsilon)\phi(\varepsilon,0)+\varepsilon^2\int_0^T\rho(\varepsilon,t)\partial_t\phi(\varepsilon,t)dt\right]=\lim_{\varepsilon\to0^+}\left[\varepsilon^2\int_0^T\rho(\varepsilon,t)\partial_t\phi(\varepsilon,t)dt\right].\end{aligned}$$

On the other hand, it is easy to get

$$\begin{aligned}\varepsilon^2\left|\int_0^T\rho(\varepsilon,t)\partial_t\phi(\varepsilon,t)dt\right| &\le \varepsilon^{2-\frac{2}{\gamma}}\left(\varepsilon^2\int_0^T\rho^\gamma(\varepsilon,t)dt\right)^{\frac{1}{\gamma}}||\partial_t\phi(\varepsilon,\cdot)||_{L^{\frac{\gamma}{\gamma-1}}}\\ &\le C_0\,\varepsilon^{2-\frac{2}{\gamma}}\to 0\end{aligned}$$

as $\varepsilon\to0^+$. Hence (4.30) holds. Similarly, one can show that

$$\lim_{\varepsilon\to0^+}\frac{3}{4}\varepsilon\int_0^T\varepsilon^2\left(\rho^{\frac{3}{4}}u_r\right)(\varepsilon,t)\phi(\varepsilon,t)=0. \tag{4.31}$$

Indeed, it follows from (3.3) and $u(\varepsilon,t)=0$ that

$$\frac{3}{4}\int_0^T\varepsilon^3\left(\rho^{\frac{3}{4}}u_r\right)(\varepsilon,t)\phi(\varepsilon,t)dt=\varepsilon^3\rho_0^{\frac{3}{4}}(\varepsilon)\phi(\varepsilon,0)+\int_0^T\varepsilon^3\rho^{\frac{3}{4}}(\varepsilon,t)\partial_t\phi(\varepsilon,t)dt.$$

Since

$$\varepsilon^3\left|\int_0^T\rho^{\frac{3}{4}}(\varepsilon,t)\partial_t\phi(\varepsilon,t)dt\right|\le\varepsilon^{3(\frac{2\gamma-1}{2\gamma})}\left(\varepsilon^2\int_0^T\rho^\gamma(\varepsilon,t)dt\right)^{\frac{3}{4\gamma}}||\partial_t\phi(\varepsilon,\cdot)||_{L^{\frac{4\gamma}{4\gamma-3}}},$$

(4.31) follows. Now (4.28) is a consequence of (4.29)–(4.31).

Now, for any $\psi = (\psi^1, \psi^2, \psi^3) \in C^2(\bar{\Omega} \times [0, T])$ satisfying $\psi(x, t) = 0$ for all $x \in \partial\Omega$ and $\psi(x, T) = 0$, we set

$$\phi(r, t) = \int_S \psi(ry, t) \cdot y \, dS_y, \tag{4.32}$$

with $S = S^2$ being the unit sphere in $\mathbb{R}^3$, and transform the terms of (4.26) into integrals in Cartesian coordinates. Noting that

$$(r^2\phi)_r = \partial_r \int_{|x| \le r} \operatorname{div} \psi(x, t) dx = r^2 \int_S (\psi^i)_{x_i}(ry, t) dS_y,$$

we have by direct calculations that

$$\begin{aligned} &- \int_0^T \int_{\varepsilon_j}^R \rho^j \left(u_r^j \phi_r + \frac{2u^j \phi}{r^2} \right) r^2 dr dt \\ =&- \int_0^T \int_{\Omega_{\varepsilon_j}} \rho^j \nabla (\mathbf{U}^j)^i : \nabla \psi^i dx dt. \end{aligned}$$

Similarly, one has

$$\int_0^T \int_{\varepsilon_j}^R \frac{3}{4} \varepsilon_j (\rho^j)^{\frac{3}{4}} \left(u_r^j + \frac{2u^j}{r} \right) \left(\phi_r + \frac{2}{r} \phi \right) r^2 dr dt = \int_0^T \int_{\Omega_{\varepsilon_j}} \frac{1}{4} \varepsilon_j (\rho^j)^{\frac{3}{4}} \operatorname{div} \mathbf{U}^j \operatorname{div} \psi dx dt,$$

and

$$- \int_0^T \int_{\varepsilon_j}^R \varepsilon_j (\rho^j)^{\frac{3}{4}} \left(2 \frac{u_r^j \phi}{r} + 2 \frac{u^j \phi_r}{r} + \frac{2}{r^2} u^j \phi \right) r^2 dr dt = - \int_0^T \int_{\Omega_{\varepsilon_j}} \varepsilon_j (\rho^j)^{\frac{3}{4}} \nabla \mathbf{U}^j : \nabla \psi dx dt.$$

Thus, it follows from (4.26) that

$$\begin{aligned} &\int_{\Omega_{\varepsilon_j}} \rho_0^j \mathbf{U}_0^j \cdot \psi(0, x) dx + \int_0^T \int_{\Omega_{\varepsilon_j}} \Big\{ \sqrt{\rho^j} (\sqrt{\rho^j} \mathbf{U}^j) \cdot \partial_t \psi + \sqrt{\rho^j} \mathbf{U}^j \otimes \sqrt{\rho^j} \mathbf{U}^j : \nabla \psi \Big\} dx dt \\ &+ \int_0^T \int_{\Omega_{\varepsilon_j}} (\rho^j)^\gamma \operatorname{div} \psi dx dt - \int_0^T \int_{\Omega_{\varepsilon_j}} \rho^j \nabla \mathbf{U}^j : \nabla \psi dx dt \\ =& \frac{1}{4} \varepsilon_j \int_0^T \int_{\Omega_{\varepsilon_j}} (\rho^j)^{\frac{3}{4}} \operatorname{div} \mathbf{U}^j \operatorname{div} \psi dx dt - \varepsilon_j \int_0^T \int_{\Omega_{\varepsilon_j}} (\rho^j)^{\frac{3}{4}} \nabla \mathbf{U}^j : \nabla \psi dx dt + \varepsilon_b^j. \end{aligned}$$

Thanks to (4.4), one has

$$\begin{aligned} &\int_\Omega \rho_0^j \mathbf{U}_0^j \cdot \psi(0, \cdot) dx + \int_0^T \int_\Omega \left\{ \sqrt{\rho^j} (\sqrt{\rho^j} \mathbf{U}^j) \cdot \partial_t \psi + \sqrt{\rho^j} \mathbf{U}^j \otimes \sqrt{\rho^j} \mathbf{U}^j : \nabla \psi \right\} dx dt \\ &\qquad + \int_0^T \int_\Omega (\rho^j)^\gamma \operatorname{div} \psi dx dt - \int_0^T \int_\Omega \rho^j \nabla \mathbf{U}^j : \nabla \psi dx dt \\ =& \int_0^T \int_{B_{\varepsilon_j}} (\rho^j)^\gamma \operatorname{div} \psi dx dt + \frac{\varepsilon_j}{4} \int_0^T \int_{\Omega_{\varepsilon_j}} (\rho^j)^{\frac{3}{4}} \operatorname{div} \mathbf{U}^j \operatorname{div} \psi dx dt \\ &- \varepsilon_j \int_0^T \int_{\Omega_{\varepsilon_j}} (\rho^j)^{\frac{3}{4}} \nabla \mathbf{U}^j : \nabla \psi dx dt + \varepsilon_b^j. \end{aligned} \tag{4.33}$$

We proceed to show that each term on the left-hand side of (4.33) converges to a corresponding term in (2.14) and each term on the right-hand side of (4.33) vanishes as $j \to \infty$.

First, the proof of the convergence of $\rho^j \mathbf{U}^j \phi_t$ is similar to that of (4.22).

Next, it holds that

$$\begin{aligned}
&\left| \int_0^T \int_\Omega \left[\sqrt{\rho^j}\mathbf{U}^j \otimes \sqrt{\rho^j}\mathbf{U}^j - \sqrt{\rho}\mathbf{U} \otimes \sqrt{\rho}\mathbf{U} \right] : \nabla\psi dxdt \right| \\
\leq & \|\nabla\psi\|_{L^\infty} \int_0^T \int_{B_{\frac{1}{n}}} (|\sqrt{\rho^j}\mathbf{U}^j|^2 + \sqrt{\rho}|\mathbf{U}|^2) dxdt \\
&+ \left| \int_0^T \int_{\Omega_{\frac{1}{n}}} \left[\sqrt{\rho^j}\mathbf{U}^j \otimes \sqrt{\rho^j}\mathbf{U}^j - (\sqrt{\rho}\mathbf{U} \otimes \sqrt{\rho}\mathbf{U}) : \nabla\psi \right] dxdt \right|
\end{aligned} \tag{4.34}$$

for any positive integer n.

By virtue of Proposition 4.3, one has

$$\begin{aligned}
\int_0^T \int_{B_{\frac{1}{n}}} |\sqrt{\rho^j}\mathbf{U}^j|^2 dxdt \leq & \left(\int_0^T \int_{B_{\frac{1}{n}}} \rho^j dxdt \right)^{\frac{\eta}{2+\eta}} \left(\int_0^T \int_{B_{\frac{1}{n}}} \rho^j |\mathbf{U}^j|^{2+\eta} dxdt \right)^{\frac{2}{2+\eta}} \\
\leq & C \left(\int_0^T \int_{B_{\frac{1}{n}}} \rho^j dxdt \right)^{\frac{\eta}{2+\eta}}.
\end{aligned}$$

As proved in Proposition 4.5, the following convergence holds:

$$\begin{aligned}
\int_0^T \int_{B_{\frac{1}{n}}} \rho^j dxdt \leq & C(T) \left(\int_0^T \int_{B_{\frac{1}{n}}} (\rho^j)^3 dxdt \right)^{\frac{1}{3}} |B_{\frac{1}{n}}|^{\frac{2}{3}} \\
\leq & C(T) |B_{\frac{1}{n}}|^{\frac{2}{3}} \to 0
\end{aligned} \tag{4.35}$$

as $n \to \infty$, where (4.8) has been used. Consequently, it holds that

$$\int_0^T \int_{B_{\frac{1}{n}}} |\sqrt{\rho^j}\mathbf{U}^j|^2 dxdt \to 0,$$

uniformly on j, as $n \to \infty$. Also,

$$\int_0^T \int_{B_{\frac{1}{n}}} |\sqrt{\rho}\mathbf{U}|^2 dxdt \leq \lim\inf_{j\to\infty} \int_0^T \int_{B_{\frac{1}{n}}} |\sqrt{\rho^j}\mathbf{U}^j|^2 dxdt \to 0$$

as $n \to \infty$. It follows from (4.34) and Proposition 4.4 that

$$\int_0^T \int_\Omega \sqrt{\rho^j}\mathbf{U}^j \otimes \sqrt{\rho^j}\mathbf{U}^j : \nabla\psi dxdt \to \int_0^T \int_\Omega \sqrt{\rho}\mathbf{U} \otimes \sqrt{\rho}\mathbf{U} : \nabla\psi dxdt \tag{4.36}$$

as $j \to \infty$. Concerning the pressure term, Proposition 4.2 implies that

$$\int_0^T \int_\Omega (\rho^j)^\gamma \operatorname{div}\psi \, dxdt \to \int_0^T \int_\Omega \rho^\gamma \operatorname{div}\psi \, dxdt, \qquad j \to \infty. \tag{4.37}$$

Concerning the diffusion terms on the left-hand side of (4.33), we obtain after integration by parts that

$$\begin{aligned}\int_0^T\int_\Omega \rho^j\nabla\mathbf{U}^j:\nabla\psi dxdt = &-\int_0^T\int_\Omega \sqrt{\rho^j}(\sqrt{\rho^j}\mathbf{U}^j)\cdot\Delta\psi dxdt\\ &-2\int_0^T\int_\Omega(\sqrt{\rho^j}\mathbf{U}^j)\cdot(\nabla\sqrt{\rho^j}\cdot\nabla)\psi dxdt.\end{aligned}\tag{4.38}$$

Using Propositions 4.2–4.4, one can prove the convergence for the first term on the right-hand side of (4.38) as follows:

$$\int_0^T\int_\Omega \sqrt{\rho^j}(\sqrt{\rho^j}\mathbf{U}^j)\Delta\phi dxdt \to \int_0^T\int_\Omega \sqrt{\rho}(\sqrt{\rho}\mathbf{U})\Delta\phi dxdt\tag{4.39}$$

as $j\to\infty$, in a similar way as in the proof of (4.22).

Due to Lemma 4.1, it holds that

$$\|\nabla\sqrt{\rho^j}\|_{L^\infty(0,T;L^2(\Omega))}\le C,$$

and hence there exists a function $g\in L^2(0,T;L^2(\Omega))$ such that

$$\nabla\sqrt{\rho^j}\rightharpoonup g \quad \text{weakly in} \quad L^2(0,T;L^2(\Omega)).$$

Meanwhile, by Proposition 4.1, up to a subsequence, $\sqrt{\rho^j}$ converges almost everywhere to $\sqrt{\rho}$. Combining the fact that $\sqrt{\rho^j}$ is uniformly bounded in $L^\infty(0,T;L^6(\Omega))$, one has

$$\sqrt{\rho^j}\rightharpoonup \sqrt{\rho} \quad \text{weakly in} \quad L^2(0,T;L^2(\Omega)),$$

and hence $g=\nabla\sqrt{\rho}$. Consequently, it yields

$$\nabla\sqrt{\rho^j}\rightharpoonup \nabla\sqrt{\rho} \quad \text{weakly in} \quad L^2(0,T;L^2(\Omega)).$$

Due to Propositions 4.3 and 4.4, we finally obtain

$$\begin{aligned}&-2\int_0^T\int_\Omega(\sqrt{\rho^j}\mathbf{U}^j)\cdot(\nabla\sqrt{\rho^j}\cdot\nabla)\psi dxdt\to\\ &-2\int_0^T\int_\Omega(\sqrt{\rho}\mathbf{U})\cdot(\nabla\sqrt{\rho}\cdot\nabla)\psi dxdt.\end{aligned}\tag{4.40}$$

Substituting (4.39) and (4.40) into (4.38) yields

$$\begin{aligned}&\int_0^T\int_\Omega \rho^j\nabla\mathbf{U}^j:\nabla\psi dxdt\to\langle\rho\nabla\mathbf{U},\nabla\psi\rangle\\ &\equiv-\int_0^T\int_\Omega\sqrt{\rho}(\sqrt{\rho}\mathbf{U})\cdot\Delta\psi dxdt-2\int_0^T\int_\Omega(\sqrt{\rho}\mathbf{U})\cdot(\nabla\sqrt{\rho}\cdot\nabla)\psi dxdt.\end{aligned}\tag{4.41}$$

Up to now, we have proved that the terms on the left-hand side of (4.33) converge to corresponding ones in (2.14) as $j\to\infty$. In the following, we prove that each term on the right-hand side of (4.33) vanishes as $j\to\infty$.

Since $\sqrt{\rho^j}$ is uniformly bounded in $L^\infty(0,T;L^6(\Omega))$ due to (4.8), it holds that

$$\left|\int_0^T\int_{B_{\varepsilon_j}}(\rho^j)^\gamma \mathrm{div}\psi dxdt\right|\le C\left(\int_0^T\int_{B_{\varepsilon_j}}(\rho^j)^3dxdt\right)^{\frac{\gamma}{3}}|B_{\varepsilon_j}|^{\frac{3-\gamma}{3}}\le C|B_{\varepsilon_j}|^{\frac{3-\gamma}{3}} \tag{4.42}$$

for $1<\gamma<3$, which tends to zero as $\varepsilon_j\to 0$.

With the help of Lemma 4.1 again, one has

$$\begin{aligned}&\left|\frac{\varepsilon_j}{4}\int_0^T\int_{\Omega_{\varepsilon_j}}(\rho^j)^{\frac34}\mathrm{div}\mathbf{U}^j\mathrm{div}\psi dxdt\right|\\ \le& C\sqrt{\varepsilon_j}\left(\varepsilon_j\int_0^T\int_{\Omega_{\varepsilon_j}}(\rho^j)^{\frac34}|\nabla\mathbf{U}^j|^2dxdt\right)^{\frac12}\left(\int_0^T\int_{\Omega_{\varepsilon_j}}(\rho^j)^{\frac34}dxdt\right)^{\frac12}\\ \le& C\sqrt{\varepsilon_j}. \end{aligned}\tag{4.43}$$

Similarly, one has

$$\left|\varepsilon_j\int_0^T\int_{\Omega_j}(\rho^j)^{\frac34}\nabla\mathbf{U}^j:\nabla\psi dxdt\right|\le C\sqrt{\varepsilon_j}. \tag{4.44}$$

It follows from (4.28) and (4.42)–(4.44) that each term on the right hand side of (4.33) converges to 0 as $j\to\infty$.

Taking the limit $j\to\infty$ in (4.33), we finish the proof of the proposition. ◻

Now we are ready to prove Theorem 2.1.

Proof of Theorem 2.1. The weak forms of the mass and momentum equations (2.13) and (2.14) follow from Propositions 4.5 and 4.6, respectively. The first part in the definition of the weak solutions (see Definition 2.1) follows from Lemmas 3.2 and 4.1 and Proposition 4.5. Moreover, $\rho\in C([0,T];L^{\frac32}(\Omega))$ and the equation of mass conservation (2.20) are obtained by Propositions 4.1 and 4.5. The estimate (2.21) is due to Lemma 4.1 and (4.18). Finally, the radial symmetry of the weak solutions is a consequence of Corollary 4.2. The proof of Theorem 2.1 is thus finished.

Acknowledgment. The authors thank the anonymous referees for their valuable suggestions to improve the presentation of the paper.

REFERENCES

[1] D. Bresch and B. Desjardins, *Existence of global weak solutions for a 2D viscous shallow water equations and convergence to the quasi-geostrophic model*, Comm. Math. Phys., 238 (2003), pp. 211–223.

[2] D. Bresch and B. Desjardins, *On the construction of approximate solutions for the 2D viscous shallow water model and for compressible Navier-Stokes models*, J. Math. Pures Appl., 86 (2006), pp. 362–368.

[3] D. Bresch, B. Desjardins, and C.-K. Lin, *On some compressible fluid models: Korteweg, lubrication, and shallow water systems*, Comm. Partial Differential Equations, 28 (2003), pp. 843–868.

[4] D. Bresch, B. Desjardins, and D. Gerard-Varet, *On compressible Navier-Stokes equations with density dependent viscosities in bounded domains*, J. Math. Pures Appl., 87 (2007), pp. 227–235.

[5] B. Ducomet, E. Feireisl, H. Petzeltová, and I. Straškraba, *Existence globale pour un fluide barotrope autogravitant*, C.R. Acad. Sci. Paris Ser. I, 332 (2001), pp. 627–632.

[6] E. Feireisl, A. Novotný, and H. Petzeltová, *On the existence of globally defined weak solutions to the Navier-Stokes equations of isentropic compressible fluids*, J. Math. Fluid Mech., 3 (2001), pp. 358–392.

[7] H. Frid and V. Shelukhin, *Vanishing shear viscosity in the equations of compressible fluids for the flows with the cylinder symmetry*, SIAM J. Math. Anal., 31 (2000), pp. 1144–1156.

[8] D. Hoff, *Spherically symmetric solutions of the Navier-Stokes equations for compressible, isothermal flow with large, discontinuous initial data*, Indiana Univ. Math. J., 41 (1992), pp. 1225–1302.

[9] D. Hoff and D. Serre, *The failure of continuous dependence on initial data for the Navier-Stokes equations of compressible flow*, SIAM J. Appl. Math., 51 (1991), pp. 887–898.

[10] D. Hoff and H. K. Jenssen, *Symmetric nonbarotropic flows with large data and forces*, Arch. Ration. Mech. Anal., 173 (2004), pp. 297–343.

[11] S. Jiang, *Global smooth solutions of the equations of a viscous, heat-conducting one-dimensional gas with density-dependent viscosity*, Math. Nachr., 190 (1998), pp. 169–183.

[12] S. Jiang and R. Racke, *Evolution equations in thermoelasticity*, Monographs and Surveys in Pure and Applied Mathematics 112, Chapman & Hall/CRC, London, 2000.

[13] S. Jiang, Z. P. Xin, and P. Zhang, *Global weak solutions to* 1*D compressible isentropy Navier-Stokes with density-dependent viscosity*, Methods Appl. Anal., 12 (2005), pp. 239–252.

[14] S. Jiang and P. Zhang, *Global spherically symmetric solutions of the compressible isentropic Navier-Stokes equations*, Comm. Math. Phys., 215 (2001), pp. 559–581.

[15] S. Jiang and P. Zhang, *Axisymmetric solutions of the* 3*D Navier-Stokes equations for compressible isentropic fluids*, J. Math. Pures Appl., 82 (2003), pp. 949–973.

[16] S. Kawashima, *Systems of a Hyperbolic-Parabolic Composite Type, with Applications to the Equations of Magnetohydrodynamics*, thesis, Kyoto University, 1983.

[17] A. V. Kazhikhov and V. V. Shelukhin, *Unique global solution with respect to time of initial-boundary value problems for one-dimensional equations of a viscous gas*, J. Appl. Math. Mech., 41 (1977), pp. 273–282; translated from Prikl. Mat. Mekh., 41 (1977), pp. 282–291.

[18] H. L. Li, J. Li, and Z. P. Xin, *Vanishing of Vacuum States and Blow-up Phenomena of the Compressible Navier-Stokes Equations*, preprint, 2006. Research Report, IMS-2006-03(139), The Institute of Mathematical Sciences, The Chinese University of Hong Kong, http://www.ims.cuhk.edu.hk.

[19] P. L. Lions, *Mathematical Topics in Fluid Dynamics* 2*, Compressible Models*, Oxford Science Publication, Oxford, 1998.

[20] T. P. Liu, Z. P. Xin, and T. Yang, *Vacuum states of compressible flow*, Discrete Contin. Dyn. Syst., 4 (1998), pp. 1–32.

[21] A. Matsumura and T. Nishida, *The initial value problem for the equations of motion of viscous and heat-conductive gases*, J. Math. Kyoto Univ., 20 (1980), pp. 67–104.

[22] A. Mellet and A. Vasseur, *On the barotropic compressible Navier-Stokes equation*, Comm. Partial Differential Equations, 32 (2007), pp. 431–452.

[23] W. Sun, S. Jiang, and Z. Guo, *Helical symmetry solutions of the* 3*-D Navier-Stokes equations for compressible isentropic fluids*, J. Differential Equations, 222 (2006), pp. 263–296.

[24] L. Sundbye, *Global existence for the Dirichlet problem for the viscous shallow water equations*, J. Math. Anal. Appl., 202 (1996), pp. 236–258.

[25] S. W. Vong, T. Yang, and C. J. Zhu, *Compressible Navier-Stokes equations with degenerate viscosity coefficient and vacuum* II, J. Differential Equations, 192 (2003), pp. 475–501.

[26] Z. P. Xin, *Blow-up of smooth solution to the compressible Navier-Stokes equations with compact density*, Comm. Pure Appl. Math., 51 (1998), pp. 229–240.

[27] T. Yang, Z. A. Yao, and C. J. Zhu, *Compressible Navier-Stokes equations with density-dependent viscosity and vacuum*, Comm. Partial Differential Equations, 26 (2001), pp. 965–981.

[28] T. Yang and C. J. Zhu, *Compressible Navier-Stokes equations with degenerate viscosity coefficient and vacuum*, Comm. Math. Phys., 230 (2002), pp. 329–363.

SIAM J. MATH. ANAL.
Vol. 39, No. 5, pp. 1428–1455

WAVE-NUMBER-EXPLICIT BOUNDS IN TIME-HARMONIC SCATTERING*

SIMON N. CHANDLER-WILDE† AND PETER MONK‡

Abstract. In this paper we consider the problem of scattering of time-harmonic acoustic waves by a bounded, sound soft obstacle in two and three dimensions, studying dependence on the wave number in two classical formulations of this problem. The first is the standard weak formulation in the part of the exterior domain contained in a large sphere, with an exact Dirichlet-to-Neumann map applied on the boundary. The second formulation is as a second kind boundary integral equation in which the solution is sought as a combined single- and double-layer potential. For the variational formulation we obtain, in the case when the obstacle is starlike, explicit upper and lower bounds which show that the inf-sup constant decreases like k^{-1} as the wave number k increases. We also give an example where the obstacle is not starlike and the inf-sup constant decreases at least as fast as k^{-2}. For the boundary integral equation formulation, if the boundary is also Lipschitz and piecewise smooth, we show that the norm of the inverse boundary integral operator is bounded independently of k if the coupling parameter is chosen correctly. The methods we use also lead to explicit bounds on the solution of the scattering problem in the energy norm when the obstacle is starlike. The dependence of these bounds on the wave number and on the geometry is made explicit.

Key words. nonsmooth boundary, a priori estimate, inf-sup constant, Helmholtz equation, oscillatory integral operator

AMS subject classifications. 35J05, 35J20, 35J25, 42B10, 78A45

DOI. 10.1137/060662575

1. Introduction.

In this paper we consider the classical problem of scattering of a time-harmonic acoustic wave by a bounded, sound soft obstacle occupying a compact set $\Omega \subset \mathbb{R}^n$ ($n = 2$ or 3). The wave propagates in the exterior domain $\Omega_e = \mathbb{R}^n \setminus \Omega$, and the boundedness of the scatterer implies that there is an $R > 0$ such that $\{x \in \mathbb{R}^n : |x| > R\} \subset \Omega_e$. We suppose that the medium of propagation outside Ω_e is homogeneous, isotropic, and at rest, and that a time-harmonic ($\mathrm{e}^{-\mathrm{i}\omega t}$ time dependence) pressure field u^i is incident on Ω. Denoting by $c > 0$ the speed of sound, we assume that u^i is an entire solution of the Helmholtz (or reduced wave) equation with wave number $k = \omega/c > 0$. Then the problem we consider is to find the resulting time-harmonic acoustic pressure field u which satisfies the Helmholtz equation

$$\Delta u + k^2 u = 0 \quad \text{in } \Omega_e \tag{1.1}$$

and the sound soft boundary condition

$$u = 0 \quad \text{on } \Gamma := \partial\Omega_e, \tag{1.2}$$

*Received by the editors June 9, 2006; accepted for publication (in revised form) July 13, 2007; published electronically January 11, 2008. This work was supported by the UK Engineering and Physical Sciences Research Council under grant GR/S67401.
http://www.siam.org/journals/sima/39-5/66257.html
†Department of Mathematics, University of Reading, Whiteknights, PO Box 220, Berkshire, RG6 6AX, UK (S.N.Chandler-Wilde@reading.ac.uk).
‡Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 (monk@math.udel.edu). This author's research was supported by AFOSR grant F49620-02-1-0071.

and is such that the scattered part of the field, $u^s := u - u^i$, satisfies the Sommerfeld radiation condition

$$\frac{\partial u^s}{\partial r} - \mathrm{i}ku^s = o(r^{-(n-1)/2}) \tag{1.3}$$

as $r := |x| \to \infty$, uniformly in $\hat{x} := x/r$. (This latter condition expresses mathematically that the scattered field u^s is outgoing at infinity; see, e.g., [14]). It is well known that this problem has exactly one solution under the constraint that u and ∇u be locally square integrable; see, e.g., [34].

The aim of this paper is to understand the behavior, in the important but difficult high frequency limit $k \to \infty$, of two standard reformulations of this problem. Both reformulations are used extensively, for theoretical analysis and for practical numerical computation. The first is a weak formulation in the bounded domain $D_R := \{x \in \Omega_e : |x| < R\}$, for some $R > R_0 := \sup_{x\in\Omega} |x|$. This formulation is expressed in terms of the Dirichlet-to-Neumann map T_R, for the canonical domain $G_R := \{x : |x| > R\}$ with boundary $\Gamma_R := \{x : |x| = R\}$. The mapping T_R takes Dirichlet data $g \in C^\infty(\Gamma_R)$ to the corresponding Neumann data $T_R g := \frac{\partial v}{\partial r}|_{\Gamma_R}$, where v denotes the solution to the Helmholtz equation in G_R which satisfies the Sommerfeld radiation condition and the boundary condition $v = g$ on Γ_R. It is standard that the mapping T_R extends to a bounded map $T_R : H^{1/2}(\Gamma_R) \to H^{-1/2}(\Gamma_R)$.

Let V_R denote the closure of $\{v|_{D_R} : v \in C_0^\infty(\Omega_e)\} \subset H^1(D_R)$ in the norm of $H^1(D_R)$. It is well known (e.g., [39]), and follows easily by integration by parts, that u satisfies the scattering problem if and only if the restriction of u to D_R satisfies the following variational problem: find $u \in V_R$ such that

$$b(u, v) = G(v), \quad v \in V_R. \tag{1.4}$$

Here G is an antilinear functional that depends on the incident field (for details see section 3), while $b(\cdot, \cdot)$ is the sesquilinear form on $V_R \times V_R$ defined by

$$b(u, v) := \int_{D_R} (\nabla u \cdot \nabla \bar{v} - k^2 u\bar{v})\, dx - \int_{\Gamma_R} \gamma\bar{v} T_R \gamma u\, ds, \tag{1.5}$$

where $\gamma : V_R \to H^{1/2}(\Gamma_R)$ is the usual trace operator. Equation (1.4) is our first standard reformulation of the scattering problem.

To introduce our second reformulation, let $\Phi(x, y)$ denote the standard free-space fundamental solution of the Helmholtz equation given, in the two-dimensional (2D) and three-dimensional (3D) cases, by

$$\Phi(x, y) := \begin{cases} \dfrac{\mathrm{i}}{4} H_0^{(1)}(k|x - y|), & n = 2, \\[2ex] \dfrac{\mathrm{e}^{\mathrm{i}k|x-y|}}{4\pi|x - y|}, & n = 3, \end{cases}$$

for $x, y \in \mathbb{R}^n$, $x \neq y$. It was proposed independently by Brakhage and Werner [4], Leis [33], and Panich [40], as a means to obtain an integral equation uniquely solvable at all wave numbers, to look for a solution to the scattering problem in the form of the *combined single- and double-layer potential*

$$u^s(x) \;:=\; \int_\Gamma \frac{\partial \Phi(x, y)}{\partial \nu(y)} \varphi(y)\, ds(y) - \mathrm{i}\eta \int_\Gamma \Phi(x, y)\varphi(y)\, ds(y), \quad x \in \Omega_e, \tag{1.6}$$

for some nonzero value of the *coupling parameter* $\eta \in \mathbb{R}$. (In this equation $\partial/\partial\nu(y)$ is the derivative in the normal direction, the unit normal $\nu(y)$ directed into Ω_e.) It follows from standard boundary trace results for single- and double-layer potentials that u^s, given by (1.6), satisfies the scattering problem if and only if φ satisfies a second kind boundary integral equation on Γ (see section 4 for details). This integral equation, in operator form, is

$$(I + K - \mathrm{i}\eta S)\varphi = 2g, \tag{1.7}$$

where I is the identity operator, S and K are single- and double-layer potential operators, defined by (4.1) and (4.2) below, and $g := -u^i|_\Gamma$ is the Dirichlet data for the scattered field on Γ.

Choosing $\eta \neq 0$ ensures that (1.6) is uniquely solvable. Precisely,

$$A := I + K - \mathrm{i}\eta S$$

is invertible as an operator on $C(\Gamma)$ when Γ is sufficiently smooth, e.g., of class C^2 (see [4] or [14]). The case of nonsmooth (Lipschitz) Γ has been considered recently in [9] (see also [37]), where it is shown that A is invertible as an operator on the Sobolev space $H^s(\Gamma)$ for $0 \leq s \leq 1$.

While it is established that each of these formulations is well-posed, precisely that A^{-1} is a bounded operator on $H^s(\Gamma)$, $0 \leq s \leq 1$, in the case of (1.6), and that the sesquilinear form $b(\cdot,\cdot)$ satisfies the inf-sup condition, that

$$\alpha := \inf_{0 \neq u \in V_R} \sup_{0 \neq v \in V_R} \frac{|b(u,v)|}{\|u\|_{V_R}\|v\|_{V_R}} > 0, \tag{1.8}$$

in the case of the formulation (1.4), there is little information in the literature on how the stability constants $\|A^{-1}\|$ and α depend on k, particularly in the limit as $k \to \infty$.

This lack of theoretical understanding is unfortunate for a number of reasons. In the first place, both formulations (and similar formulations for other boundary conditions on Γ) are used extensively for numerical computation. Much research in recent years has been aimed at efficient solvers in the difficult high frequency case, where the scatterer Γ, and thus the region D_R, are large in diameter compared to the wavelength, so that the solution u is highly oscillatory and standard discretization methods require very many degrees of freedom. This effort has included many important developments for the solution of (1.6) and similar integral equations, including higher order boundary element or Nyström schemes (e.g., [22]), fast multipole methods (e.g., [17]), generalized boundary element methods using oscillatory basis functions (e.g., [5, 32, 19]), and preconditioners for iterative solvers (e.g., [12]). Similarly, for the solution of (1.4) at high frequency, important recent developments have included the use of higher order *hp*-finite element methods (e.g., [2, 18]), the use of oscillatory basis functions (e.g., [31] and, for methods based on more general variational formulations, [7, 21]), and ray-based techniques (e.g., [28]).

An essential ingredient in the development of numerical analysis for these methods, in particular analysis which seeks to determine the behavior of algorithms as the wave number increases, is an understanding of how the stability constants of numerical schemes depend on the wave number. Quantification of the dependence on k of $\|A^{-1}\|$ and α, i.e., of stability constants for the continuous formulation, is an important step in this direction.

An additional and important practical issue in connection to (1.6) is how to choose the parameter η. A natural criterion when using (1.6) for numerical computation is to choose η so as to minimize the condition number cond $A := \|A\| \, \|A^{-1}\|$ (e.g., Kress and Spassov [30] and Kress [29]). To determine this optimal choice, information on the dependence of $\|A^{-1}\|$ on k and η is required and will be obtained in section 4.

Given the practical importance of the questions we will address, it is not surprising that a number of relevant investigations have been carried out previously. In particular, a number of authors have studied (1.6), or related integral equations, in the canonical case when Γ is a cylinder or sphere, i.e., $\Gamma = \Gamma_R$ for some $R > 0$, especially with the aim of determining η so as to minimize the $L^2(\Gamma)$ condition number of A [30, 29, 3, 23, 6, 19]. Particularly relevant are the results of Giebermann [23], which have recently been completed and put on a rigorous footing by Dominguez, Graham, and Smyshlyaev [19]. It is shown in [19] that, in the 2D case, if the choice $\eta = k$ is made, then

$$\|A^{-1}\|_2 \leq 1 \tag{1.9}$$

for all sufficiently large k (we are using $\|\cdot\|_2$ to denote both the norm on $L^2(\Gamma)$ and the induced operator norm on the space of bounded linear operators on $L^2(\Gamma)$). This result is obtained as a consequence of the coercivity result that

$$\Re(A\psi, \psi) \geq \|\psi\|_2^2 \quad \forall \psi \in L^2(\Gamma), \tag{1.10}$$

where $(\cdot, \cdot)$ is the usual scalar product on $L^2(\Gamma)$. The same coercivity result, but without an explicit value for the constant, is shown in the 3D case [19], so that, for the case when Γ is a sphere, it also holds that $\|A^{-1}\|_2 = O(1)$ as $k \to \infty$. We note that, even for these canonical cases, establishing such bounds is not straightforward and depends on explicit calculations of the spectrum of A and careful estimates of Bessel functions uniformly in argument and order.

Research of relevance to the wave number dependence of (1.4) has also been carried out. Indeed an explicit estimate of the dependence of the inf-sup constant on the wave number has been made previously in two cases. The first is what may be thought of as a one-dimensional (1D) analogue of (1.4), with $V_R := \{u \in H^1(0,1) : u(0) = 0\}$ and $b(\cdot, \cdot)$ defined by $b(u, v) := \int_0^1 u'\bar{v}' - k^2 u\bar{v} dx - \mathrm{i}k\bar{v}(1)u(1)$. The results for this case, due to Ihlenburg and Babuška [26, 27], summarized in [25], are obtained via explicit calculations of the Green's function for the corresponding boundary value problem, i.e., the solution of $u'' + k^2 u = \delta_y$ on $(0, 1)$ with $u(0) = 0$, $u'(1) = \mathrm{i}ku(1)$, where δ_y is the delta distribution supported at $y \in (0, 1)$. For this 1D problem it is shown that, for some constants $C_1 \leq C_2$, the inf-sup constant given by (1.8), with $\|u\|_{V_R}^2 = \int_0^1 |u'|^2 dx$, satisfies

$$\frac{C_1}{k} \leq \alpha \leq \frac{C_2}{k}. \tag{1.11}$$

Closer still to the results of this paper is the work of Melenk [35] (see also Cummings and Feng [16]), who considers the Helmholtz equation in a bounded domain D, which is either convex or sufficiently smooth and starlike, with the impedance boundary condition $\frac{\partial u}{\partial \nu} = \mathrm{i}k\eta u$ on ∂D, with the normal directed out of D and $\eta > 0$. The sesquilinear form in their case is $b : H^1(D) \times H^1(D) \to \mathbb{C}$ given by

$$b(u, v) := \int_D (\nabla u \cdot \nabla \bar{v} - k^2 u\bar{v}) \, dx - \mathrm{i}k\eta \int_\Gamma \gamma\bar{v}\gamma u \, ds. \tag{1.12}$$

With this definition of $b(\cdot,\cdot)$ they show that their inf-sup constant α satisfies

$$\alpha \geq \frac{C}{k}, \tag{1.13}$$

for some constant $C > 0$. The technique of argument used in [35] and [16] is to derive a Rellich-type identity, this technique being of wide applicability in obtaining a priori estimates for solutions of boundary value problems for strongly elliptic systems of PDEs; see, e.g., [38, 36, 34]. This approach, essentially a carefully chosen application of the divergence theorem, appears to depend essentially on the starlike nature of the domain to obtain the wave-number-explicit bound (1.13).

The arguments of [35] and [16] will be one ingredient of the methods we use in this paper. The general structure of the arguments, though little of the detail, will borrow heavily from two of our own recent papers [10, 8], where we show results analogous to those presented here, but for the case of *rough surface scattering*, i.e., the case where Γ is unbounded and is the graph of some bounded continuous function, and Ω_e is its epigraph. Assuming that the axes are oriented so that Γ is bounded in the x_n-direction, i.e., $f_- \leq x_n \leq f_+$ for $x = (x_1, \ldots, x_n) \in \Gamma$, for some constants f_- and f_+, the analogous sesquilinear form to (1.5) for this case is given by the same formula, provided that one redefines D_R and Γ_R by $D_R := \{x \in \Omega_e : x_n < R\}$ and $\Gamma_R := \{x \in \Omega_e : x_n = R\}$, chooses $R \geq f_+$, and sets T_R to be the Dirichlet-to-Neumann map for the Helmholtz equation in the upper half-space $\{x : x_n > f_+\}$. This Dirichlet-to-Neumann map is given explicitly as a composition of a multiplication operator and Fourier transform operators. With this definition of the sesquilinear form $b(\cdot,\cdot)$ and with the inf-sup constant defined by (1.8) with

$$\|u\|_{V_R} := \left\{ \int_{D_R} \left(|\nabla u|^2 + k^2 |u|^2 \right) dx \right\}^{1/2}, \tag{1.14}$$

we show in [10] the explicit bound for the rough surface problem

$$\alpha \geq (1 + \sqrt{2}\,\kappa\,(\kappa + 1)^2)^{-1}, \tag{1.15}$$

where $\kappa := k(R - f_-)$. In [8] we study an integral equation formulation for the same problem in the case when, additionally, the function of which Γ is the graph is continuously differentiable. For the integral equation formulation (1.7) for this problem (with the twist that S and K are defined with the standard fundamental solution $\Phi(x, y)$ replaced by the Dirichlet Green's function for a half-space containing Ω_e), we show the bound

$$||A^{-1}||_2 < 2 + 2L + 4L^2 + \frac{k}{\eta} \left(2 + 5L + 3L^{3/2} \right), \tag{1.16}$$

where L is the maximum surface slope.

We note that Claeys and Haddar [13] have recently adapted the arguments of [10] to study 3D acoustic scattering from an unbounded sound soft rough tubular surface, as an initial model of electromagnetic scattering by an infinite wire with a perturbed surface. They study a weak formulation which can be written in the form (1.4) with a sesquilinear form which can be written as (1.5), provided that one redefines Γ_R to be the infinite cylinder $\Gamma_R := \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 = R^2\}$, D_R to be that part of the region outside the tubular surface but inside Γ_R, and T_R to be the appropriate Dirichlet-to-Neumann map for the Helmholtz equation in the region exterior to Γ_R. Their emphasis is on showing well-posedness for this problem, including showing that the inf-sup condition (1.8) holds, rather than on obtaining explicitly the k-dependence, but their results do imply a lower bound on α, that $\alpha^{-1} = O(k^3)$ as $k \to \infty$, the same k-dependence as for (1.15).

In this paper we will obtain analogous bounds to (1.11), (1.15), and (1.16) for the problem of scattering by a bounded sound soft obstacle. A major obstacle in achieving this aim is understanding the behavior of the Dirichlet-to-Neumann map T_R in sufficient detail. We address this issue in section 2, where our main new result is Lemma 2.1, a subtle property of radiating solutions of the Helmholtz equation, whose proof depends on a detailed understanding of monotonicity properties of Bessel functions. This lemma is essential to our results and, we expect, will be of value in deducing explicit bounds for a range of other wave scattering problems.

In section 3 we study the formulation (1.4). Our main results are, first, the upper bound on the inf-sup constant (1.8), which holds with no constraint on Γ, that

$$\alpha \leq \frac{C_1}{kR} + \frac{C_2}{k^2R^2},$$

where the constants $C_1 \geq 2\sqrt{2}$ and C_2 depend on the shape of the domain. (Our norm $\|\cdot\|_{V_R}$ in (1.8) is the wave number dependent norm given by (1.14).) In the case that the scattering obstacle Ω is starlike in the sense that $x \in \Omega$ implies $sx \in \Omega$ for $0 \leq s < 1$, we also show a lower bound, so that it holds that

$$\frac{1}{5 + 4\sqrt{2}kR} \leq \alpha \leq \frac{C_1}{kR} + \frac{C_2}{k^2R^2}. \tag{1.17}$$

We note that this bound establishes that, when Ω is starlike, α decreases like k^{-1} as $k \to \infty$ (cf. (1.11)). Finally, we produce an example (scattering by two parallel plates) for which

$$\alpha \leq \frac{C}{k^2R^2}$$

for some constant C and unbounded sequence of values of k, showing that the lower bound in (1.17) need not hold if Ω is not starlike. We emphasize that these appear to be the first bounds on the inf-sup constant in the literature for any problem of time-harmonic scattering by a bounded obstacle in more than one dimension that make the dependence on the wave number explicit.

We turn in section 4 to the integral equation formulation (1.7). We restrict our attention to the case when Ω is starlike and Γ is Lipschitz and piecewise smooth (e.g., a starlike polyhedron). Our main result is a bound on $\|A^{-1}\|_2$ (Theorem 4.3) as a function of three geometrical parameters and the ratio k/η of the wave number to the coupling parameter. Importantly this bound shows that, if the ratio k/η is kept fixed, then $\|A^{-1}\|_2$ remains bounded as $k \to \infty$. In particular, if the choice $\eta = k$ is made, then, for $kR_0 \geq 1$,

$$\|A^{-1}\|_2 \leq \frac{1}{2}\left(1 + \theta(4\theta + 4n + 1)\right), \tag{1.18}$$

where $n = 2$ or 3 is the dimension, $\theta := R_0/\delta_-$, and $\delta_- > 0$ is the essential infimum of $x \cdot \nu$ over the surface Γ (for example, $\theta = 1$ for a sphere, and $\theta = \sqrt{3}$ for a cube). A sharper (but more complicated) bound is given in Corollary 4.4. We note that a value of η/k in the range $1/2 \leq \eta/k \leq 1$ has been recommended based on studies of the integral operator A for circular and spherical geometries [30, 29, 3, 23, 19] as minimizing, approximately, the L^2 condition number of A.

We emphasize that the only results comparable to (1.18) to date are the bounds discussed above for the case when Γ is a circle or sphere, obtained by methods specialized to circular/spherical geometry. Even for this geometry the only completely explicit bound is (1.9), shown to hold for a circle for all sufficiently large k. Our general methods give the bound (4.15) for this case, which is almost as sharp a result, implying that, for every $\beta > 5/2$, $\|A^{-1}\|_2 \leq \beta$ for all sufficiently large kR_0, where R_0 is the radius of the circle.

2. Preliminaries. It is convenient to separate in an initial section two key lemmas which are essential ingredients in the arguments we will make to obtain wavenumber-explicit bounds for both of our formulations of the scattering problem, and to gather here other material common to both formulations.

Our arguments in this paper will depend on explicit representations for solutions of the Helmholtz equation in the exterior of a large ball. These depend in turn on explicit properties of cylindrical and spherical Bessel functions. For $\nu \geq 0$, let J_ν and Y_ν denote the usual Bessel functions of the first and second kind of order ν (see, e.g., [1] for definitions) and let $H_\nu^{(1)} := J_\nu + \mathrm{i}Y_\nu$ denote the corresponding Hankel function of the first kind of order ν. Of course, where C_ν denotes any linear combination of J_ν and Y_ν, it holds that C_ν is a solution of Bessel's equation of order ν, i.e.,

$$z^2 C_\nu''(z) + zC_\nu'(z) + (z^2 - \nu^2)C_\nu(z) = 0. \tag{2.1}$$

In the 3D case it is convenient to work also with the spherical Bessel functions j_m, y_m, and $h_m^{(1)} := j_m + \mathrm{i}y_m$ for $m = 0, 1, \ldots$. These can be defined directly (see, e.g., Nédélec [39]) by recurrence relations which imply that $h_m^{(1)}(z) = \mathrm{e}^{\mathrm{i}z}p_m(z^{-1})$, where p_m is a polynomial of degree m with $p_m(0) = 1$. Alternatively, the spherical Bessel functions can be defined in terms of the usual Bessel functions via the relations

$$j_m(z) = \sqrt{\frac{\pi}{2z}} J_{m+1/2}(z), \quad y_m(z) = \sqrt{\frac{\pi}{2z}} Y_{m+1/2}(z). \tag{2.2}$$

It is convenient to introduce the notation

$$M_\nu(z) := |H_\nu^{(1)}(z)|, \quad N_\nu(z) := |H_\nu^{(1)\prime}(z)|.$$

The arguments we make depend on the fact that $M_\nu(z)$ is decreasing on the positive real axis for $\nu \geq 0$; indeed, for $\nu \geq \frac{1}{2}$ it holds that $zM_\nu^2(z)$ is nonincreasing [42, sect. 13.74]. This latter fact, together with the asymptotics of $M_\nu(z)$ [1, eq. (9.2.28)] that

$$M_\nu(z) = \sqrt{\frac{2}{\pi z}} + O(z^{-5/2}) \text{ as } z \to \infty, \tag{2.3}$$

implies that

$$zM_\nu^2(z) \geq \frac{2}{\pi} \quad \text{for } z > 0,\ \nu \geq \frac{1}{2}. \tag{2.4}$$

It follows easily from the Bessel equation (2.1) that

$$(z^2 - \nu^2)\frac{d}{dz}\left(M_\nu^2(z)\right) + \frac{d}{dz}\left(z^2 N_\nu^2(z)\right) = 0.$$

Thus, defining the function A_ν for $\nu \geq 0$ by

$$A_\nu(z) := M_\nu^2(z)(z^2 - \nu^2) + z^2 N_\nu^2(z) - \frac{4z}{\pi}, \quad z > 0, \tag{2.5}$$

it holds that

$$A'_\nu(z) = 2zM_\nu^2(z) - \frac{4}{\pi}. \tag{2.6}$$

Thus $A'_\nu(z) \geq 0$ for $\nu \geq \frac{1}{2}$ and $z > 0$ by (2.4). Further, from (2.3) and the same asymptotics for N_ν [1, eq. (9.2.30)], i.e.,

$$N_\nu(z) = \sqrt{\frac{2}{\pi z}} + O(z^{-5/2}) \text{ as } z \to \infty,$$

it follows that $A_\nu(z) \to 0$ as $z \to \infty$ for $\nu \geq 0$. Thus

$$A_\nu(z) \leq 0 \quad \text{for } z > 0,\ \nu \geq \frac{1}{2}. \tag{2.7}$$

It is convenient in the following key lemma and later to use the notation $G_R := \{x : |x| > R\}$ for $R > 0$. In addition, throughout this paper $\nabla_T v$ denotes the tangential component of ∇v, i.e., $\nabla_T v := \nabla v - \nu \partial v / \partial \nu$.

LEMMA 2.1. *Suppose that, for some $R_0 > 0$, $v \in C^2(G_{R_0})$ satisfies the Helmholtz equation* (1.1) *in G_{R_0} and the Sommerfeld radiation condition* (1.3)*. Then, for $R > R_0$,*

$$\Im \int_{\Gamma_R} \bar{v} \frac{\partial v}{\partial r}\, ds \geq 0, \quad \Re \int_{\Gamma_R} \bar{v} \frac{\partial v}{\partial r}\, ds \leq 0, \tag{2.8}$$

and

$$\Re \int_{\Gamma_R} \bar{v} \frac{\partial v}{\partial r}\, ds + R \int_{\Gamma_R} \left(k^2 |v|^2 + \left| \frac{\partial v}{\partial r} \right|^2 - |\nabla_T v|^2 \right) ds \leq 2kR\, \Im \int_{\Gamma_R} \bar{v} \frac{\partial v}{\partial r}\, ds. \tag{2.9}$$

Remark 2.2. The first two inequalities (2.8) are well known; see, for example, Nédélec [39]. The third inequality appears to be new, but we note that an analogous inequality [11, Lemma 6.1], [10, Lemma 2.2] has been used extensively in the mathematical analysis of problems of scattering by unbounded rough surfaces. This inequality (proved easily by Fourier transform methods) can be viewed as a (formal) limit of (2.9) in the limit $R \to \infty$. Closer still to (2.9) is the recent inequality of Claeys and Haddar [13, Lemma 4.4], who study the Dirichlet-to-Neumann map for the Helmholtz equation in the exterior of an infinite cylinder in $\mathbb{R}^3$. In fact, their inequality implies, at least formally, the following less sharp version of (2.9) in the 2D case: for every $\rho_0 > 0$ there exists a constant $C > 0$ such that, provided that $kR > \rho_0$,

$$2\Re \int_{\Gamma_R} \bar{v} \frac{\partial v}{\partial r}\, ds + R \int_{\Gamma_R} \left(k^2 |v|^2 + \left| \frac{\partial v}{\partial r} \right|^2 - |\nabla_T v|^2 \right) ds \leq C(1 + kR)\, \Im \int_{\Gamma_R} \bar{v} \frac{\partial v}{\partial r}\, ds.$$

Proof. Note first that, by standard elliptic regularity results, it holds that $v \in C^\infty(G_{R_0})$. We now deal with the 2D and 3D cases separately.

Suppose first that $n = 2$. Choose $R_1 \in (R_0, R)$. Introducing standard cylindrical polar coordinates, we expand v on Γ_{R_1} as the Fourier series

$$v(x) = \sum_{m\in\mathbb{Z}} a_m \mathrm{e}^{\mathrm{i}m\theta},$$

where (R_1, θ) are the polar coordinates of x. Since $v \in C^\infty(\Gamma_{R_1})$ it holds that the series is rapidly converging, i.e., that $a_m = o(|m|^{-p})$ as $|m| \to \infty$ for every $p > 0$. It is standard that the corresponding Fourier series representation of v in G_{R_1} is

$$v(x) = \sum_{m\in\mathbb{Z}} a_m \mathrm{e}^{\mathrm{i}m\theta} \frac{H^{(1)}_{|m|}(kr)}{H^{(1)}_{|m|}(kR_1)}, \tag{2.10}$$

where (r, θ) are now the polar coordinates of x, and that this series, and all its partial derivatives with respect to r and θ, converge absolutely and uniformly in G_{R_1}. Hence, defining $c_m := (|a_m|^2 + |a_{-m}|^2)/|H^{(1)}_m(kR_1)|^2$ and $\rho := kR$, and using the orthogonality of $\{\mathrm{e}^{\mathrm{i}m\theta} : m \in \mathbb{Z}\}$, we see that

$$\begin{aligned}
\int_{\Gamma_R} \bar{v}\frac{\partial v}{\partial r}\, ds &= 2\pi\rho \sum_{m\in\mathbb{Z}} |a_m|^2 \frac{\overline{H^{(1)}_{|m|}(\rho)}\, {H^{(1)}_{|m|}}'(\rho)}{|H^{(1)}_{|m|}(kR_1)|^2} \\
&= 2\pi\rho \sum_{m=0}^{\infty} c_m \left(\Re\left(\overline{H^{(1)}_m(\rho)}\, {H^{(1)}_m}'(\rho)\right) + \mathrm{i}\,(J_m(\rho)Y_m'(\rho) - J_m'(\rho)Y_m(\rho))\right) \\
&= \sum_{m=0}^{\infty} c_m \left(\pi\rho \frac{d}{d\rho}\left(M_m^2(\rho)\right) + 4\mathrm{i}\right), \qquad (2.11)
\end{aligned}$$

where in the last step we have used the Wronskian formula [1, eq. (9.1.16)], i.e.,

$$\pi\rho(J_\nu(\rho)Y_\nu'(\rho) - J_\nu'(\rho)Y_\nu(\rho)) = 2. \tag{2.12}$$

Since $M_m(\rho)$ is decreasing on $(0, \infty)$ we see that (2.8) holds.

Similarly, noting that $|\nabla_T v| = R^{-1}|\partial v/\partial\theta|$, we calculate that

$$R\int_{\Gamma_R} \left(k^2|v|^2 + \left|\frac{\partial v}{\partial r}\right|^2 - |\nabla_T v|^2\right) ds = 2\pi \sum_{m=0}^{\infty} c_m \left(M_m^2(\rho)(\rho^2 - m^2) + \rho^2 N_m^2(\rho)\right).$$

From this equation and (2.11), and recalling the definition (2.5), we see that we will complete the proof of (2.9) if we can show the inequality

$$\frac{\rho}{2}\frac{d}{d\rho}\left(M_m^2(\rho)\right) + A_m(\rho) \le 0 \tag{2.13}$$

for $\rho > 0$ and $m = 0, 1, \ldots$.

By (2.7) and since M_m is decreasing on $(0, \infty)$, we see that (2.13) holds for $\rho > 0$ and $m \in \mathbb{N}$. To finish the proof of (2.9) in the case $n = 2$ we need to show (2.13) for $m = 0$, i.e., that

$$A(\rho) := \frac{\rho}{2}\frac{d}{d\rho}\left(M_0^2(\rho)\right) + A_0(\rho) \le 0, \quad \rho > 0.$$

Now $A(\rho) = \rho(J_0(\rho)J_0'(\rho) + Y_0(\rho)Y_0'(\rho)) + A_0(\rho)$; thus, using (2.1) and (2.6), it follows that

$$A'(\rho) = A_0'(\rho) + \rho\left(N_0^2(\rho) - M_0^2(\rho)\right) = \rho\left(M_0^2(\rho) + N_0^2(\rho)\right) - \frac{4}{\pi} = \frac{A_0(\rho)}{\rho}.$$

Thus

$$\frac{d}{d\rho}\left(\frac{A(\rho)}{\rho}\right) = \frac{A'(\rho)}{\rho} - \frac{A(\rho)}{\rho^2} = -\frac{1}{2\rho}\frac{d}{d\rho}\left(M_0^2(\rho)\right) \geq 0.$$

Also, since from the standard large argument asymptotics of the Bessel functions, $A(\rho)/\rho \to 0$ as $\rho \to \infty$, it follows that $A(\rho) \leq 0$ for $\rho > 0$. This completes the proof for $n = 2$.

We turn now to the 3D case for which we make analogous arguments, though the details are different. Again we choose $R_1 \in (R_0, R)$. Introducing standard spherical polar coordinates (r, θ, ϕ), we expand v on Γ_{R_1} as the spherical harmonic expansion

$$v(x) = \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} a_\ell^m Y_\ell^m(\theta, \phi), \tag{2.14}$$

where (R_1, θ, ϕ) are the spherical polar coordinates of x and the functions Y_ℓ^m, $m = -\ell, \dots, \ell$, are the standard spherical harmonics of order ℓ (see, for example, [39, Theorem 2.4.4]). We recall (e.g., [39]) that $\{Y_\ell^m : \ell = 0, 1, \dots,\ m = -\ell, \dots, \ell\}$ is a complete orthonormal sequence in $L^2(S)$, where $S := \{x : |x| = 1\}$ is the unit sphere, and an orthogonal sequence in $H^1(S)$. Since $v \in C^\infty(\Gamma_{R_1}) \supset H^m(S)$, for all $m \in \mathbb{N}$, it holds that the series is rapidly converging, i.e., that $a_\ell^m = o(|\ell|^{-p})$ as $|\ell| \to \infty$ for every $p > 0$ [39].

The solution of the Dirichlet problem for the Helmholtz equation in the exterior of a sphere is discussed in detail in [39]. It follows from (2.14) and [39, eq. (2.6.55)] that, for $x \in G_{R_1}$,

$$v(x) = \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} a_\ell^m Y_\ell^m(\theta, \phi)\frac{h_\ell^{(1)}(kr)}{h_\ell^{(1)}(kR_1)}, \tag{2.15}$$

where (r, θ, ϕ) are now the polar coordinates of x, and hence that [39, eqs. (2.6.70)–(2.6.74)]

$$\frac{\partial v}{\partial r}(x) = k\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} a_\ell^m Y_\ell^m(\theta, \phi)\frac{{h_\ell^{(1)}}'(kr)}{h_\ell^{(1)}(kR_1)}, \tag{2.16}$$

$$\nabla_T v(x) = \frac{1}{r}\sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} a_\ell^m \nabla_S Y_\ell^m(\theta, \phi)\frac{h_\ell^{(1)}(kr)}{h_\ell^{(1)}(kR_1)},$$

where ∇_S is the surface gradient operator on S and

$$\int_S |\nabla_S Y_\ell^m|^2\, ds = \ell(\ell + 1). \tag{2.17}$$

Hence, using the orthonormality in $L^2(S)$ of the spherical harmonics Y_ℓ^m, we see that, where $c_\ell := |h_\ell^{(1)}(kR_1)|^{-2} \sum_{m=-\ell}^{\ell} |a_\ell^m|^2$ and $\rho := kR$,

$$\int_{\Gamma_R} \bar{v} \frac{\partial v}{\partial r}\, ds = R^2 \int_S \bar{v}(R\hat{x}) \frac{\partial v}{\partial r}(R\hat{x})\, ds(\hat{x}) = R\rho \sum_{\ell=0}^{\infty} c_\ell\, \overline{h_\ell^{(1)}(\rho)} h_\ell^{(1)\prime}(\rho)$$

$$= R \sum_{\ell=0}^{\infty} c_\ell \left(\frac{\rho}{2} \frac{d}{d\rho} \left(|h_\ell^{(1)}(\rho)|^2 \right) + \frac{\mathrm{i}}{\rho} \right), \tag{2.18}$$

where in the last step we have used (2.2) and (2.12). Recalling that $|h_\ell^{(1)}(\rho)| = \sqrt{\pi/(2\rho)} M_{\ell+1/2}(\rho)$ is decreasing on $(0,\infty)$, we see that (2.8) holds.

Similarly, but also using the orthogonality of the surface gradients $\nabla_S Y_\ell^m$ in $L^2(S)$ and (2.17), we calculate that

$$\int_{\Gamma_R} \left(k^2 |v|^2 + \left| \frac{\partial v}{\partial r} \right|^2 - |\nabla_T v|^2 \right) ds$$
$$= \sum_{\ell=0}^{\infty} c_\ell \left(|h_\ell^{(1)}(\rho)|^2 (\rho^2 - \ell(\ell+1)) + \rho^2 |h_\ell^{(1)\prime}(\rho)|^2 \right).$$

Thus

$$\Re \int_{\Gamma_R} \bar{v} \frac{\partial v}{\partial r}\, ds + R \int_{\Gamma_R} \left(k^2 |v|^2 + \left| \frac{\partial v}{\partial r} \right|^2 - |\nabla_T v|^2 \right) ds - 2kR\, \Im \int_{\Gamma_R} \bar{v} \frac{\partial v}{\partial r}\, ds$$
$$= \frac{1}{k} \sum_{\ell=0}^{\infty} c_\ell B_\ell(\rho),$$

where

$$B_\ell(\rho) := \frac{\rho^2}{2} \frac{d}{d\rho} \left(|h_\ell^{(1)}(\rho)|^2 \right) + |h_\ell^{(1)}(\rho)|^2 \rho (\rho^2 - \ell(\ell+1)) + \rho^3 |h_\ell^{(1)\prime}(\rho)|^2 - 2\rho.$$

But straightforward calculations, using the definitions (2.2), yield that

$$B_\ell(\rho) = \frac{\pi}{2} A_{\ell+1/2}(\rho), \quad \rho > 0,\ \ell = 0, 1, \ldots.$$

Thus, applying (2.7), we see that $B_\ell(\rho) \le 0$ for $\ell = 0, 1, \ldots$ and $\rho > 0$, which completes the proof of (2.9). ∎

The following lemma is another key component in obtaining our wave-number-explicit bounds. Of course, the first equation is just a special case of Green's first theorem. The second is a Rellich–Payne–Weinberger identity, essentially that used in [35] to obtain an estimate for the solution of the Helmholtz equation with impedance boundary condition in an interior domain (see also [16]). In the case $k = 0$ it is a special case of a general identity for second order strongly elliptic operators given in Lemma 4.22 of [34] (see also [38, Chapter 5]). For completeness, we include the short proof of this key step in our arguments.

LEMMA 2.3. *Suppose that $G \subset \mathbb{R}^n$ is a bounded Lipschitz domain and that $v \in H^2(G)$. Then, for every $k \geq 0$, where $g := \Delta v + k^2 v$ and the unit normal vector ν is directed out of G, it holds that*

$$\int_G \left(|\nabla v|^2 - k^2|v|^2 + g\bar{v}\right)\, dx = \int_{\partial G} \bar{v}\,\frac{\partial v}{\partial \nu}\, ds \tag{2.19}$$

and

$$\int_G \left((2-n)|\nabla v|^2 + nk^2|v|^2 + 2\Re\left(gx\cdot\nabla\bar{v}\right)\right)\, dx$$

$$= \int_{\partial G} \left(x\cdot\nu \left(k^2|v|^2 + \left|\frac{\partial v}{\partial \nu}\right|^2 - |\nabla_T v|^2 \right) + 2\Re\left(x\cdot\nabla_T \bar{v}\,\frac{\partial v}{\partial \nu}\right)\right)\, ds. \tag{2.20}$$

Proof. In the case $v \in C^2(\bar{G})$, these equations are a consequence of the divergence theorem, $\int_G \nabla\cdot F dx = \int_{\partial G} f\cdot\nu ds$, which holds for every vector field $F \in C^1(\bar{G})$ (see, e.g., McLean [34] for the case when G is Lipschitz). Equation (2.19) follows by applying the divergence theorem to the identity $|\nabla v|^2 - k^2|v|^2 + g\bar{v} = \nabla\cdot(\bar{v}\nabla v)$ integrated over G. Equation (2.20) follows by applying the divergence theorem to the identity

$$(2-n)|\nabla v|^2 + nk^2|v|^2 + 2\Re\left(gx\cdot\nabla\bar{v}\right) = \nabla\cdot\left[x\left(k^2|v|^2 - |\nabla v|^2\right) + 2\Re\left(x\cdot\nabla\bar{v}\nabla v\right)\right]$$

integrated over G, and then noting that $x\cdot\nabla v = x\cdot\nu\frac{\partial v}{\partial\nu} + x\cdot\nabla_T v$ on ∂G. The extension from $C^2(\bar{G})$ to $H^2(G)$ follows by the density of $C^2(\bar{G})$ in $H^2(G)$ and by the continuity of the trace operator $\gamma : H^1(G) \to H^{1/2}(\partial G)$. ∎

3. The scattering problem and weak formulation. In this section we formulate the scattering problem precisely, state its weak formulation, and obtain explicit lower bounds on the inf-sup constant (1.8), using the results of the previous section.

To state precisely the scattering problem we wish to solve, let $H_0^1(\Omega_e) \subset H^1(\Omega_e)$ denote the closure of $C_0^\infty(\Omega_e)$, the set of C^∞ functions on Ω_e that are compactly supported, in the norm of the Sobolev space $H^1(\Omega_e)$. Let $H_0^{1,\mathrm{loc}}(\Omega_e)$ denote the set of those functions, v, that are locally integrable on Ω_e and satisfy that $\psi\chi \in H_0^1(\Omega_e)$ for every compactly supported $\chi \in C^\infty(\bar{\Omega}_e) := \{v|_{\bar{\Omega}_e} : v \in C^\infty(\mathbb{R}^n)\}$. Then our scattering problem can be stated as follows. For simplicity of exposition we restrict our attention throughout to two specific cases. The first is when the incident wave u^i is the plane wave

$$u^i(x) := \mathrm{e}^{\mathrm{i}kx_1}, \quad x \in \mathbb{R}^n. \tag{3.1}$$

THE PLANE WAVE SCATTERING PROBLEM. *Given $k > 0$, find $u \in H_0^{1,\mathrm{loc}}(\Omega_e) \cap C^2(\Omega_e)$ such that u satisfies the Helmholtz equation* (1.1) *in Ω_e, and $u^s := u - u^i$ satisfies the Sommerfeld radiation condition* (1.3) *as $r = |x| \to \infty$, uniformly in $\hat{x} = x/r$.*

The above is the scattering problem that we will focus on in this paper. But it is essential to our methods of argument in this section to also consider the following scattering problem where the source of the acoustic excitation is due to a compactly supported source region in Ω_e.

THE DISTRIBUTED SOURCE SCATTERING PROBLEM. *Given $k > 0$ and $g \in L^2(\Omega_e)$ which is compactly supported, find $u \in H_0^{1,\mathrm{loc}}(\Omega_e)$ such that u satisfies the inhomogeneous Helmholtz equation*

$$\Delta u + k^2 u = g \quad \text{in } \Omega_e \tag{3.2}$$

in a distributional sense, and u satisfies the Sommerfeld radiation condition (1.3) *as $r = |x| \to \infty$, uniformly in $\hat{x} = x/r$.*

Recall that, for $R > R_0 := \sup_{x\in\Omega} |x|$, we define $D_R := \{x \in \Omega_e : |x| < R\}$ and $V_R := \{v|_{D_R} : v \in H_0^1(\Omega_e)\} \subset H^1(D_R)$. We note that V_R is a closed subspace of $H^1(D_R)$. Throughout this section, $(\cdot,\cdot)$ will denote the standard scalar product on $L^2(D_R)$ and $\|\cdot\|_2$ the corresponding norm; i.e.,

$$(u, v) := \int_{D_R} u\bar{v}\,dx, \quad \|v\|_2 := (v, v)^{1/2} = \left\{\int_{D_R} |v|^2\,dx\right\}^{1/2}.$$

It is convenient to equip V_R with a wave number dependent norm, equivalent to the usual norm on $H^1(D_R)$ and defined by (1.14).

As discussed in section 1, our first reformulation of the plane wave scattering problem is the following weak formulation: find $u \in V_R$ such that

$$b(u, v) = G(v), \quad v \in V_R. \tag{3.3}$$

In this equation b is the bounded sesquilinear form on V_R given by (1.5), and $G \in V_R^*$, the dual space of V_R, is given by

$$G(v) = \int_{\Gamma_R} \bar{v}\left(\frac{\partial u^i}{\partial r} - T_R u^i\right) ds. \tag{3.4}$$

As defined in section 1, the operator $T_R : H^{1/2}(\Gamma_R) \to H^{-1/2}(\Gamma_R)$, which occurs in both (1.5) and (3.4), is the Dirichlet-to-Neumann map. Explicitly, in the 2D case, if $\phi \in H^{1/2}(\Gamma_R)$ has the Fourier series expansion

$$\phi(x) = \sum_{m\in\mathbb{Z}} a_m \mathrm{e}^{\mathrm{i}m\theta},$$

where (R, θ) are the polar coordinates of x, then (see (2.10) or [25])

$$T_R\phi(x) = k \sum_{m\in\mathbb{Z}} a_m \mathrm{e}^{\mathrm{i}m\theta} \frac{H_{|m|}^{(1)\prime}(kr)}{H_{|m|}^{(1)}(kR)}, \quad x \in \Gamma_R. \tag{3.5}$$

Similarly, in the 3D case, if $\phi \in H^{1/2}(\Gamma_R)$ has the expansion in spherical harmonics

$$\phi(x) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_\ell^m Y_\ell^m(\theta, \phi),$$

where (R, θ, ϕ) are the spherical polar coordinates of x, then (see (2.16) or [39])

$$T_R\phi(x) = k \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_\ell^m Y_\ell^m(\theta, \phi) \frac{h_\ell^{(1)\prime}(kr)}{h_\ell^{(1)}(kR)}, \quad x \in \Gamma_R, \tag{3.6}$$

where both series (3.5) and (3.6) are convergent in the norm of $H^{-1/2}(\Gamma_R)$. Moreover, from Lemma 2.1, we have the following key properties of T_R (see [39]).

COROLLARY 3.1. *For all $R > 0$ and all $\phi \in H^{1/2}(\Gamma_R)$ it holds that*

$$\Re \int_{\Gamma_R} \bar{\phi} T_R \phi \, ds \leq 0 \text{ and } \Im \int_{\Gamma_R} \bar{\phi} T_R \phi \, ds \geq 0.$$

That the plane wave scattering problem and the weak formulation (3.3) are equivalent is standard. Precisely, we have the following result (see, e.g., [25] or [39]).

THEOREM 3.2. *If u is a solution to the plane wave scattering problem, then $u|_{D_R} \in V_R$ satisfies* (3.3). *Conversely, suppose $u \in V_R$ satisfies* (3.3), *let $F_R := \gamma u^s$ be the trace of $u^s = u - u^i$ on Γ_R, and extend the definition of $u = u^i + u^s$ to Ω_e by setting $u^s|_{G_R}$ to be the solution of the Dirichlet problem in G_R, with data F_R on Γ_R (this solution is given explicitly by* (2.10) *and* (2.15), *in the cases $n = 2$ and $n = 3$, respectively). Then this extended function satisfies the plane wave scattering problem.*

In the case that $\text{supp}(g) \subset \overline{D_R}$, the distributed source scattering problem is equivalent, in the same precise sense as in the above theorem, to the following variational problem: find $u \in V_R$ such that

$$b(u, v) = -(g, v), \quad v \in V_R. \tag{3.7}$$

It is well known that both scattering problems have exactly one solution. Indeed this follows from the above equivalence and the fact that the variational problem (3.3) has exactly one solution $u \in V_R$ for every $G \in V_R^*$ (see, e.g., [25, 39]). In turn, this follows from uniqueness for the scattering problem (which follows from Rellich's lemma [14]) and from the fact that $b(\cdot, \cdot)$ satisfies a Gårding inequality [25, 39] (the first inequality in Corollary 3.1 plays a role here together with the compactness of the embedding operator from V_R to $L^2(D_R)$). Further, we have the following standard stability estimate (see [25, Remark 2.20]).

LEMMA 3.3. *The inf-sup condition* (1.8) *holds and, for all $u \in V_R$ and $\mathcal{G} \in V_R^*$ satisfying* (3.3), *it holds that*

$$\|u\|_{V_R} \leq C \|\mathcal{G}\|_{V_H^*}, \tag{3.8}$$

with $C = \alpha^{-1}$. Conversely, if there exists $C > 0$ such that, for all $u \in V_R$ and $\mathcal{G} \in V_R^*$ satisfying (3.3), *the bound* (3.8) *holds, then the inf-sup condition* (1.8) *holds with $\alpha \geq C^{-1}$.*

The second part of the above lemma shows that we obtain a lower bound on the inf-sup constant α if we show the bound (3.8) for all $u \in V_R$ and $\mathcal{G} \in V_R^*$ satisfying (3.3), and this will be the strategy that we will employ to obtain wave-number-explicit lower bounds on α. The following lemma reduces the problem of establishing (3.8) to that of establishing an a priori bound for solutions of the special case (3.7). The proof (very close to that of [10, Lemma 4.5]) depends on the observation that if $u \in V_R$ satisfies (3.3), then $u = u_0 + w$, where $u_0, w \in V_R$ satisfy

$$b_0(u_0, v) = \mathcal{G}(v) \quad \text{ and } \quad b(w, v) = 2k^2(u_0, v) \quad \forall v \in V_R,$$

where $b_0 : V_R \times V_R \to \mathbb{C}$ is defined by

$$b_0(u, v) = (\nabla u, \nabla v) + k^2(u, v) - \int_{\Gamma_R} \gamma \overline{v} \, T_R \gamma u \, ds, \quad u, v \in V_R.$$

It follows from Corollary 3.1 that $\Re\, b_0(v,v) \geq \|v\|^2_{V_R}$, $v \in V_R$, so that $\|u_0\|_{V_R} \leq \|\mathcal{G}\|_{V_R^*}$ by Lax–Milgram, and (3.9) and (1.14) imply that $\|w\|_{V_R} \leq 2k\tilde{C}\|u_0\|_2 \leq 2\tilde{C}\|\mathcal{G}\|_{V_R^*}$.

LEMMA 3.4. *Suppose there exists $\tilde{C} > 0$ such that, for all $u \in V_R$ and $g \in L^2(D_R)$ satisfying* (3.7), *it holds that*

$$\|u\|_{V_R} \leq k^{-1}\tilde{C}\,\|g\|_2. \tag{3.9}$$

Then, for all $u \in V_R$ and $\mathcal{G} \in V_R^*$ satisfying (3.3), *the bound* (3.8) *holds with $C \leq 1 + 2\,\tilde{C}$.*

We have reduced the problem of obtaining lower bounds on the inf-sup constant to the problem of obtaining a bound on the solution to a scattering problem, namely the distributed source scattering problem stated above. We will shortly bootstrap to the case where we require no smoothness on Γ, but our first bound on the solution to this problem is restricted to the case where Γ is smooth and Ω is starlike. Specifically, we require the following assumption.

Assumption 1. Let $S := \{x \in \mathbb{R}^n : |x| = 1\}$. For some $f \in C^\infty(S, \mathbb{R})$ with $\min_{\hat{x}\in S} f(\hat{x}) > 0$, it holds that $\Gamma = \{f(\hat{x})\hat{x} : \hat{x} \in S\}$.

LEMMA 3.5. *Suppose that Assumption* 1 *holds, u is a solution to the distributed source scattering problem, $R > R_0$, and* $\mathrm{supp}(g) \subset \overline{D_R}$. *Then*

$$k\|u\|_{V_R} \leq (n - 1 + 2\sqrt{2}kR)\|g\|_2. \tag{3.10}$$

Proof. Since D_R is a smooth domain, by standard elliptic regularity results [24] we have that $u \in H^{2,\mathrm{loc}}(D_R)$. Thus we can apply Lemma 2.3 to u in D_R to get, by adding $n-1$ times the real part of (2.19) to (2.20), that

$$\int_{D_R} \left(|\nabla u|^2 + k^2|u|^2 + \Re\left(g(2x\cdot\nabla\bar{u} + (n-1)\bar{u})\right)\right)\,dx = -\int_\Gamma x\cdot\nu\left|\frac{\partial u}{\partial\nu}\right|^2 ds$$
$$+ \int_{\Gamma_R}\left(R\left(k^2|u|^2 + \left|\frac{\partial u}{\partial r}\right|^2 - |\nabla_T u|^2\right) + \Re\left((n-1)\bar{u}\,\frac{\partial u}{\partial r}\right)\right)\,ds,$$

where we have also used the Dirichlet boundary condition (1.2), that is, $u = 0$ on Γ.

Since $x\cdot\nu > 0$ on Γ, applying Lemma 2.1 and then using (2.19), we see that

$$\int_{D_R}\left(|\nabla u|^2 + k^2|u|^2 + \Re\left(g(2x\cdot\nabla\bar{u} + (n-1)\bar{u})\right)\right)\,dx$$
$$\leq 2kR\,\Im\int_{\Gamma_R}\bar{u}\frac{\partial u}{\partial r}\,ds = 2kR\,\Im\int_{D_R} g\bar{u}\,dx.$$

Applying the Cauchy–Schwarz inequality and noting that

$$2ab \leq \epsilon a^2 + \frac{b^2}{\epsilon}, \tag{3.11}$$

for $a, b \geq 0$, $\epsilon > 0$, we deduce that

$$\|u\|^2_{V_R} \leq (n-1+2kR)\|g\|_2\|u\|_2 + 2R\|g\|_2\|\nabla u\|_2$$
$$\leq \frac{1}{2}\|u\|^2_{V_R} + \frac{\|g\|_2^2}{2k^2}(4k^2R^2 + (n-1+2kR)^2).$$

Thus

$$k^2\|u\|_{V_R}^2 \le \|g\|_2^2(4k^2R^2 + (n-1+2kR)^2),$$

from which (3.10) follows. ∎

Combining Lemmas 3.3, 3.4, and 3.5, we obtain the following corollary.

COROLLARY 3.6. *If Assumption* 1 *is satisfied, then the inf-sup condition* (1.8) *holds with* $\alpha^{-1} \le 1 + 2(n-1+2\sqrt{2}kR) \le 5 + 4\sqrt{2}kR$.

We proceed now to establish that Lemma 3.5 and Corollary 3.6 hold if Ω is starlike. Precisely, we require only the following, relaxed version of Assumption 1.

Assumption 2. It holds that $0 \notin \Omega_e$ and, if $x \in \Omega_e$, then $sx \in \Omega_e$ for every $s > 1$.

To establish these generalizations we first prove the following technical lemma (cf. [10, Lemma 4.10]).

LEMMA 3.7. *If Assumption* 2 *holds, then, for every* $\phi \in C_0^\infty(\Omega_e)$ *and* $R > R_0$, *there exists* $f \in C^\infty(S,\mathbb{R})$ *with* $\min_{\hat{x}\in S} f(\hat{x}) > 0$ *such that*

$$\operatorname{supp}\phi \subset \Omega_e' := \{sf(\hat{x}) \in \mathbb{R}^n : \hat{x} \in S,\ s > 1\}$$

and $\overline{G_R} \subset \Omega_e' \subset \Omega_e$.

Proof. Clearly, it is sufficient to consider the case when $R = 1$. So suppose $R = 1$, let $U := \operatorname{supp}\phi \cup \Gamma_1$, let $B := \{sx : x \in U,\ s \ge 1\}$, and let $\delta := \operatorname{dist}(U,\Gamma)/4$, so that $\operatorname{dist}(B,\Gamma) = \operatorname{dist}(U,\Gamma) = 4\delta$ and $0 < \delta \le \frac{1}{4}$. Let $B_\delta := \{x \in \mathbb{R}^n : \operatorname{dist}(x,B) < 2\delta\}$.

Let $N \in \mathbb{N}$ and $S_j \subset S$, $j = 1,\dots,N$, be such that each S_j is measurable and nonempty, $S_j \cap S_m = \emptyset$ for $j \ne m$, $S = \bigcup_{j=1}^N S_j$, and $\operatorname{diam}(S_j) \le \delta$, $j = 1,\dots,N$. For $j = 1,\dots,N$ choose $\hat{x}_j \in S_j$ and let

$$f_j := \inf\{|x| : x \in B_\delta,\ x/|x| \in S_j\}.$$

Then $2\delta \le f_j \le 1 - 2\delta$, $j = 1,\dots,N$. Define $\tilde{f} : S \to \mathbb{R}$ by

$$\tilde{f}(\hat{x}) := f_j \quad \text{ if } \hat{x} \in S_j,\ j = 1,\dots,N.$$

Then $\tilde{f} \in L^\infty(S,\mathbb{R})$; in fact, $\tilde{f}$ is a simple function and $2\delta \le \tilde{f}(\hat{x}) \le 1 - 2\delta$, $\hat{x} \in S$. Choose ϵ with $0 < \epsilon < \delta$ and let $J \in C^\infty[0,2]$ be such that $J \ge 0$, $J(t) = 0$ if $\epsilon^2/2 \le t \le 2$ and, where $e_3 := (0,0,1)$, such that $\int_S J(1 - e_3\cdot\hat{y})\,ds(\hat{y}) = 1$, so that $\int_S J(1 - \hat{x}\cdot\hat{y})\,ds(\hat{y}) = 1$, $\hat{x} \in S$. Define $f \in C^\infty(S,\mathbb{R})$ by

$$f(\hat{x}) := \int_S J(1 - \hat{x}\cdot\hat{y})\tilde{f}(\hat{y})\,ds(\hat{y}), \quad \hat{x} \in S,$$

and let Ω_e' be defined as in the statement of the lemma. Then f and Ω_e' have the properties claimed.

To see that this is true first note that, since $J(1 - \hat{x}\cdot\hat{y}) = 0$ if $|\hat{x} - \hat{y}| \ge \epsilon$,

$$\min_{|\hat{y}-\hat{x}|\,<\,\epsilon} |\tilde{f}(\hat{y})| \le f(\hat{x}) \le \max_{|\hat{y}-\hat{x}|\,<\,\epsilon} |\tilde{f}(\hat{y})|, \quad \hat{x} \in S, \tag{3.12}$$

so that $\overline{G_1} \subset \Omega_e'$. Now every $\hat{y} \in S$ is an element of S_j, for some $j \in \{1,\dots,N\}$, and $\tilde{f}(\hat{y}) = f_j$ and $|\hat{y} - \hat{x}_j| \le \delta$. Thus it follows from (3.12) that, for every $\hat{x} \in S$, $f(\hat{x}) \le f_m$ for some m for which $|\hat{x}_m - \hat{x}| < \epsilon + \delta$. Now let $x = f_m\hat{x}$, $y = f_m\hat{x}_m$. Then $|x - y| \le |\hat{x} - \hat{x}_m| < \epsilon + \delta$, and $\operatorname{dist}(y,B) = 2\delta$, so that

$$\operatorname{dist}(x,B) \ge \operatorname{dist}(y,B) - |x - y| \ge 2\delta - (\epsilon + \delta) > 0.$$

Thus $x \notin B$, and so $f(\hat{x})\hat{x} \notin B$. Thus $U \subset B \subset \Omega'_e$, and so $\operatorname{supp}\phi \subset U \subset \Omega'_e$.

Arguing similarly, for all $\hat{x} \in S$, $f(\hat{x}) \geq f_m$ for some m for which $|\hat{x}_m - \hat{x}| < \epsilon + \delta$. Defining $x = f_m\hat{x}$ and $y = f_m\hat{x}_m$, it holds that

$$\operatorname{dist}(x, B) \leq \operatorname{dist}(y, B) + |x - y| \leq 2\delta + \epsilon + \delta < 4\delta,$$

so that $x \in \Omega_e$ and hence $f(\hat{x})\hat{x} \in \Omega_e$. Thus, for all $\hat{x} \in S$, $s\hat{x} \in \Omega_e$ for $s > f(\hat{x})$; i.e., $\Omega'_e \subset \Omega_e$. $\square$

With this preliminary lemma we can proceed to show that Lemma 3.5 holds whenever Assumption 2 holds. In this final lemma (cf. [10, Lemma 4.11]) we use explicitly the fact that $b(\cdot, \cdot)$ is bounded. In fact, examining the definition (1.5), clearly we have that

$$|b(u, v)| \leq c\|u\|_{V_R}\|v\|_{V_R}, \quad u, v \in V_R, \tag{3.13}$$

where $c := 1 + \|\gamma\|^2\|T_R\|$ and $\|\gamma\|$ denotes the norm of the trace operator $\gamma : V_R \to H^{1/2}(\Gamma_R)$, while $\|T_R\|$ denotes the norm of T_R as a mapping from $H^{1/2}(\Gamma_R)$ to $H^{-1/2}(\Gamma_R)$.

Lemma 3.8. *Suppose that Assumption* 2 *holds, u is a solution to the distributed source scattering problem, $R > R_0$, and* $\operatorname{supp}(g) \subset \overline{D_R}$. *Then the bound* (3.10) *holds.*

Proof. Let $\tilde{V} := \{\phi|_{D_R} : \phi \in C_0^\infty(\Omega_e)\}$. Then $\tilde{V}$ is dense in V_R. Recall that the distributed source scattering problem is equivalent to (3.7). Suppose u satisfies (3.7) and choose a sequence $(u_m) \subset \tilde{V}$ such that $\|u_m - u\|_{V_R} \to 0$ as $m \to \infty$. Then $u_m = \phi_m|_{D_R}$, with $\phi_m \in C_0^\infty(\Omega_e)$, and, by Lemma 3.7, there exists $f_m \in C^\infty(S, \mathbb{R})$ with $\min f > 0$ such that $\operatorname{supp}\phi_m \subset \Omega_e^{(m)}$ and $\overline{G_R} \subset \Omega_e^{(m)} \subset \Omega_e$, where $\Omega_e^{(m)} := \{sf_m(\hat{x}) \in \mathbb{R}^n : \hat{x} \in S, s > 1\}$. Let $V_R^{(m)}$ and b_m denote the space and sesquilinear form corresponding to the domain $\Omega_e^{(m)}$. That is, where $D_R^{(m)} := \Omega_e^{(m)} \setminus \overline{G_R}$, $V_R^{(m)}$ is defined by $V_R^{(m)} := \{\phi|_{D_R^{(m)}} : \phi \in H_0^1(\Omega_e^{(m)})\}$ and b_m is given by (1.5) with D_R and V_R replaced by $D_R^{(m)}$ and $V_R^{(m)}$, respectively. Then $D_R^{(m)} \subset D_R$ and, if $v_m \in V_R^{(m)}$ and v denotes v_m extended by zero from $D_R^{(m)}$ to D_R, it holds that $v \in V_R$. Via this extension by zero, we can regard $V_R^{(m)}$ as a subspace of V_R and regard u_m as an element of $V_R^{(m)}$.

For all $v \in V_R^{(m)} \subset V_R$, we have

$$b_m(u_m, v) = b(u_m, v) = -(g, v) - b(u - u_m, v).$$

Let u'_m and $u''_m \in V_R^{(m)}$ be the unique solutions of

$$b_m(u'_m, v) = -(g, v), \quad b_m(u''_m, v) = -b(u - u_m, v) \quad \forall v \in V_R^{(m)}.$$

Clearly $u_m = u'_m + u''_m$ and, by Lemma 3.5, $\|u'_m\|_{V_R^{(m)}} \leq k^{-1}\tilde{C}\|g\|_2$, where $\tilde{C} = n - 1 + 2\sqrt{2}kR$, while, by (3.13), Lemma 3.3, and Corollary 3.6,

$$\|u''_m\|_{V_R^{(m)}} \leq c(1 + 2\tilde{C})\|u - u_m\|_{V_R}.$$

Thus $\|u\|_{V_R} = \lim_{m\to\infty} \|u_m\|_{V_R^{(m)}} \leq k^{-1}\tilde{C}\|g\|_2$. $\square$

Combining Lemmas 3.3, 3.4, and 3.8, we obtain the following generalization of Corollary 3.6, which is our main lower bound on the inf-sup constant and the main result of this section.

COROLLARY 3.9. *If Assumption 2 is satisfied, then the inf-sup condition* (1.8) *holds with* $\alpha^{-1} \le 1 + 2(n-1+2\sqrt{2}kR) \le 5 + 4\sqrt{2}kR$.

We finish the section by obtaining two upper bounds on the inf-sup constant, which will show, among other things, that the above bound is sharp in its dependence on k in the limit $k \to \infty$.

To obtain these bounds we modify the construction of Ihlenburg [25] for a weak formulation of a 1D Helmholtz problem. We note first that, for every nonzero $w \in V_R$,

$$\alpha \le \sup_{0 \ne v \in V_R} \frac{|b(w,v)|}{\|w\|_{V_R}\|v\|_{V_R}}.$$

Now choose $w \in V_R \cap H^2(D_R)$ such that $w = \nabla w = 0$ on Γ_R. Then, integrating by parts, for $v \in V_R$,

$$|b(w,v)| = \left| \int_{D_R} \left(\nabla w \cdot \nabla \bar{v} - k^2 w \bar{v} \right) dx \right| = \left| \int_{D_R} \left(\Delta w + k^2 w \right) \bar{v}\, dx \right|.$$

Thus, and recalling the definition (1.14),

$$\frac{|b(w,v)|}{\|w\|_{V_R}\|v\|_{V_R}} \le \frac{\|\Delta w + k^2 w\|_2 \|v\|_2}{\|w\|_{V_R}\|v\|_{V_R}} \le \frac{\|\Delta w + k^2 w\|_2}{k^2\|w\|_2}. \tag{3.14}$$

Now define $u(x) = \mathrm{e}^{\mathrm{i}kx_1} w(x)$. Then the above bound holds with w replaced by u, and

$$\Delta u(x) + k^2 u(x) = \left(2\mathrm{i}k \frac{\partial w(x)}{\partial x_1} + \Delta w(x) \right) \mathrm{e}^{\mathrm{i}kx_1}$$

so that

$$\frac{|b(u,v)|}{\|u\|_{V_R}\|v\|_{V_R}} \le \frac{\|\Delta u + k^2 u\|_2}{k^2\|u\|_2} = \frac{\|2\mathrm{i}k \frac{\partial w}{\partial x_1} + \Delta w\|_2}{k^2\|w\|_2}.$$

We have shown most of the following lemma.

LEMMA 3.10. *Suppose* $w \in V_R \cap H^2(D_R)$ *is such that* $\gamma w = \gamma \nabla w = 0$ *and* w *is nonzero. Then the inf-sup constant* (1.8) *is bounded above by*

$$\alpha \le \frac{C_1}{kR} + \frac{C_2}{k^2R^2},$$

where $C_1 := 2R\|\frac{\partial w}{\partial x_1}\|_2/\|w\|_2$, $C_2 := R^2\|\Delta w\|_2/\|w\|_2$, *and* $C_1 \ge 2\sqrt{2} \approx 2.83$.

Proof. It remains only to show the last inequality. Since $\gamma w = 0$, we can approximate w in the $H^1(D_R)$ norm arbitrarily closely by $\tilde{w} \in C_0^\infty(D_R)$. Then $C_1 \ge 2\sqrt{2}$ follows by a standard Friedrichs inequality (e.g. [10, Lemma 3.4]) which gives that $\|\tilde{w}\|_2 \le (R/\sqrt{2})\|\frac{\partial \tilde{w}}{\partial x_1}\|_2$. ∎

We note that in the case that Assumption 2 holds so that Corollary 3.9 applies, we have both upper and lower bounds on the inf-sup constant, namely (1.17), where C_1 and C_2 are as defined in Lemma 3.10.

The left-hand bound in (1.17) holds for every domain Ω_e satisfying Assumption 2. To check its sharpness, let us consider the case when $\Omega = \{0\}$ and $D = \mathbb{R}^n \setminus \{0\}$. In this special case, $V_R = H^1(D_R)$, and the solution of the plane wave scattering problem is just $u = u^i$; i.e., the scattered field is zero. Taking in this case $w(x) = F(|x|/R)$, where $F(t) := (1-t^2)^2$, we calculate that

$$C_1 = 2\sqrt{\frac{\int_0^1 (F'(t))^2 t^{n-1} dt}{n\int_0^1 (F(t))^2 t^{n-1} dt}} = \begin{cases} 2\sqrt{30}/3 \approx 3.65, & n = 2, \\ 2\sqrt{33}/3 \approx 3.83, & n = 3. \end{cases}$$

Thus, defining $c_- := (4\sqrt{2})^{-1}$, for this example the bounds (1.17) bracket $kR\alpha$ fairly tightly, predicting that, in the limit $kR \to \infty$, $kR\alpha$ is in the range $[c_-, C_1]$ with $C_1/c_- \le 8\sqrt{66}/3 \approx 21.7$.

The above results show that $k\alpha$ is bounded above in the limit $k \to \infty$ and is also bounded below if Assumption 2 holds. If Assumption 2 does not hold, then α may not be bounded below by a multiple of k^{-1}. The following example shows this behavior. It is convenient in this example to write $x \in \mathbb{R}^n$ as $x = (\tilde{x}, x_n)$, where $\tilde{x} = (x_1, \dots, x_{n-1})$.

Choose $A > 0$ and let $\Omega := \Omega_+ \cup \Omega_-$, where $\Omega_\pm := \{x \in \mathbb{R}^n : x_n = \pm A,\ |\tilde{x}| \le A\}$, so that Ω consists of two parallel lines of length $2A$ distance $2A$ apart in the 2D case, and two parallel disks of radius A in the 3D case. Choose $R > R_0 = \sqrt{2}A$ and define the function w by

$$w(x) := \begin{cases} \cos(kx_n)F(|\tilde{x}|/A), & |\tilde{x}| \le A,\ |x_n| \le A, \\ 0 & \text{otherwise,} \end{cases}$$

and suppose that $k \in \Lambda := \{(m + 1/2)\pi/A : m \in \mathbb{N}\}$. Then $w \in V_R \cap H^2(D_R)$ and $w = \nabla w = 0$ on Γ_R, so that (3.14) holds. Further,

$$\Delta w(x) + k^2 w(x) = \begin{cases} A^{-2}\cos(kx_n)\tilde{F}(|\tilde{x}|/A), & |\tilde{x}| \le A,\ |x_n| \le A, \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{F}(t) := F''(t) + (n-2)F'(t)/t$. Thus, for $k \in \Lambda$, the inf-sup constant is bounded above by

$$\alpha \le \frac{\|\Delta w + k^2 w\|_2}{k^2\|w\|_2} = \frac{C_n^*}{k^2A^2}, \tag{3.15}$$

where $C_n^* := \sqrt{\int_0^1 \tilde{F}^2(t)t^{n-2}dt / \int_0^1 F^2(t)t^{n-2}dt}$.

4. Integral equation formulations. In this section we will obtain estimates explicit in the wave number for integral equation formulations of scattering problems, focusing on the plane wave scattering problem introduced in section 3 and on the integral equation (1.7) and its adjoint.

Throughout this section we assume, essential to the integral equation method, a degree of regularity of the domain, namely that Ω_e is Lipschitz (which implies that the interior of Ω, denoted $\Omega_i = \mathbb{R}^n \setminus \bar{\Omega}_e$ is also Lipschitz). We note that the invertibility of the integral equation (1.7) and its adjoint for the general Lipschitz case has recently been established in [9, section 2], by combining known results for layer-potentials on Lipschitz domains in [41, 20, 36, 34].

Given a domain G, let $H^1(G;\Delta) := \{v \in H^1(G) : \Delta v \in L^2(G)\}$ (Δ being the Laplacian in a weak sense). This is a Hilbert space with the norm $\|v\|_{H^1(G;\Delta)} := \{\int_G [|v|^2 + |\nabla v|^2 + |\Delta v|^2]dx\}^{1/2}$. If G is Lipschitz, then there is a well-defined normal derivative operator [34], the unique bounded linear operator $\partial_\nu : H^1(G;\Delta) \to H^{-1/2}(\partial G)$ which satisfies

$$\partial_\nu v = \frac{\partial v}{\partial \nu} := \nu \cdot \nabla v$$

almost everywhere on Γ, when $v \in C^\infty(\bar{G})$.

Our integral equation formulations will be based on standard acoustic layer potentials and their normal derivatives. In the case when the domain Ω_e is Lipschitz, for $\varphi \in L^2(\Gamma)$ we define the *single-layer potential operator* by

$$S\varphi(x) \;:=\; 2\int_\Gamma \Phi(x,y)\varphi(y)\,ds(y), \quad x \in \Gamma, \tag{4.1}$$

and the *double-layer potential operator* by

$$K\varphi(x) \;:=\; 2\int_\Gamma \frac{\partial\Phi(x,y)}{\partial\nu(y)}\varphi(y)\,ds(y), \quad x \in \Gamma, \tag{4.2}$$

where the normal ν is directed into Ω_e. We also define the operator K', which arises from taking the normal derivative of the single-layer potential, by

$$K'\varphi(x) = 2\int_\Gamma \frac{\partial\Phi(x,y)}{\partial\nu(x)}\varphi(y)\,ds(y), \quad x \in \Gamma. \tag{4.3}$$

We note that the right-hand sides of (4.1)–(4.3) are well defined at least for almost all $x \in \Gamma$, where (4.1) is understood in a Lebesgue sense (that $S\varphi(x)$ is well defined in this sense for almost all $x \in \Gamma$ follows from Young's inequality), while the double-layer potential and $K'\varphi$ must be understood as Cauchy principal values [36]. Further, all three operators are bounded operators on $L^2(\Gamma)$ [36]. In fact [34], it holds that, for $|s| \le 1/2$,

$$\begin{aligned} S &: H^{s-1/2}(\Gamma) \to H^{s+1/2}(\Gamma),\\ K &: H^{s+1/2}(\Gamma) \to H^{s+1/2}(\Gamma),\\ K' &: H^{s-1/2}(\Gamma) \to H^{s-1/2}(\Gamma), \end{aligned}$$

and these mappings are bounded.

These operators can also be characterized as traces on Γ of single- and double-layer potentials defined in Ω_e and Ω_i. Introducing, temporarily, the notation $\partial_\nu^\pm$ with ∂_ν^+ and ∂_ν^- denoting the exterior and interior normal derivative operators, mapping, respectively, $H^1(\Omega_e;\Delta)$ and $H^1(\Omega_i;\Delta)$ to $H^{-1/2}(\Gamma)$, it holds that [34]

$$K'\varphi = \left(\partial_\nu^+\mathcal{S} + \partial_\nu^-\mathcal{S}\right)\varphi, \quad \varphi \in H^{-1/2}(\Gamma), \tag{4.4}$$

where $\mathcal{S}$ is defined by

$$\mathcal{S}\varphi(x) \;:=\; \int_\Gamma \Phi(x,y)\varphi(y)\,ds(y), \quad x \in \mathbb{R}^n. \tag{4.5}$$

It is shown in, e.g., McLean [34] that $\mathcal{S} : H^{-1/2}(\Gamma) \to H^{1,\mathrm{loc}}(\mathbb{R}^n)$, and clearly $(\Delta + k^2)\mathcal{S}\varphi = 0$ in $\mathbb{R}^n \setminus \Gamma$, so that the right-hand side of (4.5) defines a bounded operator on $H^{-1/2}(\Gamma)$.

From [15, Theorem 3.12] and [34, Theorems 7.15, 9.6] it follows that, if u satisfies the plane wave scattering problem, then a form of Green's representation theorem holds, namely

$$u(x) = u^i(x) - \int_\Gamma \Phi(x,y)\partial_\nu u(y)\,ds(y), \quad x \in \Omega_e. \tag{4.6}$$

Two integral equations for $\partial_\nu u$ can be obtained by taking the trace and the normal derivative, respectively, of (4.6) on Γ, namely $0 = u^i - \mathcal{S}\partial_\nu u$ and $\partial_\nu u = \partial_\nu u^i - \partial_\nu \mathcal{S}\partial_\nu u$. Note that, to simplify the notation, we have not explicitly used the trace operator γ in these equations or later in this section. Its presence is assumed implicitly. Since [34] we have the jump relations that on Γ we have $2\mathcal{S}\varphi = S\varphi$ and $2\partial_\nu \mathcal{S}\varphi = -\varphi + K'\varphi$, for $\varphi \in H^{-1/2}(\Gamma)$, we can write these equations as

$$S\partial_\nu u = 2u^i, \quad \partial_\nu u + K'\partial_\nu u = 2\partial_\nu u^i.$$

It is well known (e.g., [14]) that each of these integral equations fails to be uniquely solvable if $-k^2$ is an eigenvalue of the Laplacian in Ω for, respectively, Dirichlet and Neumann boundary conditions, but that a uniquely solvable integral equation is obtained by taking an appropriate linear combination of the above equations. Clearly, for every $\eta \in \mathbb{R}$ it follows from the above equations that

$$A'\partial_\nu u = f, \tag{4.7}$$

where

$$A' := I + K' - \mathrm{i}\eta S,$$

I is the identity operator, and

$$f(x) := 2\frac{\partial u^i}{\partial \nu}(x) - 2\mathrm{i}\eta u^i(x), \quad x \in \Gamma.$$

We have shown the first part of the following theorem, which is standard (e.g., [14]) in the case when Γ is smooth; for the extension to the case of Lipschitz Γ see [9].

THEOREM 4.1. *If u satisfies the plane wave scattering problem, then, for every $\eta \in \mathbb{R}$, $\partial_\nu u \in H^{-1/2}(\Gamma)$ satisfies the integral equation* (4.7). *Conversely, if $\phi \in H^{-1/2}(\Gamma)$ satisfies $A'\phi = f$, for some $\eta \in \mathbb{R} \setminus \{0\}$, and u is defined in Ω_e by* (4.6), *with $\partial_\nu u$ replaced by ϕ, then u satisfies the plane wave scattering problem and $\partial_\nu u = \phi$.*

Note that, since we know that the plane wave scattering problem is uniquely solvable, this theorem implies that the integral equation (4.7) has exactly one solution in $H^{-1/2}(\Gamma)$.

The integral equation (4.7) is an example of a so-called *direct* integral equation formulation, obtained by applying Green's theorem to the original scattering problem. A related, *indirect* integral equation formulation, dating back to Brakhage and Werner [4], Leis [33], and Panich [40], is obtained by looking for a solution to the scattering problem in the form (1.6) for some density $\varphi \in H^{1/2}(\Gamma)$ and some $\eta \in \mathbb{R} \setminus \{0\}$. This combined single- and double-layer potential is in $C^2(\Omega_e)$, satisfies the Helmholtz equation and Sommerfeld radiation condition, and is in $H^{1,\mathrm{loc}}(\Omega_e)$ [34]. Thus it satisfies the plane wave scattering problem if and only if it satisfies the boundary condition that $u^s = -u^i$ on Γ. Using the standard jump relations for Lipschitz domains [34], we see that this holds if and only if the integral equation (1.7) is satisfied, i.e., if and only if

$$A\varphi = 2g,$$

where $g(x) := -u^i(x)$, $x \in \Gamma$, is the required Dirichlet data on Γ, and $A := I + K - \mathrm{i}\eta S$. This is the integral equation formulation introduced in [4, 33, 40].

Note that the above mapping properties of S, K, and K' imply that, for $|s| \le 1/2$,

$$A : H^{s+1/2}(\Gamma) \to H^{s+1/2}(\Gamma), \quad A' : H^{s-1/2}(\Gamma) \to H^{s-1/2}(\Gamma), \tag{4.8}$$

and these mappings are bounded. It is shown, moreover, in [9] (or see [37] for the case A' and $s = 0$), by combining the standard arguments for these integral equations when Γ is smooth (see, e.g., [14]) with known properties of integral operators on Lipschitz domains [41, 20, 36, 34], that, for $\eta \in \mathbb{R} \setminus \{0\}$, these mappings are bijections, which of course implies that their inverses are bounded by the Banach theorem. Further [36], K' is the adjoint of K, and S is self-adjoint, so that A' is the adjoint of A in the same sense, namely that

$$(\phi, A\psi)_\Gamma = (A'\phi, \psi)_\Gamma \quad \text{for } \phi \in L^2(\Gamma),\ \psi \in L^2(\Gamma), \tag{4.9}$$

where $(\phi, \psi)_\Gamma := \int_\Gamma \phi\psi ds$. Since $H^1(\Gamma)$ is dense in $H^{-1}(\Gamma)$ and the mappings (4.8) are bounded, it follows by density that the duality relation (4.9) holds, more generally, for $\phi \in H^{-s-1/2}(\Gamma)$ and $\psi \in H^{s+1/2}(\Gamma)$, provided that $|s| \le 1/2$. This implies that the norms of A and A^{-1} as operators on $H^{s+1/2}(\Gamma)$ coincide with those of A' and ${A'}^{-1}$, respectively, as operators on $H^{-s-1/2}(\Gamma)$ for $|s| \le 1/2$. In particular, we note that

$$\|A^{-1}\|_2 = \|{A'}^{-1}\|_2, \tag{4.10}$$

where, here and in the remainder of the paper, $\|\cdot\|_2$ denotes both the norm on $L^2(\Gamma) = H^0(\Gamma)$ and the induced norm on the space of bounded linear operators on $L^2(\Gamma)$.

Following this preparation, we show now the main result of this section, which is an explicit bound on $\|A^{-1}\|_2 = \|{A'}^{-1}\|_2$ in terms of the geometry of Γ and the wave number, in the case when Ω is starlike and Lipschitz. For brevity and to simplify the arguments somewhat we also assume that Γ is piecewise smooth. Precisely, we make the following assumption, which is intermediate between Assumptions 1 and 2 introduced in section 3.

Assumption 3. *For some $f \in C^{0,1}(S, \mathbb{R})$ with $\min_{\hat{x} \in S} f(\hat{x}) > 0$, it holds that $\Gamma = \{f(\hat{x})\hat{x} : \hat{x} \in S\}$. Further, for some $M \in \mathbb{N}$, it holds that $S = \overline{\cup_{j=1}^M S_j}$, with each S_j open in S, $S_{sing} := S \setminus \cup_{j=1}^M S_j$ a set of zero (surface) measure, and $f|_{S_j} \in C^2(S_j, \mathbb{R})$ for $j = 1, \ldots, M$.*

Remark 4.2. As an important example, we note that Assumption 3 is satisfied if Γ is a polyhedron, provided that the interior of Ω, $\Omega_i = \mathbb{R}^n \setminus \bar{\Omega}_e$, is starlike with respect to the origin; i.e., $x \in \Omega_i$ implies $sx \in \Omega_i$ for $0 \le s \le 1$. Explicitly the function f is then defined by $f(\hat{x}) := \max\{s > 0 : s\hat{x} \in \Omega\}$, and, if $\Gamma_1, \ldots, \Gamma_M$ denote the sides of Γ (each Γ_j open in Γ) and $\Gamma_{sing} := \Gamma \setminus \cup_{j=1}^M \Gamma_j$ denote the edges and corners of Γ, then Assumption 3 holds with $S_j := f^{-1}(\Gamma_j)$, $j = 1, \ldots, M$ and $\Gamma_{sing} = f(S_{sing})$.

Note that, if Assumption 3 holds (and, more generally, whenever Γ is piecewise smooth), the integrals (4.2) and (4.3) are well defined in the ordinary Lebesgue sense almost everywhere on Γ, in fact, provided that $x \notin \Gamma_{sing} = f(S_{sing})$. Note also that if Assumption 3 holds, then $0 < \delta_- \le \delta_+ \le R_0$, where

$$\delta_- := \inf_{x \in \Gamma \setminus \Gamma_{sing}} (x \cdot \nu), \quad \delta_+ := \sup_{x \in \Gamma \setminus \Gamma_{sing}} (x \cdot \nu),$$

and $R_0 = \max_{x \in \Gamma} |x|$. Let us also define

$$\delta^* := \sup_{x \in \Gamma \setminus \Gamma_{sing}} |x - (x \cdot \nu)\nu| \le R_0.$$

The main result of this section is the following theorem. We postpone the proof until the end of the section.

THEOREM 4.3. *Suppose that Assumption* 3 *holds and* $\eta \in \mathbb{R} \setminus \{0\}$. *Then*

$$\|A^{-1}\|_2 = \|A'^{-1}\|_2 \leq B, \tag{4.11}$$

where

$$B := \frac{1}{2} + \left[\left(\frac{\delta_+}{\delta_-} + \frac{4\delta^{*2}}{\delta_-^2}\right)\left[\frac{\delta_+}{\delta_-}\left(\frac{k^2}{\eta^2}+1\right) + \frac{n-2}{\delta_-|\eta|} + \frac{\delta^{*2}}{\delta_-^2}\right] + \frac{(1+2kR_0)^2}{2\delta_-^2\eta^2}\right]^{1/2}.$$

To help make the expression for B more comprehensible, let us consider some examples. Suppose first that Γ is a circle or sphere; i.e., $\Gamma = \{x : |x| = R_0\}$. Then $\delta_- = \delta_+ = R_0$ and $\delta^* = 0$, and thus

$$B = \frac{1}{2} + \left[1 + \frac{k^2}{\eta^2} + \frac{n-2}{R_0|\eta|} + \frac{(1+2kR_0)^2}{2R_0^2\eta^2}\right]^{1/2}. \tag{4.12}$$

In the 2D case that Γ is a regular polygon (centered on the origin) with M sides, $\delta_- = \delta_+ = R_0\cos(\pi/M)$ and $\delta^* = R_0\sin(\pi/M)$; thus

$$B = \frac{1}{2} + \left[\left(1 + 4\tan^2\frac{\pi}{M}\right)\left[1 + \frac{k^2}{\eta^2} + \tan^2\frac{\pi}{M}\right] + \frac{(1+2kR_0)^2}{2R_0^2\eta^2\cos^2(\pi/M)}\right]^{1/2}. \tag{4.13}$$

In the limit $M \to \infty$ this recovers (4.12), and for a square ($M = 4$) this simplifies to

$$B = \frac{1}{2} + \left[10 + 5\frac{k^2}{\eta^2} + \frac{(1+2kR_0)^2}{R_0^2\eta^2}\right]^{1/2}. \tag{4.14}$$

Similarly, for the cube $\Omega = \{x : |x_j| \leq a,\ j = 1,2,3\}$ of side-length $2a$ we have $\delta_- = \delta_+ = a$, $\delta^* = \sqrt{2}a$, and $R_0 = \sqrt{3}a$, so that

$$B = \frac{1}{2} + 3\left[3 + \frac{k^2}{\eta^2} + \frac{1}{a|\eta|} + \frac{(1+2\sqrt{3}ka)^2}{18a^2\eta^2}\right]^{1/2}.$$

We note that (4.12) can be compared with the results of Dominguez, Graham, and Smyshlyaev [19], who have shown, when Γ is a circle, the bound (1.9) for all sufficiently large k if the choice $\eta = k$ is made. Our results Theorem 4.3 and (4.12) predict for the circle that, if we choose $\eta = k$,

$$\|A^{-1}\|_2 \leq \frac{1}{2} + \left[2 + \frac{(1+2kR_0)^2}{2k^2R_0^2}\right]^{1/2}. \tag{4.15}$$

The right-hand side of this equation is a decreasing function of kR_0 on $(0,\infty)$ which approaches the limit 2.5 as $kR_0 \to \infty$. Thus our results show for a circle that, for every $\theta > 2.5$, $\|A^{-1}\|_2 \leq \theta$ for all sufficiently large kR_0. This bound is close to the result of [19] although we use much more general methods than the explicit calculation of eigenfunctions and eigenvectors used in [19], which are only available for a circular geometry. On the other hand, the authors also show, importantly, the coercivity (1.10), which our methods do not seem to be well adapted to obtain.

If we follow Dominguez, Graham, and Smyshlyaev [19] and choose $\eta = k$, we obtain the following simplification of the bound in Theorem 4.3 for the case $kR_0 \geq 1$. To obtain the second inequality we use that $\delta_+/\delta_- \leq \theta$ and $\delta^*/\delta_- \leq \theta$.

COROLLARY 4.4. *If Assumption* 3 *holds, $\eta = k$, and $kR_0 \geq 1$, then*

$$\|A^{-1}\|_2 = \|A'^{-1}\|_2 \leq \frac{1}{2} + \left[\left(\frac{\delta_+}{\delta_-} + \frac{4\delta^{*2}}{\delta_-^2}\right)\left[2\frac{\delta_+}{\delta_-} + \frac{(n-2)R_0}{\delta_-} + \frac{\delta^{*2}}{\delta_-^2}\right] + \frac{9R_0^2}{2\delta_-^2}\right]^{1/2}$$
$$\leq \frac{1}{2}\left(1 + \theta(4\theta + 4n + 1)\right),$$

where $\theta := R_0/\delta_-$.

We finish the section by providing a proof of Theorem 4.3. Clearly, given that we already know that A and A' are invertible as operators on $L^2(\Gamma)$ and we have (4.10), this theorem is implied as a corollary of the following lemma (cf. [8, Lemma 3.3]).

LEMMA 4.5. *Suppose that Assumption* 3 *holds and $\eta \in \mathbb{R} \setminus \{0\}$. Then, for all $\varphi \in L^2(\Gamma)$,*

$$||A'\varphi||_2 \geq B^{-1}||\varphi||_2. \tag{4.16}$$

Proof. Let $Y \subset L^2(\Gamma)$ denote the set of those functions φ that are Hölder continuous and are supported in $\Gamma \setminus \Gamma_{sing}$. Since Y is dense in $L^2(\Gamma)$ and A' is bounded on $L^2(\Gamma)$, it is sufficient to show that (4.16) holds for all $\varphi \in Y$.

Thus suppose $\varphi \in Y$, and define the single-layer potential u by

$$u(x) := \int_\Gamma \Phi(x,y)\varphi(y)ds(y) = \int_{\tilde{\Gamma}} \Phi(x,y)\varphi(y)ds(y), \quad x \in \mathbb{R}^n,$$

where $\tilde{\Gamma} \subset \Gamma \setminus \Gamma_{sing}$ is the support of φ. From standard properties of the single-layer potential (e.g., [14]) we have that $u \in C(\mathbb{R}^3) \cap C^2(\mathbb{R}^n \setminus \tilde{\Gamma})$. Further, it follows from [14, Theorem 2.17] that ∇u can be continuously extended from Ω_e to $\bar{\Omega}_e$ and from Ω_i to Ω, with limiting values on Γ given by

$$\nabla u_\pm(x) = \int_{\tilde{\Gamma}} \nabla_x \Phi(x,y)\varphi(y)ds(y) \mp \frac{1}{2}\varphi(x)\nu(x), \quad x \in \Gamma, \tag{4.17}$$

where, as before, $\nu(x)$ is the unit normal vector at x, directed into Ω_e, and

$$\nabla u_\pm(x) := \lim_{\epsilon \to 0+} \nabla u(x \pm \epsilon\nu(x)), \quad x \in \Gamma.$$

We note from (4.17) that the tangential part of ∇u, $\nabla_T u$, is continuous across Γ. On the other hand, the normal derivative jumps across Γ, with

$$\frac{\partial u_\pm}{\partial \nu}(x) = \frac{1}{2}\left[K'\varphi(x) \mp \varphi(x)\right], \quad x \in \Gamma \setminus \Gamma_{sing}. \tag{4.18}$$

Since also $u(x) = \frac{1}{2}S\varphi(x)$, $x \in \Gamma$, defining

$$g := \frac{1}{2}A'\varphi = \frac{1}{2}(I + K' - \mathrm{i}\eta S)\varphi,$$

we see that

$$\frac{\partial u_-}{\partial \nu}(x) - \mathrm{i}\eta u(x) = g(x), \quad x \in \Gamma \setminus \Gamma_{sing}. \tag{4.19}$$

Further, from (4.18) we see that

$$\frac{\partial u_-}{\partial \nu}(x) - \frac{\partial u_+}{\partial \nu}(x) = \varphi(x), \quad x \in \Gamma \setminus \Gamma_{sing}. \tag{4.20}$$

Note that to complete the proof we have to show that

$$||\varphi||_2 \leq 2B||g||_2. \tag{4.21}$$

We will achieve this by bounding the normal derivatives of u on Γ via applications of Lemma 2.3 in Ω_i and in D_R, for some $R > R_0$.

Before proceeding we note first that (2.19) and (2.20) do hold with v replaced by u and $G = \Omega_i$ or $G = D_R$, although we have not shown that $u \in H^2(G)$ so that we cannot apply Lemma 2.3 directly. To derive these equations when $G = \Omega_i$, we can first apply Lemma 2.3 with $v = u$ and $G = s\Omega_i$, for $s \in (0, 1)$, and then take the limit $s \to 1^-$, noting that $\Delta u + k^2 u = 0$ in Ω_i and $u \in C^1(\Omega)$. Arguing similarly, these equations also hold with v replaced by u and $G = D_R$. Thus, recalling that our normal vector ν on Γ points out of Ω_e, we have, taking the imaginary part of (2.19) with $v = u$ and $G = \Omega_i$ or $G = D_R$, the identities

$$\Im \int_\Gamma \bar{u} \, \frac{\partial u_-}{\partial \nu} \, ds = 0, \tag{4.22}$$

$$\Im \int_\Gamma \bar{u} \, \frac{\partial u_+}{\partial \nu} \, ds = \Im \int_{\Gamma_R} \bar{u} \, \frac{\partial u}{\partial r} \, ds. \tag{4.23}$$

Taking $v = u$ and $G = \Omega_i$ and adding (2.20) to $(n-2)$ times the real part of (2.19) gives

$$\begin{aligned} 2k^2 \int_\Omega |u|^2 \, dx = \int_\Gamma \Bigg(& x \cdot \nu \left(k^2|u|^2 + \left| \frac{\partial u_-}{\partial \nu} \right|^2 - |\nabla_T u|^2 \right) \\ & + \Re \left([(n-2)\bar{u} + 2x \cdot \nabla_T \bar{u}] \, \frac{\partial u_-}{\partial \nu} \right) \Bigg) \, ds. \end{aligned} \tag{4.24}$$

Finally, taking $v = u$ and $G = D_R$, for some $R > R_0$, and adding (2.20) to the real part of (2.19), we have

$$\begin{aligned} & \int_{D_R} \left((3-n)|\nabla u|^2 + (n-1)k^2|u|^2 \right) \, dx \\ &= -\int_\Gamma \left(x \cdot \nu \left(k^2|u|^2 + \left| \frac{\partial u_+}{\partial \nu} \right|^2 - |\nabla_T u|^2 \right) + \Re \left([\bar{u} + 2x \cdot \nabla_T \bar{u}] \, \frac{\partial u_+}{\partial \nu} \right) \right) \, ds \\ & \quad + \int_{\Gamma_R} \left(\left(k^2|u|^2 + \left| \frac{\partial u}{\partial r} \right|^2 - |\nabla_T u|^2 \right) + \Re \left(\bar{u} \, \frac{\partial u}{\partial r} \right) \right) \, ds. \end{aligned} \tag{4.25}$$

Using these four identities and Lemma 2.1 we will complete the proof.

We start by using (4.19) to replace $\partial u_- / \partial \nu$ in (4.22). Applying Cauchy–Schwarz, we see that

$$|\eta| \, \|u\|_2^2 = \left| \Im \int_\Gamma \bar{u} g ds \right| \leq \|u\|_2 \|g\|_2,$$

so that

$$\|u\|_2 \le |\eta|^{-1}\|g\|_2. \tag{4.26}$$

Alternatively, from (4.22) we have that

$$\Re \int_\Gamma \mathrm{i}\eta\bar{u}\frac{\partial u_-}{\partial \nu}\,ds = 0,$$

and, using (4.19) and Cauchy–Schwarz, we see that

$$\left\|\frac{\partial u_-}{\partial \nu}\right\|_2 \le \|g\|_2. \tag{4.27}$$

It remains to bound the L^2 norm of $\partial u_+/\partial\nu$ in terms of $\|g\|_2$. To achieve this goal we first bound $\|\nabla_T\|_2$ using (4.24). From this equation we have that

$$\begin{aligned}\delta_-\|\nabla_T u\|_2^2 &\le \int_\Gamma x\cdot\nu|\nabla_T u|^2\,ds\\ &\le \delta_+k^2\|u\|_2^2 + \delta_+\left\|\frac{\partial u_-}{\partial \nu}\right\|_2^2 + [(n-2)\|u\|_2 + 2\delta^*\|\nabla_T u\|_2]\left\|\frac{\partial u_-}{\partial \nu}\right\|_2,\end{aligned}$$

where we have used that $|x\cdot\nabla_T u| = |(x-(x\cdot\nu)\nu)\cdot\nabla_T u|$. From (4.26) and (4.27) it follows that

$$\delta_-\|\nabla_T u\|_2^2 \le \left[\delta_+\left(\frac{k^2}{\eta^2}+1\right) + \frac{n-2}{|\eta|}\right]\|g\|_2^2 + 2\delta^*\|\nabla_T u\|_2\|g\|_2.$$

Finally, applying (3.11) to the last term on the right-hand side, we deduce that

$$\frac{\delta_-}{2}\|\nabla_T u\|_2^2 \le \left[\delta_+\left(\frac{k^2}{\eta^2}+1\right) + \frac{n-2}{|\eta|} + 2\frac{\delta^{*2}}{\delta_-}\right]\|g\|_2^2$$

so that

$$\|\nabla_T u\|_2 \le \left[2\frac{\delta_+}{\delta_-}\left(\frac{k^2}{\eta^2}+1\right) + \frac{2(n-2)}{\delta_-|\eta|} + 2\frac{\delta^{*2}}{\delta_-^2}\right]^{1/2}\|g\|_2. \tag{4.28}$$

To finish the proof, we start from (4.25), apply Lemma 2.1, which is valid since u is a radiating solution of the Helmholtz equation, and then use (4.23) to see that

$$\delta_-\left\|\frac{\partial u_+}{\partial \nu}\right\|_2^2 \le \int_\Gamma x\cdot\nu\left|\frac{\partial u_+}{\partial \nu}\right|^2\,ds$$

$$\le \int_\Gamma\left(x\cdot\nu|\nabla_T u|^2 + \Re\left([\bar{u}+2x\cdot\nabla_T\bar{u}]\,\frac{\partial u_+}{\partial \nu}\right) + 2kR\Im\left(\bar{u}\frac{\partial u_+}{\partial \nu}\right)\right)ds.$$

Applying Cauchy–Schwarz and (3.11), we see that

$$\begin{aligned}\delta_-\left\|\frac{\partial u_+}{\partial \nu}\right\|_2^2 &\le \delta_+\|\nabla_T u\|_2^2 + (1+2kR)\|u\|_2\left\|\frac{\partial u_+}{\partial \nu}\right\|_2 + 2\delta^*\|\nabla_T u\|_2\left\|\frac{\partial u_+}{\partial \nu}\right\|_2\\ &\le \left(\delta_+ + \frac{4\delta^{*2}}{\delta_-}\right)\|\nabla_T u\|_2^2 + \frac{\delta_-}{2}\left\|\frac{\partial u_+}{\partial \nu}\right\|_2^2 + \frac{(1+2kR)^2}{\delta_-}\|u\|_2^2.\end{aligned}$$

Hence, and using (4.26) and (4.28),

$$\begin{aligned}\left\|\frac{\partial u_+}{\partial \nu}\right\|_2^2 &\le 2\left(\frac{\delta_+}{\delta_-}+\frac{4\delta^{*2}}{\delta_-^2}\right)\|\nabla_T u\|_2^2 + \frac{2(1+2kR)^2}{\delta_-^2}\|u\|_2^2 \\ &\le 4\left[\left(\frac{\delta_+}{\delta_-}+\frac{4\delta^{*2}}{\delta_-^2}\right)\left[\frac{\delta_+}{\delta_-}\left(\frac{k^2}{\eta^2}+1\right)+\frac{(n-2)}{\delta_-|\eta|}+\frac{\delta^{*2}}{\delta_-^2}\right]+\frac{(1+2kR)^2}{2\delta_-^2\eta^2}\right]\|g\|_2^2.\end{aligned}$$

This bound holds for all $R > R_0$ and hence also for $R = R_0$. Combining this bound with (4.27) we see that we have shown (4.21) and thus have finished the proof of the lemma. ∎

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] M. Ainsworth, *Discrete dispersion relation for hp-version finite element approximation at high wave number*, SIAM J. Numer. Anal., 42 (2004), pp. 553–575.

[3] S. Amini, *On the choice of the coupling parameter in boundary integral equation formulations of the exterior acoustic problem*, Appl. Anal., 35 (1990), pp. 75–92.

[4] H. Brakhage and P. Werner, *Über das Dirichletsche Außenraumproblem für die Helmholtzsche Schwingungsgleichung*, Arch. Math., 16 (1965), pp. 325–329.

[5] O. P. Bruno, C. A. Geuzaine, J. A. Monro, Jr., and F. Reitich, *Prescribed error tolerances within fixed computational times for scattering problems of arbitrarily high frequency: The convex case*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 362 (2004), pp. 629–645.

[6] A. Buffa and S. Sauter, *On the acoustic single layer potential: Stabilization and Fourier analysis*, SIAM J. Sci. Comput., 28 (2006), pp. 1974–1999.

[7] O. Cessenat and B. Després, *Using plane waves as base functions for solving time harmonic equations with the ultra weak variational formulation*, J. Comput. Acoust., 11 (2003), pp. 227–238.

[8] S. N. Chandler-Wilde, E. Heinemeyer, and R. Potthast, *A well-posed integral equation formulation for 3D rough surface scattering*, Proc. R. Soc. Lond. Ser. A, 462 (2006), pp. 3683–3705.

[9] S. N. Chandler-Wilde and S. Langdon, *A Galerkin boundary element method for high frequency scattering by convex polygons*, SIAM J. Numer. Anal., 45 (2007), pp. 610–640.

[10] S. N. Chandler-Wilde and P. Monk, *Existence, uniqueness, and variational methods for scattering by unbounded rough surfaces*, SIAM J. Math. Anal., 37 (2005), pp. 598–618.

[11] S. N. Chandler-Wilde and B. Zhang, *Electromagnetic scattering by an inhomogenous conducting or dielectric layer on a perfectly conducting plate*, Proc. R. Soc. Lond. Ser. A, 454 (1998), pp. 519–542.

[12] S. H. Christiansen and J. C. Nédélec, *Preconditioners for the numerical solution of boundary integral equations from electromagnetism*, C.R. Acad. Sci. Paris Sér I Math., 331 (2000), pp. 733–738.

[13] X. Claeys and H. Haddar, *Scattering from infinite rough tubular surfaces*, Math. Methods Appl. Sci., 30 (2006), pp. 389–414.

[14] D. L. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.

[15] D. L. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer, Berlin, 1998.

[16] P. Cummings and X. Feng, *Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations*, Math. Models Methods Appl. Sci., 16 (2006), pp. 139–160.

[17] E. Darve and P. Havé, *A fast multipole method for Maxwell equations stable at all frequencies*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 362 (2004), pp. 603–628.

[18] L.F. Demkowicz and K. Gerdes, *Solution of 3D Laplace and Helmholtz equations in exterior domains using hp infinite element methods*, Comput. Methods Appl. Mech. Engrg., 137 (1996), pp. 239–274.

[19] V. Dominguez, I. G. Graham, and V. P. Smyshlyaev, *A hybrid numerical-asymptotic boundary integral method for high-frequency acoustic scattering*, Numer. Math., 106 (2007), pp. 471–510.

[20] J. Elschner, *The double-layer potential operator over polyhedral domains* I: *Solvability in weighted Sobolev spaces*, Appl. Anal., 45 (1992), pp. 117–134.
[21] C. Farhat, I. Harari, and U. Hetmaniuk, *A discontinuous Galerkin method with Lagrange multipliers for the solution of Helmholtz problems in the mid-frequency regime*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 1389–1419.
[22] M. Ganesh and I. G. Graham, *A high-order algorithm for obstacle scattering in three dimensions*, J. Comput. Phys., 198 (2004), pp. 211–242.
[23] K. Giebermann, *Schnelle Summationsverfahren zur numerischen Lösung von Integralgleichungen für Streuprobleme im* $\mathbb{R}^3$, Ph.D. thesis, Universität Karlsruhe, Karlsruhe, Germany, 1997.
[24] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer, Berlin, 1983.
[25] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering*, Springer, Berlin, 1998.
[26] F. Ihlenburg and I. Babuška, *Finite element solution of the Helmholtz equation with high wave number.* I. *The h-version of the FEM*, Comput. Math. Appl., 30 (1995), pp. 9–37.
[27] F. Ihlenburg and I. Babuška, *Finite element solution of the Helmholtz equation with high wave number Part* II: *The h-p version of the FEM*, SIAM J. Numer. Anal., 34 (1997), pp. 315–358.
[28] S. J. Kim and J. B. Keller, *High-frequency asymptotics for the numerical solution of the Helmholtz equation*, Appl. Math. Lett., 18 (2005), pp. 797–804.
[29] R. Kress, *Minimizing the condition number of boundary integral-operators in acoustic and electromagnetic scattering*, Quart. J. Mech. Appl. Math., 38 (1985), pp. 323–341.
[30] R. Kress and W. T. Spassov, *On the condition number of boundary integral operators for the exterior Dirichlet problem for the Helmholtz equation*, Numer. Math., 42 (1983), pp. 77–95.
[31] O. Laghrouche, P. Bettess, and R. J. Astley, *Modelling of short wave diffraction problems using approximating systems of plane waves*, Internat. J. Numer. Methods Engrg., 54 (2002), pp. 1501–1533.
[32] S. Langdon and S. N. Chandler-Wilde, *A wavenumber independent boundary element method for an acoustic scattering problem*, SIAM J. Numer. Anal., 43 (2006), pp. 2450–2477.
[33] R. Leis, *Zur Dirichtletschen Randwertaufgabe des Aussenraums der Schwingungsgleichung*, Math. Z., 90 (1965), pp. 205–211.
[34] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.
[35] J. M. Melenk, *On Generalized Finite Element Methods*, Ph.D. thesis, University of Maryland, College Park, MD, 1995.
[36] Y. Meyer and R. Coifman, *Wavelets: Calderón-Zygmund and Multilinear Operators*, Cambridge University Press, Cambridge, UK, 2000.
[37] M. Mitrea, *Boundary value problems and Hardy spaces associated to the Helmholtz equation in Lipschitz domains*, J. Math. Anal. Appl., 202 (1996), pp. 819–842.
[38] J. Necas, *Les Methodes Directes en Theorie des Equations Elliptiques*, Academia, Prague and Masson, Paris, 1967.
[39] J.-C. Nédélec, *Acoustic and Electromagnetic Equations: Integral Representations for Harmonic Problems*, Springer, New York, 2001.
[40] O. I. Panich, *On the question of the solvability of the exterior boundary-value problems for the wave equation and Maxwell's equations*, Uspekhi Mat. Nauk, 20A (1965), pp. 221–226.
[41] G. Verchota, *Layer potentials and regularity for the Dirichlet problem for Laplace's equation in Lipschitz domains*, J. Funct. Anal., 59 (1984), pp. 572–611.
[42] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, UK, 1944.

SIAM J. MATH. ANAL.
Vol. 39, No. 5, pp. 1456–1471

ON THE CAUCHY PROBLEM FOR THE p-SYSTEM AT A JUNCTION*

RINALDO M. COLOMBO† AND MAURO GARAVELLO‡

Abstract. We present a model for the description of a nonviscous isentropic or isothermal fluid crossing a junction. Aiming at an extension of the usual Euler equations, we neglect the effects of friction against the walls of the pipes, but the reaction constraints at the junction are considered. The well posedness of the Cauchy problem is proved, and some qualitative properties of the model are described.

Key words. p-system, hyperbolic systems of conservation laws, Cauchy problem, compressible fluids, gas dynamics

AMS subject classifications. 35L65, 76N10

DOI. 10.1137/060665841

1. Introduction. Consider n rectilinear tubes exiting a junction. For $l = 1, \ldots, n$, the direction and section of the lth tube are described, respectively, by the direction and the norm of a vector $\nu_l \in \mathbb{R}^3 \setminus \{0\}$. All tubes are filled with the same nonviscous isentropic (or isothermal) fluid, and friction along the walls is neglected. The resulting system can be modeled through n copies of the one-dimensional p-system in Eulerian coordinates:

$$\left\{ \begin{array}{l} \partial_t \rho_l + \partial_x q_l = 0, \\ \partial_t q_l + \partial_x \left(\dfrac{q_l^2}{\rho_l} + p(\rho_l) \right) = 0, \end{array} \right. \qquad \begin{array}{rcl} t & \in & [0, +\infty[\,, \\ x & \in & [0, +\infty[\,, \\ l & \in & \{1, \ldots, n\}\,. \end{array} \tag{1.1}$$

Here, t is time, and, along the lth tube, x is the abscissa, ρ_l is the fluid density, and q_l is the linear momentum density. The pressure law $p = p(\rho)$ is the same for all tubes.

This paper aims to study the Cauchy problem for (1.1). Therefore, we introduce below coupling conditions among the tubes that depend on the relative sizes and positions of the tubes. These conditions neglect friction at the junction; nevertheless the total linear momentum is *not* conserved due to the presence of the junction, and its variation depends on $\nu_1, \ldots, \nu_n$ (see also Example 1 below).

The key condition, written using the linear momentum flow $P(\rho, q) = \frac{q^2}{\rho} + p(\rho)$, reads

$$\lim_{x \to 0+} P\left(q_l(t,x), \rho_l(t,x)\right) = \lim_{x \to 0+} P\left(q_h(t,x), \rho_h(t,x)\right) \tag{P}$$

for all $l, h = 1, \ldots, n$, and for almost every (a.e.) $t > 0$. In the static case $q_l = 0$, it amounts to the equality of the hydrostatic pressure. Note that in (P) the geometry of the junction is hidden in the vectors ν_l that identify the direction of the tube. Indeed,

*Received by the editors July 24, 2006; accepted for publication (in revised form) July 16, 2007; published electronically January 11, 2008.
http://www.siam.org/journals/sima/39-5/66584.html
†Department of Mathematics, University of Brescia, 25123 Brescia, Italy (rinaldo@ing.unibs.it).
‡Department of Mathematics and Applications, University of Milano–Bicocca, 20125 Milano, Italy. Current address: Dipartimento di Scienze e Tecnologie Avanzate, Università del Piemonte Orientale, 15100 Alessandria, Italy (mauro.garavello@mfn.unipmn.it).

the connection between (1.1) and the full three-dimensional (3D) Euler equations is achieved through the functions

$$\rho(t;\mathbf{x}) = \sum_{l=1}^{n} \rho_l(t, \mathbf{x}\cdot\hat{\nu}_l)\,\chi_{T_l}(\mathbf{x}) \quad \text{and} \quad \mathbf{q}(t;\mathbf{x}) = \sum_{l=1}^{n} q_l(t, \mathbf{x}\cdot\hat{\nu}_l)\,\chi_{T_l}(\mathbf{x})\,\hat{\nu}_l\,,$$

where T_l is the region in the lth tube, $\mathbf{x}$ is the space variable, and $\hat{\nu}_l = \nu_l/\|\nu_l\|$. As usual, χ_T is the characteristic function of T; i.e., $\chi_T(\mathbf{x}) = 1$ if $\mathbf{x} \in T$, while $\chi_T(\mathbf{x}) = 0$ if $\mathbf{x} \notin T$. Here we approximate the intrinsically 3D nature of the real junction assuming that the fluid speed in the lth pipe is parallel to ν_l and that the different pipes interact at a unique single point. The total linear momentum is then

$$Q(t) = \int_{\mathbb{R}^3} \mathbf{q}(t,\mathbf{x})\,d\mathbf{x} = \sum_{l=1}^{n} \int_0^{+\infty} q_l(t,x)\,\nu_l\,dx.$$

Straightforward computations based on the 3D Euler equations (see [9]) show that if (ρ_l, q_l) is in $\mathbf{L^1} \cap \mathbf{BV}$, then the variation of the total linear momentum on $\bigcup_{l=1}^{n} T_l$ during the time interval $[t_1, t_2]$ is

$$\begin{aligned} Q(t_2) - Q(t_1) &= \iiint_{\bigcup_l T_l} \left(\mathbf{q}(t_2;\mathbf{x}) - \mathbf{q}(t_1;\mathbf{x})\right) d\mathbf{x} \\ &= \int_{t_1}^{t_2} \sum_{l=1}^{n} P\left(\rho_l(t,0+), q_l(t,0+)\right) \nu_l\,dt\,. \end{aligned}$$

The latter term above is the variation in the total linear momentum and is due to the junction. Condition (P) says that

$$Q(t_2) - Q(t_1) = \int_{t_1}^{t_2} \sum_{l=1}^{n} P\left(\rho_l(t,0+), q_l(t,0+)\right) \nu_l dt = \left(\int_{t_1}^{t_2} P_*(t)\,dt\right) \sum_{l=1}^{n} \nu_l$$

for a suitable function P_* of time. Hence, (P) implies that the reaction constraint due to the junction is directed along $\sum_l \nu_l$; i.e.,

$$\text{(Q)} \qquad \left(\sum_{l=1}^{n} P\left(\rho_l(t,0+), q_l(t,0+)\right) \nu_l\right) \in \left(\mathbb{R} \sum_{l=1}^{n} \nu_l\right).$$

This condition depends heavily on the geometry of the junction and is equivalent to (P) as soon as $\nu_1, \ldots, \nu_n$ are linearly independent; see Lemma 2.2.

In the case of $n = 2$ ducts with different sections, our model can be compared to the limit situation of [18, 19] corresponding to a delta source; see [14]. The case of a "kink," i.e., $n = 2$ nonparallel ducts having the same section, was considered in [17]. In these papers, the variation in the total linear momentum is quantified through a source term. In the present model, no source term is specified, and the modulus of the reaction constraint is not assumed to be known. On the contrary, here only the *direction* of the reaction constraint is used, and it is assumed to depend only on the geometry of the junction.

The friction against the pipes' walls is neglected here. When considered, it can be described through standard source terms along the pipes and a Dirac delta at the junction, using, for example, the techniques in [14, 17]. The present results are also preliminary to the study of more complicated networks. However, as the results in [13] also show, well posedness depends on the geometry of the network.

Our choice of considering the subsonic case is motivated by the relevance of this situation for applications. From the analytic point of view, as will be clear by the proofs below, the only relevant constraints are that (i) the total number of positive characteristic speeds must be equal to the number of pipes, and (ii) no transonic wave arises.

The main technique we use is the so-called wave-front tracking method; see [4, 10]. Therefore, we base our analysis on the solution to the Riemann problem at the junction, as in [8, 9]. This concept of solution is a natural extension of the classical Lax solution and reduces to it in the case of two ducts with the same section. Furthermore, it ensures the well posedness for the Riemann problem; hence the present definition is different from that introduced in [2, 3].

Similarly to the well posedness results found in the literature for the general p-system [4, 10], the well posedness of the Cauchy problem at a junction is proved below in the case where the initial data are a sufficiently small perturbation of a stationary solution.

There are a large number of papers dealing with various fluid-dynamical models in networks. Concerning car traffic in a network, we refer to the book [13] and to the papers [5, 6, 9, 11, 12, 16]. Concerning the evolution of gas in a network, see, for example, [2, 3, 8, 9]. Networks of open channels are studied with different techniques in [15].

Section 2 describes the model and recalls the solution to the Riemann problem at the junction, as defined in [8, 9]. Qualitative properties of the solution to the Riemann problem and numerical examples are also provided. Section 3 is devoted to the well posedness of solutions to the Cauchy problem. Finally, section 4 contains the technical details of the proofs.

2. The model. This section is devoted to definitions and results of the Riemann problem at a junction; see also [8, 9]. The physics of the fluid is described by the following usual condition on the pressure law in (1.1).

(EoS) The **E**quation **o**f **S**tate of the fluid, i.e., the pressure law $p = p(\rho)$, satisfies $p \in \mathbf{C^2}(\mathbb{R}^+; \mathbb{R}^+)$, $p(0) = 0$, $p' > 0$, and $p'' \geq 0$.

We denote $\mathbb{R}^+ = [0, +\infty[$ and $\mathring{\mathbb{R}}^+ =]0, +\infty[$. A typical example is the γ-law

$$p(\rho) = p_* \cdot (\rho/\rho_*)^\gamma \tag{2.1}$$

for fixed constants $\rho_*, p_* > 0$ and $\gamma \geq 1$. Other quantities relevant in the study of (1.1) are

$$\begin{aligned}
\text{flow of the linear momentum:} \quad & P(\rho, q) = \frac{q^2}{\rho} + p(\rho),\\
\text{total energy:} \quad & E(\rho, q) = \frac{q^2}{2\rho} + \rho \int_{\rho_*}^{\rho} \frac{p(r)}{r^2}\, dr,\\
\text{flow of the total energy:} \quad & F(\rho, q) = \frac{q}{\rho} \cdot \left(E(\rho, q) + p(\rho)\right).
\end{aligned}$$

We refer below to P as to the *dynamic pressure*. As is well known (see [10, formula (3.3.21)]), the pair (E, F) plays the role of the (mathematical) entropy-entropy flux pair.

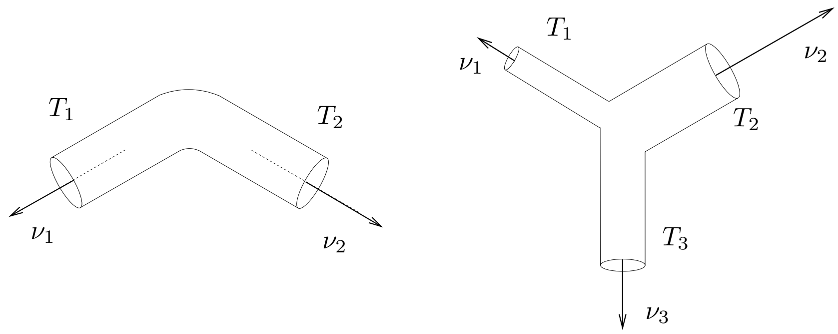
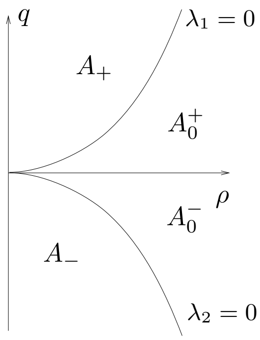


FIG. 2.1. *Examples of junctions and notation used in section* 3.

We introduce the regions



$$
\begin{aligned}
A_- &= \left\{\mathbb{R}^+ \times \mathbb{R} \colon \lambda_2(\rho, q) < 0\right\},\\
A_0^- &= \left\{\mathbb{R}^+ \times \mathbb{R} \colon \lambda_2(\rho, q) \geq 0,\ q \leq 0\right\},\\
A_0^+ &= \left\{\mathbb{R}^+ \times \mathbb{R} \colon \lambda_1(\rho, q) \leq 0,\ q \geq 0\right\},\\
A_+ &= \left\{\mathbb{R}^+ \times \mathbb{R} \colon \lambda_1(\rho, q) > 0\right\},\\
A_0 &= A_0^- \cup A_0^+.
\end{aligned}
$$

Above, as usual, λ_i is the ith characteristic speed; see (4.1) for its expression and for other relations holding for the p-system. We shall often refer to A_0 as the *subsonic* region.

Consider n ducts exiting a single junction. Each tube is modeled by $\mathbb{R}^+$, and the junction is at $x = 0$. The lth pipe is described by a vector ν_l parallel to it that exits the junction and whose norm $\|\nu_l\|$ is equal to the section of the duct; see Figure 2.1. For instance, standard Riemann problems correspond to (2.2) with $n = 2$ and $\nu_1 + \nu_2 = 0$. This choice makes the geometry of the junction intrinsic to the structure of the model; see Example 1 below.

Assigning at time $t = 0$ a constant initial state $(\bar{\rho}_l, \bar{q}_l) \in \mathbb{R}^+ \times \mathbb{R}$ in each of the n ducts exiting a junction ($l \in \{1, \ldots, n\}$), we have a *Riemann problem at the junction*:

$$
(2.2) \qquad \left\{
\begin{array}{ll}
\partial_t \rho_l + \partial_x q_l = 0, & t \in \mathbb{R}^+,\\
\partial_t q_l + \partial_x \left(\dfrac{q_l^2}{\rho_l} + p(\rho_l)\right) = 0, & \begin{array}{l} x \in \mathbb{R}^+,\\ l \in \{1, \ldots, n\},\end{array}\\
(\rho_l, q_l)(0, x) = (\bar{\rho}_l, \bar{q}_l), & (\bar{\rho}_l, \bar{q}_l) \in \mathring{\mathbb{R}}^+ \times \mathbb{R}.
\end{array}
\right.
$$

DEFINITION 2.1. *A* solution *to the Riemann problem* (2.2) *is a self-similar function* $(\rho, q) \in \mathbf{BV}(\mathbb{R}^+ \times \mathbb{R}^+; (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n)$ *such that the following hold:*

(L) *For all* $l = 1, \ldots, n$, *the function* $(t, x) \to (\rho_l, q_l)(t, x)$ *is self-similar and coincides with the restriction to* $x \in \mathring{\mathbb{R}}^+$ *of the* ***L****ax solution to the standard Riemann* *problem*

$$\begin{cases} \partial_t \rho_l + \partial_x q_l = 0, \\ \partial_t q_l + \partial_x \left(\dfrac{q_l^2}{\rho_l} + p(\rho_l) \right) = 0, \\ (\rho_l, q_l)(0,x) = \begin{cases} (\bar\rho_l, \bar q_l) & \textit{if } x > 0, \\ (\rho_l, q_l)(1, 0+) & \textit{if } x < 0 \,. \end{cases} \end{cases}$$

(M) ***M**ass is conserved at the junction; i.e., for a.e. $t > 0$,*

$$\sum_{l=1}^{n} \|\nu_l\| \, q_l(t, 0+) = 0 \,.$$

(P) *The trace of the dynamic **P**ressure P is the same along all tubes; i.e., there exists a positive P_* such that for $l = 1, \ldots, n$,*

$$P\left(\rho_l(t, 0+), q_l(t, 0+)\right) = P_* \,.$$

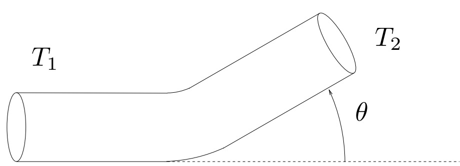
(E) *At the junction, **E**ntropy may not decrease; i.e., for a.e. $t > 0$,*

$$\sum_{l=1}^{n} \|\nu_l\| \, F\left(\rho_l(t, 0+), q_l(t, 0+)\right) \leq 0 \,.$$

Condition **(L)** implies that waves exiting the junction are standard Lax solutions to suitable Riemann problems. Hence, waves produced by the solution to (2.2) have positive speed in each duct, and the trace at the junction is constant in time.

The next example shows that, in Definition 2.1, geometry is hidden in the vectors $\nu_1, \ldots, \nu_n$.

Example 1. Consider an elbow of an angle $\pi - \theta$ between two tubes having the same section, so that $n = 2$, $\nu_1 = [-1 \quad 0]^T$, and $\nu_2 = [\cos\theta \quad \sin\theta]^T$.



Choose initial data $(\bar\rho_1, \bar q_1)$ and $(\bar\rho_2, \bar q_2)$. According to the conservation of mass **(M)**, which is part of Definition 2.1, if these data are in *equilibrium* (in the sense that no wave arises in the corresponding Riemann problem (2.2)), then the variation in the total linear momentum is $\Delta Q = \bar q_1 \nu_1 - \bar q_2 \nu_2 = \bar q_1 \left(\nu_1 + \nu_2\right)$. Note that ΔQ is directed along $\nu_1 + \nu_2$, coherently with (Q) and, moreover, $\|\Delta Q\| = |\bar q_1| \sqrt{2(1 - \cos\theta)}$, exactly as in [17, formula (0.3)].

As anticipated in the introduction, the independence of **(P)** from the geometry is only apparent. Indeed, **(P)** implies the more explicitly geometric condition (Q) and is equivalent to it as soon as $\nu_1, \ldots, \nu_n$ are linearly independent.

LEMMA 2.2. *For any n-tuple of pipes,* **(P)** $\Longrightarrow$ (Q). *Moreover, if $\nu_1, \ldots, \nu_n$ are linearly independent, then* **(P)** $\Longleftrightarrow$ (Q).

The proof is immediate. Note that **(P)** gives $n - 1$ conditions, while (Q) provides a number of conditions equal to the dimension of $\left(\sum \nu_l\right)^\perp$. In particular, with more than three tubes, **(P)** and (Q) may not be equivalent.

The results in [8, 9] can be slightly modified to cover the present case and ensure that the Riemann problem (2.2) with Definition 2.1 enjoys the same properties of standard Riemann problems with the standard Lax solutions [4, section 5.3]. In fact, we have the following proposition.

PROPOSITION 2.3. *Let p satisfy* (EoS). *Fix $n \in \mathbb{N}$ with $n \geq 2$ and a positive P_*. Choose n initial states satisfying*

$$(2.3) \qquad (\bar{\rho}_i, \bar{q}_i) \in \mathring{A}_0, \quad \sum_{i=1}^{n} \|\nu_i\| \bar{q}_i = 0, \quad P(\bar{\rho}_i, \bar{q}_i) = P_*, \quad \sum_{i=1}^{n} \|\nu_i\| F(\bar{\rho}_i, \bar{q}_i) < 0.$$

Then, for every $C > 0$, there exists $\delta > 0$ such that for all n-tuples of initial states $(\widetilde{\rho}, \widetilde{q}) \in (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n$ with $|\bar{\rho}_i - \widetilde{\rho}_i| + |\bar{q}_i - \widetilde{q}_i| < \delta$, the Riemann problem (2.2) *admits a unique solution $(\rho, q) = (\rho, q)(t, x)$ in the sense of Definition* 2.1 *satisfying*

$$(2.4) \qquad |\rho_i(t,x) - \bar{\rho}_i| + |q_i(t,x) - \bar{q}_i| < C,$$

for every $i \in \{1, \ldots, n\}$, $t \in \mathbb{R}^+$, and $x \in \mathbb{R}^+$. Moreover, in the case $n = 2$, $\nu_1 + \nu_2 = 0$, the map $(t, x) \to (\rho_*, q_*)(t, x)$ is the standard Lax solution to (1.1) *with data*

$$(\rho, q)(0, x) = \begin{cases} (\bar{\rho}_1, -\bar{q}_1) & \text{if} \quad x < 0, \\ (\bar{\rho}_2, \bar{q}_2) & \text{if} \quad x > 0 \end{cases}$$

if and only if the map

$$(2.5) \qquad (t, x) \to \begin{cases} (\rho_*, -q_*)(t, -x) & \text{if} \quad x \leq 0, \\ (\rho_*, q_*)(t, x) & \text{if} \quad x \geq 0 \end{cases}$$

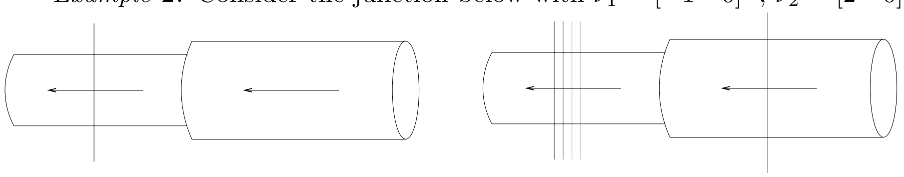
solves (2.2) *with data $(\bar{\rho}_1, \bar{q}_1)$ and $(\bar{\rho}_2, \bar{q}_2)$ in the sense of Definition* 2.1.

The proof is as in [8, 9]. Here, aiming both at the well posedness of the Riemann problem and at the qualitative properties of the solutions to (2.2), we consider some effects of interactions at the junction.

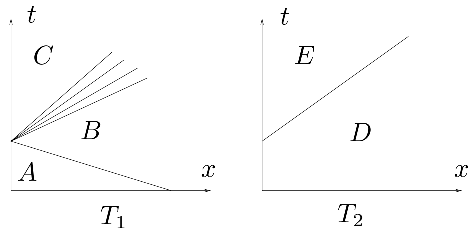
PROPOSITION 2.4. *Let p satisfy* (EoS). *Fix $n \in \mathbb{N}$ with $n \geq 2$ and a positive P_*. Choose n initial states $(\bar{\rho}_l, \bar{q}_l)$ satisfying* (2.3) *and thus producing a stationary solution for the Riemann problem. Then, there exists $\delta > 0$ such that if in the lth duct a rarefaction (resp., a shock) of the first family connecting $(\bar{\rho}_l, \bar{q}_l)$ with a state $(\rho_l^r, q_l^r) \in \mathring{A}_0$ satisfying $|(\bar{\rho}_l, \bar{q}_l) - (\rho_l^r, q_l^r)| < \delta$ reaches the junction, then a rarefaction (resp., a shock) propagates in all the other pipes.*

The proof is deferred to section 4. Here we give three numerical examples, where we used the γ-law (2.1) with $\gamma = 1.4$, $\rho_* = 1$, and $p_* = 1$.

Example 2. Consider the junction below with $\nu_1 = [-1 \quad 0]^T$, $\nu_2 = [2 \quad 0]^T$.



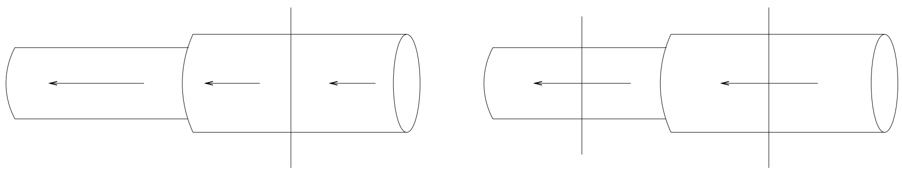
The arrows indicate the direction of the fluid. A 1-shock heads from the left toward the junction. Before the interaction the states (ρ, q) from left to right are as shown below:



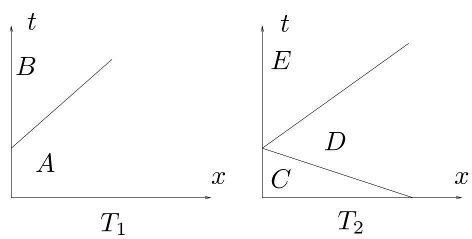
$$
\begin{aligned}
A &= (2.00001, 1.29999),\\
B &= (2.10000, 1.22504),\\
C &= (2.09816, 1.21891),\\
D &= (2.34829, -0.649996),\\
E &= (2.38388, -0.609455).
\end{aligned}
$$

After the interaction, a shock propagates in the pipe to the right, and a rarefaction is reflected to the left.

Example 3. Consider the same pipes as in Example 2.



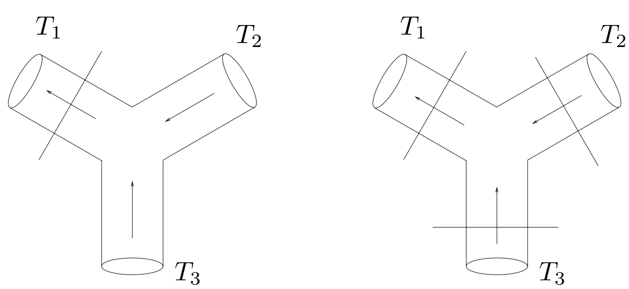
A 1-shock moves from the larger duct toward the smaller one:



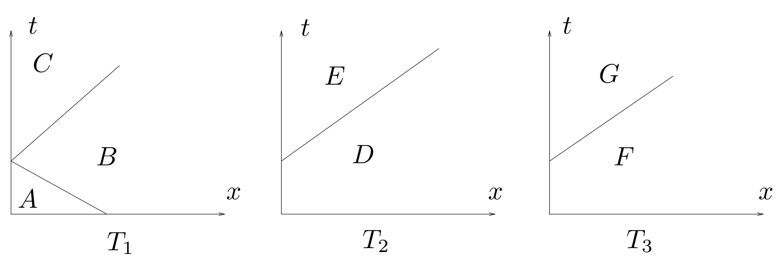
$$
\begin{aligned}
A &= (2.00001, 1.29999),\\
B &= (2.13796, 1.58483),\\
C &= (2.34829, -0.649996),\\
D &= (2.50000, -0.913070),\\
E &= (2.61032, -0.792455).
\end{aligned}
$$

After the interaction, a shock is reflected back to the right, and another shock is refracted in the smaller tube.

Example 4. We now consider an example with three pipes. Let $\nu_1 = [-1 \quad 0]^T$, $\nu_2 = [1/2 \quad \sqrt{3}/2]^T$, and $\nu_3 = [1/2 \quad -\sqrt{3}/2]^T$; i.e., all tubes have the same sections, and $\sum \nu_l = 0$.



The tubes are in equilibrium. A 1-shock approaches the junction along tube 1.



$$\begin{array}{ll} A = (1.0363, 1.0802), & D = (1.6273, -0.57179), \\ B = (1.6000,\ 0.79276), & E = (1.7694, -0.42680), \\ C = (1.6007,\ 0.79407), & F = (1.6528, -0.50842), \\ & G = (1.7847, -0.36827). \end{array}$$

Then, the resulting interaction at the junction forms a shock in each tube.

3. Analytical results. Consider the *Cauchy problem at a junction* with n pipes, i.e., the problem

$$\left\{\begin{array}{ll} \partial_t \rho_l + \partial_x q_l = 0, & t \in \mathbb{R}^+, \\ \partial_t q_l + \partial_x \left(\dfrac{q_l^2}{\rho_l} + p(\rho_l)\right) = 0, & x \in \mathbb{R}^+, \\ & l \in \{1, \ldots, n\}, \\ (\rho, q)(0, x) = (\rho_o, q_o)(x), & (\rho_o, q_o) \in \mathbf{L^1}\left(\mathbb{R}^+; (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n\right). \end{array}\right. \tag{3.1}$$

For $(\rho, q) \in \mathbf{L^1}(\mathbb{R}^+; (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n)$ attaining values in a neighborhood of a fixed $(\hat\rho, \hat q) \in (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n$, with a slight abuse of notation, we introduce

$$\|(\rho, q)\|_{\mathbf{L^1}} = \sum_{l=1}^n \|(\rho_l, q_l)\|_{\mathbf{L^1}} \quad \text{and} \quad \mathrm{TV}\,(\rho, q) = \sum_{l=1}^n \mathrm{TV}\,(\rho_l, q_l)\,.$$

The natural extension of the usual definition of a weak entropy solution to the present case is the following.

DEFINITION 3.1. *Fix $(\hat\rho, \hat q) \in (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n$ and $T \in\,]0, +\infty]$. A weak solution to* (3.1) *is a map $(\rho, q) \in \mathbf{C^0}([0, T]; (\hat\rho, \hat q) + \mathbf{L^1}(\mathbb{R}^+; (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n))$ such that for all $\varphi \in \mathbf{C_c^\infty}(]-\infty, T[\, \times \mathbb{R}^+; \mathbb{R})$*

$$\sum_{l=1}^n \left(\int_0^T \!\! \int_{\mathbb{R}^+} (\rho_l\, \partial_t \varphi + q_l\, \partial_x \varphi)\, dx\, dt + \int_{\mathbb{R}^+} \rho_{o,l}(x)\, \varphi(0, x)\, dx\right) \|\nu_l\| = 0,$$

$$\left(\int_0^T \!\! \int_{\mathbb{R}^+} (q_l \partial_t \varphi + P_l \partial_x \varphi)\, dx dt + \int_{\mathbb{R}^+} q_{o,l}(x)\, \varphi(0, x)\, dx\right) \|\nu_l\| = \int_0^T P_*(t) \varphi(t, 0) dt$$

for all $l = 1, \ldots, n$ and for a suitable $P_* \in \mathbf{L^1}([0, T]; \mathbb{R}^+)$. The weak solution (ρ, q) is entropic if for all $\varphi \in \mathbf{C_c^\infty}(] - \infty, T[\times \mathbb{R}^+; \mathbb{R}^+)$

$$\sum_{l=1}^n \left(\int_0^T \!\! \int_{\mathbb{R}^+} (E_l\, \partial_t \varphi + F_l\, \partial_x \varphi)\, dx\, dt + \int_{\mathbb{R}^+} E(\rho_{o,l}, q_{o,l})\, \varphi(0, x)\, dx\right) \|\nu_l\| \geq 0,$$

where $E_l = E(\rho_l, q_l)$ and $F_l = F(\rho_l, q_l)$.

Above, the condition on the second equation clearly reflects **(P)**. A similar condition related to (Q) is

$$\sum_{l=1}^n \left(\int_0^T \!\! \int_{\mathbb{R}^+} (q_l\, \partial_t \varphi + P_l\, \partial_x \varphi)\, dx\, dt + \int_{\mathbb{R}^+} q_{o,l}(x)\, \varphi(0, x)\, dx\right) \nu_l \in \left(\mathbb{R} \sum_{l=1}^n \nu_l\right).$$

A simple condition for (ρ, q) to be a weak entropy solution is provided by the following lemma.

LEMMA 3.2. *Let $n \in \mathbb{N}$ with $n \geq 2$ and $\nu_1, \ldots, \nu_n \in \mathbb{R}^3$ be pairwise distinct. Fix a state $(\hat\rho, \hat q) \in (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n$ and $T \in\,]0, +\infty]$. If the map $(\rho, q) \in \mathbf{C^0}([0, T]; (\hat\rho, \hat q) + \mathbf{L^1}(\mathbb{R}^+; (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n))$ satisfies*

1. *for* $l = 1, \ldots, n$, *each map* (ρ_l, q_l) *is a weak entropy solution to* (1.1) *for* $x \in \mathring{\mathbb{R}}^+$ *and* $t \in \mathbb{R}^+$,
2. *for a.e.* $t \in \mathring{\mathbb{R}}^+$, *the trace* $(\bar\rho_l, \bar q_l)(t) = \lim_{x \to 0+} (\rho_l, q_l)(t, x)$ *is the initial data of a stationary solution to the Riemann problem* (2.2) *in the sense of Definition* 2.1,

then it is a weak solution to (3.1).

The proof is an adaptation of [9, Propositions 2.1, 2.2, and 3.2] and is omitted. We proceed to the main result of this paper: the well posedness of the solution to the Cauchy problem for the p-system at a junction.

THEOREM 3.3. *Fix an n-tuple of subsonic states* $(\hat\rho, \hat q) \in (\mathring{A}_0)^n$, *giving a stationary solution to the Riemann problem* (2.2), *in the sense of Definition* 2.1, *with entropy flux strictly negative. Then, there exist positive constants* δ_0, L *and a map* $S\colon [0, +\infty[\times \mathcal{D} \to \mathcal{D}$, *with the following properties*:

1. $\mathcal{D} \supseteq \{(\rho, q) \in (\hat\rho, \hat q) + \mathbf{L^1}(\mathbb{R}^+; (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n)\colon \mathrm{TV}(\rho, q) \le \delta_0\}$.
2. *For* $(\rho, q) \in \mathcal{D}$, $S_0(\rho, q) = (\rho, q)$ *and for* $s, t \ge 0$, $S_s S_t(\rho, q) = S_{s+t}(\rho, q)$.
3. *For* $(\rho, q), (\rho', q') \in \mathcal{D}$ *and* $s, t \ge 0$,
$$\|S_t(\rho, q) - S_s(\rho', q')\|_{\mathbf{L^1}} \le L \cdot (\|(\rho, q) - (\rho', q')\|_{\mathbf{L^1}} + |t - s|).$$
4. *If* $(\rho, q) \in \mathcal{D}$ *is piecewise constant, then for* $t > 0$ *sufficiently small,* $S_t(\rho, q)$ *coincides with the juxtaposition of the solutions to Riemann problems centered at the points of jumps or at the junction.*

Moreover, for every $(\rho, q) \in \mathcal{D}$, *the map* $t \to S_t(\rho, q)$ *is a weak entropy solution to the Cauchy problem* (3.1) *in the sense of Definition* 3.1.

Note that 1–4 above are the natural extension of the definition of the standard Riemann semigroup [4, Definition 9.1] to the Cauchy problem (3.1). In general, the linear momentum fails to be conserved. Indeed, the following estimate holds.

PROPOSITION 3.4. *Let* $n \in \mathbb{N}$ *with* $n \ge 2$ *and* $\nu_1, \ldots, \nu_n \in \mathbb{R}^3$ *be pairwise distinct. Fix* $(\hat\rho, \hat q) \in (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n$ *and* $T \in]0, +\infty]$. *Let* $(\rho, q) \in \mathbf{C^0}([0, T]; (\hat\rho, \hat q) + \mathbf{L^1}(\mathbb{R}^+; (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n))$ *be a solution to* (1.1) *in the sense of Definition* 2.1. *Then,*

$$Q(t_2) - Q(t_1) = \left(\int_{t_1}^{t_2} P_*(t)\, dt \right) \sum_{l=1}^{n} \nu_l,$$

where, for $l = 1, \ldots, n$ *and a.e.* $t_1, t_2 \in [0, T]$, $P_*(t) = P(\rho_l(t, 0^+), q_l(t, 0^+))$.

The proof is similar to that of [9, Proposition 3.2] and is omitted. The term on the right-hand side above is the impulse of the reaction constraint and describes the defect in the conservation of the total linear momentum.

4. Technical details. The first part of this section deals with the basic properties of the p-system and with the proof of Proposition 2.4. Subsection 4.2 describes the wave-front tracking construction and the estimates for the existence and well posedness of the Cauchy problem.

In this section, we use the notation $u = [\rho \quad q]^T$.

4.1. Properties of the *p*-system. We recall here basic formulas of the p-system (1.1) valid for a pressure law satisfying **(EoS)**. Throughout, $c(\rho) = \sqrt{p'(\rho)}$ denotes the sound speed. Let λ_i be the ith eigenvalue corresponding to the ith right eigenvector r_i of the Jacobian of the flow $f(\rho, q) = [q \;\; q^2/\rho + p(\rho)]^T$. We have

$$
\begin{aligned}
&\lambda_1(\rho, q) = \tfrac{q}{\rho} - c(\rho)\,, && \lambda_2(\rho, q) = \tfrac{q}{\rho} + c(\rho)\,,\\
&r_1(\rho, q) = \begin{bmatrix} \rho \\ q - \rho c(\rho) \end{bmatrix}, && r_2(\rho, q) = \begin{bmatrix} \rho \\ q + \rho c(\rho) \end{bmatrix},\\
&\nabla\lambda_1 \cdot r_1 = -c(\rho) - \rho c'(\rho), && \nabla\lambda_2 \cdot r_2 = c(\rho) + \rho c'(\rho).
\end{aligned}
\tag{4.1}
$$

The speeds of $1,2$-shock waves between (ρ_o, q_o) and the state at density ρ are

$$\Lambda_{\frac{1}{2}}(\rho, \rho_o, q_o) = \frac{q_o}{\rho_o} \mp \sqrt{\frac{\rho}{\rho_o} \cdot \frac{p(\rho) - p(\rho_o)}{\rho - \rho_o}}.$$

The (forward) $1,2$-Lax curves have the expressions

$$
L_1(\rho; \rho_o, q_o) = \begin{cases} \dfrac{\rho}{\rho_o} q_o - \rho \displaystyle\int_{\rho_o}^{\rho} \frac{c(r)}{r}\, dr & \text{if } \rho < \rho_o, \\ \dfrac{\rho}{\rho_o} q_o - \sqrt{\dfrac{\rho}{\rho_o} (\rho - \rho_o)\,(p(\rho) - p(\rho_o))} & \text{if } \rho > \rho_o, \end{cases}
$$

$$
L_2(\rho; \rho_o, q_o) = \begin{cases} \dfrac{\rho}{\rho_o} q_o - \sqrt{\dfrac{\rho}{\rho_o} (\rho - \rho_o)\,(p(\rho) - p(\rho_o))} & \text{if } \rho < \rho_o, \\ \dfrac{\rho}{\rho_o} q_o + \rho \displaystyle\int_{\rho_o}^{\rho} \frac{c(r)}{r}\, dr & \text{if } \rho > \rho_o \end{cases}
$$

(see Figure 4.1, right), while the reversed $1,2$-Lax curves exiting $(\bar\rho, \bar q)$ are

$$
\begin{aligned}
L_1^-(\rho; \bar\rho, \bar q) &= \begin{cases} \dfrac{\rho}{\bar\rho} \bar q + \sqrt{\dfrac{\rho}{\bar\rho} (\bar\rho - \rho)\,(p(\bar\rho) - p(\rho))} & \text{if } \rho < \bar\rho, \\ \dfrac{\rho}{\bar\rho} \bar q - \rho \displaystyle\int_{\bar\rho}^{\rho} \frac{c(r)}{r} dr & \text{if } \rho > \bar\rho, \end{cases}\\
L_2^-(\rho; \bar\rho, \bar q) &= \begin{cases} \dfrac{\rho}{\bar\rho} \bar q + \sqrt{\dfrac{\rho}{\bar\rho} (\rho - \bar\rho)\,(p(\rho) - p(\bar\rho))} & \text{if } \rho > \bar\rho, \\ \dfrac{\rho}{\bar\rho} \bar q - \rho \displaystyle\int_{\rho}^{\bar\rho} \frac{c(r)}{r} dr & \text{if } \rho < \bar\rho. \end{cases}
\end{aligned}
\tag{4.2}
$$

Proof of Proposition 2.4. For simplicity denote with u_l the states $(\bar\rho_l, \bar q_l)$. Assume that in a pipe, say 1, a shock wave connecting u_1 with $u_1^r = (\rho_1^r, q_1^r)$ arrives at the junction. Clearly $\rho_1 < \rho_1^r$. We have

$$\frac{d}{d\rho} P(\rho, L_1(\rho; u_1))_{|\rho=\rho_1} = (\lambda_1(u_1))^2 > 0,$$

and thus $P(u_1) < P(u_1^r)$. Denote with $\tilde u_1 \in A_0$ the point of the reversed Lax curve of the second family through u_1^r having dynamical pressure equal to $P(u_1)$. The n-tuple $(\tilde u_1, u_2, \ldots, u_n)$ does not provide a stationary solution to the Riemann problem at the junction, since the sum in **(M)** is less than 0. Therefore the solution to the Riemann problem has a dynamical pressure $P_* > P(u_1)$. Since the dynamical pressure P is strictly increasing along the reversed Lax curves of the second family [8, Lemma 1], we conclude that shock waves are generated in all pipes, except the first one.

Similar considerations hold in the case of a rarefaction wave. ∎

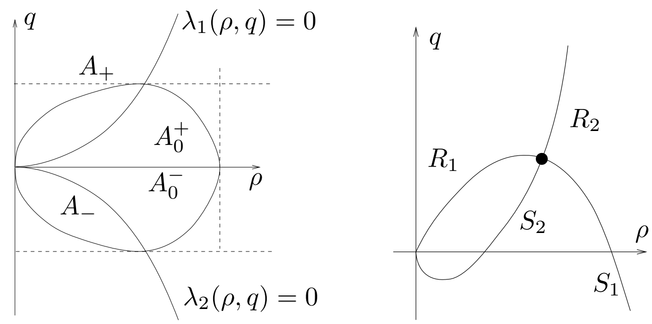


FIG. 4.1. *Left*: *the regions* A_-, $A_0^{\pm}$, A_+ *and a level curve of the dynamic pressure. Right*: *the Lax forward curves for* (1.1).

4.2. Wave-front tracking with the junction. This subsection is devoted to the proof of Theorem 3.3. To do this, we introduce a wave-front tracking approximation for our Cauchy problem and a functional measuring the distance in $\mathbf{L^1}$ between two piecewise constant solutions. The construction of a solution to (3.1) is given adapting the wave-front tracking technique; see [4, Chapter 7] or [10, Chapter 14].

Define $\hat{u}_l = (\hat{\rho}_l, \hat{q}_l)$ for $l = 1, \ldots, n$. Let $\hat{\delta} > 0$ be such that $B(\hat{u}_l, \hat{\delta}) \subset \mathring{A}_0$ for $l = 1, \ldots, n$ and introduce the compact set $\mathcal{B} = \prod_{l=1}^n \overline{B(\hat{u}_l, \hat{\delta})}$.

Approximate the initial datum u_o with a sequence $u_{o,\nu}$ of piecewise constant initial data each having a finite number of discontinuities so that $\lim_{\nu \to +\infty} \|u_{o,\nu} - u_o\|_{\mathbf{L^1}} = 0$. Then, at the junction and at each point of jump in the approximate initial datum along the ducts, we solve the corresponding Riemann problem. The Riemann problem at the junction is solved according to Definition 2.1. If the total variation of the initial datum is sufficiently small, then Theorem 2 in [8] assures the existence and uniqueness of the solution to the Riemann problem. We approximate each rarefaction wave with a rarefaction fan, i.e., by means of (nonentropic) shock waves traveling at the characteristic speed of the state to the right of the shock.

This construction can be extended up to the first time $\bar{t}_1$ at which two waves interact in a duct or a wave hits the junction. Clearly, at time $\bar{t}_1$ the functions so constructed are piecewise constant with a finite number of discontinuities. Hence, at any subsequent interaction or collision with the junction, we repeat the previous construction with the following provisions:

1. No more than two waves interact at the same point or at the junction.
2. A rarefaction fan of the ith family produced by the interaction between an ith rarefaction and any other wave is *not* split any further.
3. When the product of the strengths of two interacting waves falls below a threshold $\check{\varepsilon}$, then we let the waves cross each other, their size being unaltered, and introduce a *nonphysical* wave with speed $\hat{\lambda}$, with $\hat{\lambda} > \sup_u \lambda_2(u)$; see [4, Chapter 7] and the refinement [1].

In the present case, we have to complete the above algorithm by stating how the Riemann problem at the junction is to be solved. At time $t = 0$ and whenever a physical wave with size greater than $\check{\varepsilon}$ hits the junction, the accurate solver is used; i.e., the exact solution as in Definition 2.1 is approximated by replacing rarefaction waves with rarefaction fans. When a wave with strength smaller than $\check{\varepsilon}$ hits the junction, then we let it be reflected into a nonphysical wave with speed $\hat{\lambda}$, and no wave in any other duct is produced.

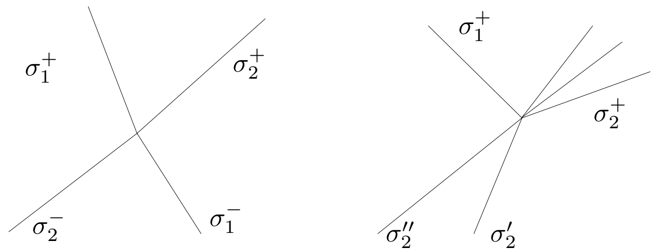


FIG. 4.2. *Interactions in the* (t, x) *plane and notation for the standard interaction estimates in Lemma* 4.1.

Repeating this procedure recursively, we construct a wave-front tracking sequence of approximate solutions u_ν in the sense of [4, Definition 7.1].

A key role in wave-front tracking is played by interaction estimates. At interactions of waves in a duct, we have the following classical result.

LEMMA 4.1. *There exists a constant* K *with the following property.*

1. *If there is an interaction in a duct between two waves* σ_1^- *and* σ_2^-, *respectively of the first and the second family, producing the waves* σ_1^+ *and* σ_2^+ *(see Figure* 4.2, *left), then*

$$\left|\sigma_1^+ - \sigma_1^-\right| + \left|\sigma_2^+ - \sigma_2^-\right| \leq K \cdot \left|\sigma_1^- \sigma_2^-\right| . \tag{4.3}$$

2. *If there is an interaction in a duct between two waves* σ_i' *and* σ_i'' *of the same* i*th family producing waves of total size* σ_1^+ *and* σ_2^+ *(see Figure* 4.2, *right, for the case* $i = 2$*), then*

$$\begin{aligned} \left|\sigma_1^+ - (\sigma_1'' + \sigma_1')\right| + \left|\sigma_2^+\right| &\leq K \cdot |\sigma_1' \sigma_1''| \qquad \text{if } i = 1, \\ \left|\sigma_1^+\right| + \left|\sigma_2^+ - (\sigma_2'' + \sigma_2')\right| &\leq K \cdot |\sigma_2' \sigma_2''| \qquad \text{if } i = 2. \end{aligned}$$

3. *If there is an interaction in a duct between two physical waves* σ_1^- *and* σ_2^- *producing a nonphysical wave* σ_3^+ *(see Figure* 4.3, *left), then*

$$\left|\sigma_3^+\right| \leq K \cdot \left|\sigma_1^- \sigma_2^-\right|.$$

4. *If there is an interaction in a duct between a physical wave* σ *and a nonphysical wave* σ_3^- *producing a physical wave* σ *and a nonphysical wave* σ_3^+ *(see Figure* 4.3, *right), then*

$$\left|\sigma_3^+\right| - \left|\sigma_3^-\right| \leq K \cdot \left|\sigma \sigma_3^-\right|.$$

For a proof of this result see [4, Chapter 7]. Remember that nonphysical waves cannot interact with the junction or with other nonphysical waves.

In the case of the junction, with the notation in Figure 4.4, we have the following result.

PROPOSITION 4.2. *Let* (EoS) *hold. There exist* $\delta_J > 0$ *and* $K_J \geq 1$ *with the following property. For any* $\bar{u} \in \mathcal{B}$ *that yields a stationary solution to the Riemann problem* (2.2), *for any* 1*-waves* $\sigma_l^- \in \,]-\delta_J, \delta_J[$ *hitting the junction and producing the* 2*-waves* σ_l^+, *it holds that*

$$\sum_{l=1}^{n} \left|\sigma_l^+\right| \leq K_J \cdot \sum_{l=1}^{n} \left|\sigma_l^-\right| . \tag{4.4}$$

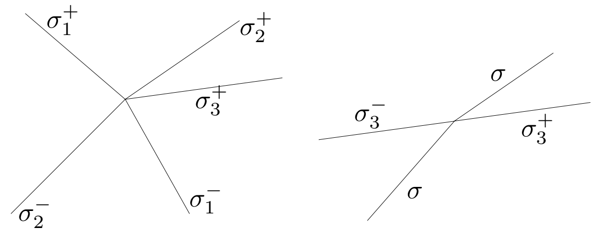


FIG. 4.3. *Left: in the* (t, x) *plane, a nonphysical wave arises, and right: a nonphysical wave interacts with a physical one.*

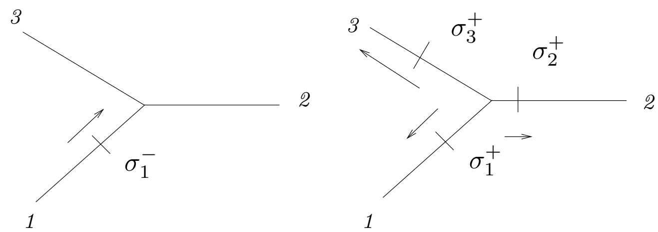


FIG. 4.4. *Notation for Proposition* 4.2: *in the physical space, before the interaction (left) and after the interaction (right).*

Proof. We recall that, similarly to the proof of [8, Theorem 1], the densities ρ_l^+ after the interaction solve

$$\sum_{l=1}^{n} \|\nu_l\| \, L_2^-(\rho_l^+; \rho_l^-, L_1(\rho_l^-; \bar{u})) = 0,$$

$$P(\rho_l^+, L_2^-(\rho_l^+; L_1(\rho_l^-; \bar{u}))) = P(\rho_{l-1}^+, L_2^-(\rho_{l-1}^+; L_1(\rho_{l-1}^-; \bar{u}))).$$

Applying the implicit function theorem in a neighborhood of $\rho_l^- = \rho_l^+ = \bar{\rho}$, we obtain $(\rho_1^+, \dots, \rho_n^+)$ as a function of $(\rho_1^-, \dots, \rho_n^-)$. Moreover,

$$\sum_{l=1}^{n} \left|\rho_l^+ - \bar{\rho}\right| \le \tilde{K} \cdot \sum_{l=1}^{n} \left|\rho_l^- - \bar{\rho}\right|.$$

Pass now to the arc-length parametrization, since in $\mathcal{B}$ the arc-length is a bi-Lipschitz function of the density variation. More precisely, the unknowns $\partial_{\rho_1^-} \rho_l^+$ solve the linear system $Ax = b$, where, setting $\lambda_i(l) = \lambda_i(\bar{\rho}_l, \bar{q}_l)$,

$$A = \begin{bmatrix} \|\nu_1\|\lambda_2(1) & \|\nu_2\|\lambda_2(2) & \|\nu_3\|\lambda_2(3) & \dots & \|\nu_n\|\lambda_2(n) \\ (\lambda_2(1))^2 & -(\lambda_2(2))^2 & 0 & \dots & 0 \\ (\lambda_2(1))^2 & 0 & -(\lambda_2(3))^2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ (\lambda_2(1))^2 & 0 & 0 & \dots & -(\lambda_2(n))^2 \end{bmatrix},$$

$$b = \|\nu_1\|\lambda_1(1) \begin{bmatrix} 1 & \lambda_2(1) + 2\frac{q}{\rho} & \lambda_2(1) + 2\frac{q}{\rho} & \cdots & \lambda_2(1) + 2\frac{q}{\rho} \end{bmatrix}^T.$$

This concludes the proof. $\square$

Define $K_1 = 2KK_J + 1$. Fix a wave-front tracking approximate solution u_ν. For $t > 0$ and $l \in \{1, \dots, n\}$, we denote with $x_{l,\alpha}$ $(\alpha \in J_l(u))$ the positions of the discontinuities of the approximate solution in the lth duct and with $\sigma_{l,1,\alpha}$, $\sigma_{l,2,\alpha}$, $\sigma_{l,3,\alpha}$ the strengths of the waves of the first family, the second family, and the nonphysical waves at $x_{l,\alpha}$, respectively. Introduce the Glimm-type functionals

$$V(t) = \sum_{l=1}^{n} \sum_{\alpha \in \mathcal{J}_l} [2\,K_J \cdot |\sigma_{l,1,\alpha}| + |\sigma_{l,2,\alpha}| + |\sigma_{l,3,\alpha}|],$$

$$Q(t) = \sum_{l=1}^{n} \sum \{ |\sigma_{l,i,\alpha}\,\sigma_{l,j,\beta}| \colon (\sigma_{l,i,\alpha}, \sigma_{l,j,\beta}) \in \mathcal{A}_l \},$$

$$\Upsilon(t) = V(t) + K_1 \cdot Q(t), \tag{4.5}$$

where $\mathcal{A}_l$ denotes the set of approaching waves in the lth duct; see [4].

The functionals above are well defined for every $t > 0$ at which no interaction happens. Suppose now that at a time $\tau > 0$ there is an interaction between $k \in \{1, \ldots, n\}$ waves $\sigma_{k,1,\alpha}$ of the first family and the junction. This interaction produces n waves $\sigma'_{l,2,\alpha}$ of the second family. Thus

$$\Delta V(\tau) \le \sum_{l=1}^{n} [\,|\sigma'_{l,2,\alpha}| - 2\,K_J\,|\sigma_{l,1,\alpha}|\,] \;\le\; -K_J \sum_{l=1}^{n} |\sigma_{l,1,\alpha}|,$$

$$\Delta Q(\tau) \le K_J V(\tau-) \sum_{l=1}^{n} |\sigma_{l,1,\alpha}|,$$

$$\Delta \Upsilon(\tau) \le K_J \cdot [K_1\, V(\tau-) - 1] \cdot \sum_{l=1}^{n} \left|\sigma^-_{l,1,\alpha}\right|.$$

Suppose now that an interaction between two waves $\sigma_{l,i,\alpha}$, $\sigma'_{l,j,\beta}$ happens in a duct at time τ. We deduce that

$$\Delta V(\tau) \le 2KK_J \left|\sigma_{l,i,\alpha}\sigma'_{l,j,\beta}\right|,$$

$$\Delta Q(\tau) \le K \left|\sigma_{l,i,\alpha}\sigma'_{l,j,\beta}\right| V(\tau-) - \left|\sigma_{l,i,\alpha}\sigma'_{l,j,\beta}\right|,$$

$$\Delta \Upsilon(\tau) \le \left|\sigma_{l,i,\alpha}\sigma'_{l,j,\beta}\right| [K(2K_J + K_1 V(\tau-)) - K_1].$$

We have thus proved the following basic result.

PROPOSITION 4.3. *Let* $\delta_1 = (1/(2K_1)) \min\{1/K, 1\}$. *At any interaction time* $\tau > 0$*, if* $V(\tau-) < \delta_1$*, then* $\Delta\Upsilon(\tau) < 0$ *with* Υ *defined in* (4.5).

Proof of Theorem 3.3. Let δ_1 be as in Proposition 4.3 and define

$$\tilde{\mathcal{D}} = \{ u \in (\hat\rho, \hat q) + (\mathbf{PC} \cap \mathbf{L^1}(\mathbb{R}^+; (\mathring{\mathbb{R}}^+ \times \mathbb{R})^n)) \colon \Upsilon(u) \le \delta_1 \}.$$

There exists $C_1 > 0$ such that $\frac{1}{C_1}\,\mathrm{TV}\,(u)(t,\cdot) \le V(t) \le C_1\,\mathrm{TV}\,(u)(t,\cdot)$ for all $u \in \tilde{\mathcal{D}}$. Any initial datum in $\tilde{\mathcal{D}}$ yields an approximate solution to (1.1) attaining values in $\tilde{\mathcal{D}}$ by Proposition 4.3, and thus, by classical arguments (see [4]), for all initial data in $\mathcal{D}$ a solution to (3.1) exists for every $t > 0$ and attains values in $\mathcal{D}$, where $\mathcal{D}$ is the $\mathbf{L^1}$ closure of $\tilde{\mathcal{D}}$. Hence, statements 1 and 2 in Theorem 3.3 clearly hold. The limit orbits $t \to S_t u$ are indeed solutions in the sense of Definition 3.1; note that the existence of the trace $P_*(t)$ at the junction follows from Definition 2.1 and [7, Proposition 5.3].

We pass now to the well posedness of the Cauchy problem. Consider two ε–wave-front tracking approximate solutions u_1 and u_2. Define the functional

$$\Phi(u_1, u_2) = \sum_{l=1}^{n} \sum_{i=1}^{2} \int_0^{+\infty} |s_{l,i}(x)|\, W_{l,i}(x)\, dx\,, \tag{4.6}$$

where $s_{l,i}(x)$ measures the strengths of the ith shock wave in the lth duct at point x (see [4, Chapter 8]) and the weights $W_{l,i}$ are defined by

$$W_{l,i}(x) = 1 + \kappa_1 \, A_{l,i}(x) + \kappa_1 \, \kappa_2 \, [\Upsilon(u_1) + \Upsilon(u_2)].$$

Here Υ is the functional defined in (4.5), while the $A_{l,i}$ are defined by

$$\begin{aligned} A_{l,i}(x) = {} & \sum \left\{ |\sigma_{l,k_\alpha,\alpha}| : \begin{array}{l} x_\alpha < x, \ i < k_\alpha \leq 2, \\ x_\alpha > x, \ 1 \leq k_\alpha < i, \end{array} \right\} \\ & + \left\{ \begin{array}{ll} \sum \left\{ |\sigma_{l,i,\alpha}| : \begin{array}{l} x_\alpha < x, \ \alpha \in J(u_1), \\ x_\alpha > x, \ \alpha \in J(u_2), \end{array} \right\} & \text{if } q_{l,i}(x) < 0, \\ \sum \left\{ |\sigma_{l,i,\alpha}| : \begin{array}{l} x_\alpha < x, \ \alpha \in J(u_2), \\ x_\alpha > x, \ \alpha \in J(u_1), \end{array} \right\} & \text{if } q_{l,i}(x) \geq 0; \end{array} \right. \end{aligned}$$

see [4, Chapter 8]. We first fix κ_1, κ_2 so that δ_0 in the definition of $\mathcal{D}$ can be chosen to satisfy $1 \leq W_{l,i}(x) \leq 2$ for every $l \in \{1, \ldots, n\}$, $i \in \{1, 2\}$, and $x \geq 0$. Hence the functional Φ is equivalent to the $\mathbf{L^1}$ distance; i.e.,

$$\Phi(u_1, u_2) \geq \frac{1}{C_2} \cdot \|u_1 - u_2\|_{\mathbf{L^1}} \quad \text{and} \quad \Phi(u_1, u_2) \leq C_2 \cdot \|u_1 - u_2\|_{\mathbf{L^1}}$$

for a positive constant C_2. The same calculations as in [4, Chapter 8] show that, at any time $t > 0$ when an interaction happens neither in u_1 nor in u_2,

$$\frac{d}{dt} \Phi(u_1(t), u_2(t)) \leq C_3 \, \varepsilon \,, \tag{4.7}$$

where C_3 is a suitable positive constant depending only on a bound on the total variation of the initial data. If $t > 0$ is an interaction time for u_1 or u_2, then, by Proposition 4.3, $\Delta[\Upsilon(u_1(t)) + \Upsilon(u_2(t))] < 0$ and, choosing κ_2 large enough, we obtain

$$\Delta \Phi(u_1(t), u_2(t)) < 0 \,. \tag{4.8}$$

Thus, $\Phi(u_1(t), u_2(t)) - \Phi(u_1(s), u_2(s)) \leq C_2 \, \varepsilon \, (t - s)$ for every $0 \leq s \leq t$, proving statement 3. Statement 4 follows by standard arguments. $\square$

Acknowledgments. The authors thank Graziano Guerra and Giovanni Russo for several useful discussions. The referees' reports also provided valuable constructive criticisms.

REFERENCES

[1] P. Baiti and H. K. Jenssen, *On the front-tracking algorithm*, J. Math. Anal. Appl., 217 (1998), pp. 395–404.

[2] M. K. Banda, M. Herty, and A. Klar, *Coupling conditions for gas networks governed by the isothermal Euler equations*, Netw. Heterog. Media, 1 (2006), pp. 295–314.

[3] M. K. Banda, M. Herty, and A. Klar, *Gas flow in pipeline networks*, Netw. Heterog. Media, 1 (2006), pp. 41–56.

[4] A. Bressan, *Hyperbolic Systems of Conservation Laws—The One-Dimensional Cauchy Problem*, Oxford Lecture Ser. Math. Appl. 20, Oxford University Press, Oxford, UK, 2000.

[5] Y. Chitour and B. Piccoli, *Traffic circles and timing of traffic lights for cars flow*, Discrete Contin. Dyn. Syst. Ser. B, 5 (2005), pp. 599–630.

[6] G. M. Coclite, M. Garavello, and B. Piccoli, *Traffic flow on a road network*, SIAM J. Math. Anal., 36 (2005), pp. 1862–1886.

[7] R. M. Colombo and A. Corli, *Sonic hyperbolic phase transitions and Chapman-Jouguet detonations*, J. Differential Equations, 184 (2002), pp. 321–347.

[8] R. M. Colombo and M. Garavello, *A well posed Riemann problem for the p-system at a junction*, Netw. Heterog. Media, 1 (2006), pp. 495–511.

[9] R. M. Colombo and M. Garavello, *On the p-system at a junction*, Contemp. Math., 426 (2007), pp. 193–217.

[10] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, 1st ed., Springer-Verlag, Berlin, 2000.

[11] M. Garavello and B. Piccoli, *Source-destination flow on a road network*, Commun. Math. Sci., 3 (2005), pp. 261–283.

[12] M. Garavello and B. Piccoli, *Traffic flow on a road network using the Aw-Rascle model*, Comm. Partial Differential Equations, 31 (2006), pp. 243–275.

[13] M. Garavello and B. Piccoli, *Traffic Flow on Networks—Conservation Laws Models*, AIMS Ser. Appl. Math. 1, American Institute of Mathematical Sciences, Springfield, MO, 2006.

[14] P. Goatin and P. G. LeFloch, *The Riemann problem for a class of resonant hyperbolic systems of balance laws*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 21 (2004), pp. 881–902.

[15] M. Gugat, G. Leugering, K. Schittkowski, and E. J. P. G. Schmidt, *Modelling, stabilization, and control of flow in networks of open channels*, in Online Optimization of Large Scale Systems, Springer-Verlag, Berlin, 2001, pp. 251–270.

[16] H. Holden and N. H. Risebro, *A mathematical model of traffic flow on a network of unidirectional roads*, SIAM J. Math. Anal., 26 (1995), pp. 999–1017.

[17] H. Holden and N. H. Risebro, *Riemann problems with a kink*, SIAM J. Math. Anal., 30 (1999), pp. 497–515.

[18] T. P. Liu, *Nonlinear stability and instability of transonic flows through a nozzle*, Comm. Math. Phys., 83 (1982), pp. 243–260.

[19] T. P. Liu, *Transonic gas flow in a duct of varying area*, Arch. Rational Mech. Anal., 80 (1982), pp. 1–18.

SIAM J. MATH. ANAL.
Vol. 39, No. 5, pp. 1472–1496

CONVERGENCE OF GIBBS MEASURES ASSOCIATED WITH SIMULATED ANNEALING*

DENNIS D. COX†, ROBERT M. HARDT‡, AND PETR KLOUČEK§

Abstract. We give a sufficient condition for a sequence of Gibbs measures dominated by Lebesgue measure to converge to a singular measure concentrated on a submanifold. The limiting measure is absolutely continuous with respect to Hausdorff (Riemannian) measure on the submanifold, and a formula for its density is given. These results have implications for simulated annealing algorithms on a continuous state space when the set of minimizers of the objective function is more complex than a finite set of points. Under regularity conditions, the limiting measure is concentrated on the highest dimensional submanifold of the set of minimizers, so that lower dimensional components of the minimizing set are essentially lost. A generalization of the main result treats multiple limits within submanifolds, which could be useful for constrained optimization with simulated annealing. An example is given which shows that if the conditions of the theorem do not hold, then unexpected results may occur.

Key words. Gibbs measure, Hausdorff measure, simulated annealing, Markov chain Monte Carlo methods, differential inclusions

AMS subject classifications. 60J22, 60J25, 82C31, 82B26, 82B80, 58D20, 28A33

DOI. 10.1137/060657169

1. Introduction.

We consider some measure theoretic problems that arise in the context of simulated annealing. Simulated annealing (SA) [22], [16] is a stochastic optimization algorithm that mimics the physical process of a thermodynamic system settling into the state of minimal energy while lowering the temperature. It is usually considered in a discrete state space setting when the objective has multiple minima, but continuous state space simulated annealing has found many applications [5], [6], [21]. The version of the SA algorithm we consider here is in the Metropolis–Hastings (MH) family of algorithms, which is one of the main methods for Markov chain Monte Carlo (MCMC). We also consider continuous state space simulated annealing with equality constraints that can be expressed as the zero set of a nonnegative function. One approach for dealing with the constraints is a relaxation method wherein one adds a nonnegative multiple of the constraint function to the objective and lets the multiplier go to infinity. The work in the present paper indicates that without appropriate conditions, the procedure may not converge to the desired constrained minimum. Other authors have considered similar approaches to constrained simulated annealing (see [28], [14]).

*Received by the editors April 13, 2006; accepted for publication (in revised form) July 3, 2007; published electronically January 22, 2008.
http://www.siam.org/journals/sima/39-5/65716.html

†Department of Statistics, Rice University, 6100 Main Street, Houston, TX 77005 (dcox@stat.rice.edu). This author's work was supported by NSF grant DMS-05-05584.

‡Department of Mathematics, Rice University, 6100 Main Street, Houston, TX 77005 (hardt@rice.edu). This author's work was supported in part by NSF grant DMS-0306294.

§Texas Learning and Computation Center, University of Houston, 4800 Calhoun, TX 77204. Current address: Institute de Mathématique, Université de Neuchâtel, Rue Emile Argand 11, CH-2007 Neuchâtel, Switzerland (petr.kloucek@unine.ch). This author's work was supported in part by NSF grants ACI-0325081 and CCR-0306503, and by the European Commission via MEXC-CT-2005-023843.

We first consider an unconstrained problem. Suppose our goal is to find the set

$$\operatorname*{Argmin}_{x\in\mathbb{R}^n} J(x),$$

where J is a continuous function, bounded below, and $J(x) \to \infty$ as $|x|_{\mathbb{R}^n} \to \infty$ sufficiently fast so that

$$Z_\lambda^{-1} \;=\; \int_{\mathbb{R}^n} e^{-\lambda J(x)}\,dx \;<\; \infty$$

for some $\lambda > 0$. We may then define a probability density function

$$f_\lambda(x) \;=\; Z_\lambda e^{-\lambda J(x)}. \tag{1.1}$$

The corresponding probability distribution, P_λ, given for any Borel subset B of $\mathbb{R}^n$ by

$$P_\lambda(B) = \int_B f_\lambda\,(x)\;dx, \tag{1.2}$$

is known as a *Gibbs measure*. In a thermodynamic setting, x represents the state of the system, and $\lambda = (kT)^{-1}$, where k is Boltzmann's constant and T is temperature. Using MCMC methods it is possible to construct ergodic Markov chains $\{\boldsymbol{X}_k\}_{k=0}^\infty$ whose distribution approaches a *stationary distribution* which is P_λ [12].

The idea behind SA is to concentrate the Gibbs measure on the set of minima by letting $T \to 0$, i.e., $\lambda \to \infty$. Then the simulated values of the Markov chain will be close to a minimum with high probability. This convergence will not necessarily be uniform on the set of minima. Note that there are two limits taking place here: we must simulate a realization from the Markov chain $\{\boldsymbol{X}_{m,\lambda} : m = 1, 2, \ldots\}$ and stop at a large value of m so that the probability distribution for $\boldsymbol{X}_{m,\lambda}$ is close to P_λ. Then we increase λ (which changes the dynamics of the Markov chain) and run, say, m_1 more simulations with the new value, increase λ again, etc.

Next consider constrained optimization problems. Suppose that J_0 and J_1 are nonnegative functions satisfying conditions similar to those for J, and that we wish to find the minimizers of J_0 when x is constrained to the set

$$M \;=\; \{x \in \mathbb{R}^n \mid J_1\,(x) \;=\; 0\}\,. \tag{1.3}$$

One approach to solving the constrained minimization problem is to apply SA to the Gibbs measure P_{λ_0,λ_1} with density

$$\begin{aligned} f\,(x;\lambda_0,\lambda_1) &\overset{\text{def}}{=} Z_{\lambda_0,\lambda_1} e^{-\lambda_0 J_0(x)-\lambda_1 J_1(x)},\\ Z^{-1}_{\lambda_0,\lambda_1} &\overset{\text{def}}{=} \textstyle\int_{\mathbb{R}^n} e^{-\lambda_0 J_0(x)-\lambda_1 J_1(x)}\,dx. \end{aligned} \tag{1.4}$$

One could then let $\lambda_1 \to \infty$ so that the Gibbs measure will concentrate on the constraint set M, and then afterwards let $\lambda_0 \to \infty$ to find the minima on the constraint set M. One question we address in this article is, Under what conditions could we expect P_{λ_0,λ_1} to concentrate on the set of constrained minima as $\lambda_1 \to \infty$ and then $\lambda_0 \to \infty$? One might conjecture that as $\lambda_1 \to \infty$, the limit of P_{λ_0,λ_1} would be a measure on the constraint set M which is absolutely continuous (dominated by) a Hausdorff measure on M and the density with respect to a Hausdorff measure being a multiple of $e^{-\lambda_0 J_0}$. The first main result of this paper, stated in section 3, gives sufficient conditions for the first part of the conjecture to hold. In particular, under regularity assumptions about the constraint function J_1, the constraint set M will be a disjoint union of submanifolds, and the limiting measure will concentrate on the highest dimensional submanifold. Further, there may be nonconstant factors in the density for the limiting measure that come from the Hessian $D^2 J_1$. In section 3 we also give a generalization which treats the case of multiple limits, with the occurrence of a new factor in the limiting measure that indicates the degree of "orthogonality" of the multiple constraints on the common constraint set. Some examples of these theorems are given in section 4. Other examples in this section, which are not covered by the results of section 3, show that the support of the limiting measure P_{λ_1,λ_2} as $\lambda_1 \to \infty$ can be a subset of the constraint set which does not include the constrained minima of J_2.

Our main objective here is to initiate the development of a theory for handling these limiting Gibbs measures in complex situations. The paper is organized as follows. We discuss the SA and constrained SA in more detail in section 2 and show how the results here pertain to the limiting behavior of these algorithms. The main results are stated in section 3. In section 4 we present three examples, two of which exemplify the main theorems and one where the limit of the Gibbs measure can be determined through ad hoc methods. Further possible applications are discussed in section 5, including the problem which motivated our study, namely, solving differential inclusions which are used for modeling functional materials. The remaining sections present technical details of the proofs of the theorems.

2. Simulated annealing. We will first define the MH algorithm, which eventually generates an approximate sample from a given probability density function $f(\cdot)$ on $\mathbb{R}^n$. Consider a family of "proposal" distributions $R(x,\cdot)$, where for each fixed x, $R(x,\cdot)$ is a Borel probability measure on $\mathbb{R}^n$. Suppose that $R(x,\cdot)$ has density $r(x,\cdot)$ with respect to Lebesgue measure. Starting from some initial state $\boldsymbol{X}_0$, we will recursively generate new states. Given the current state $\boldsymbol{X}_m$, the next state $\boldsymbol{X}_{m+1}$ is generated as follows:

Step 1. Generate a "proposal for the new state," $\boldsymbol{X} \in \mathbb{R}^n$, from the measure $R(\boldsymbol{X}_m,\cdot)$.

Step 2. Compute

$$\alpha(\boldsymbol{X}_m,\boldsymbol{X}) = \frac{f(\boldsymbol{X})r(\boldsymbol{X},\boldsymbol{X}_m)}{f(\boldsymbol{X}_m)r(\boldsymbol{X}_m,\boldsymbol{X})}.$$

Step 3. Let $W \in [0,1)$ be a random number drawn from the uniform distribution on $[0,1)$. Then

$$\boldsymbol{X}_{m+1} = \begin{cases} \boldsymbol{X} & \text{if} \quad W < \alpha(\boldsymbol{X}_m,\boldsymbol{X}), \\ \boldsymbol{X}_m & \text{otherwise.} \end{cases}$$

Note that if $\alpha(\boldsymbol{X}) \geq 1$, then the proposal $\boldsymbol{X}$ is automatically accepted, and thus there is no point in even generating the W in Step 3. Under very general conditions [26] the limiting distribution as $m \to \infty$ of $\{\boldsymbol{X}_m\}_m$ will have the probability density function f. This convergence takes place in total variation norm on measures at an exponential rate.

If the proposal density r satisfies a "reversibility" property, i.e., $r(x,y) = r(y,x)$, $x,y \in \mathbb{R}^n$, then the algorithm is identical to the original Metropolis algorithm and

$\alpha(\boldsymbol{X}) = f(\boldsymbol{X})/f(\boldsymbol{X}_m)$. An example of a reversible proposal distribution is any Gaussian distribution with the mean x, e.g.,

$$r(x, y) = (2\pi\sigma^2)^{-n/2}\ e^{-|x-y|^2/(2\sigma^2)}, \tag{2.1}$$

where $|x|$ indicates the norm of the vector $x \in \mathbb{R}^n$.

Now we describe an SA algorithm. Consider a sequence $\lambda_k \to \infty$. Starting with $k = 1$, replace f above by f_{λ_k} as given in (1.1) and run the MH algorithm until $\boldsymbol{X}_{m_k}$ has a distribution which is sufficiently close to the one with density f_{λ_k}. Now, increment k and start with $\boldsymbol{X}_{m_k}$ and repeat the MH algorithm. As long as we require that the total variation distance between the distribution of $\boldsymbol{X}_{m_k}$ and the distribution with density f_{λ_k} tends to 0 as $k \to \infty$, then it is clear that the limiting distribution of $\boldsymbol{X}_m$ will be the same as the limiting distribution of the Gibbs measures. This motivates our study of the limits of these Gibbs measures. There are various papers that consider the convergence of SA (e.g., [15], [13]). The basic idea is that $\boldsymbol{X}_m$ will converge to a global minimum of J as long as the temperature goes to 0 but not too fast.

Assuming a reversible proposal distribution for the SA algorithm described above, we see that a proposal $\boldsymbol{X}$ is automatically accepted if $f_{\lambda_k}(\boldsymbol{X}) \geq f_{\lambda_k}(\boldsymbol{X}_m)$, i.e., if the proposed $\boldsymbol{X}$ has value of the objective no larger than the current value, and the proposal is accepted with some positive probability if $0 < f_{\lambda_k}(\boldsymbol{X}) < f_{\lambda_k}(\boldsymbol{X}_m)$. This latter type of uphill step allows the Markov chain to avoid getting stuck in local minima. As the temperature go to 0 (i.e., $\lambda_k \to \infty$), the probability distribution for the state of the system becomes concentrated near the global minima of J. Keeping the temperature from cooling too quickly avoids getting stuck in local minima.

For constrained optimization problems, we can consider multiple temperatures as in (1.4). We then modify the SA algorithm so that both $\lambda_0 \to \infty$ and $\lambda_1 \to \infty$, slowly enough so that the distribution of each X_m is close to the appropriate Gibbs measure with density $f(\cdot; \lambda_0, \lambda_1)$. This methodology has been proposed in the discrete setting in [28]. The results stated in the next section show that the Gibbs measure will converge to a constrained minimum under certain conditions, and an example given in section 4 shows that this convergence can fail if the conditions of the theorems in section 3 are not valid.

3. Statement of the main results. For our first result, we consider a slight generalization of a Gibbs measure with density as in (1.1). For $\lambda > 0$ and nonnegative continuous functions h, J on $\mathbb{R}^n$ we assume that

$$Z_\lambda^{-1} \ = \ \int_{\mathbb{R}^n} h(x) e^{-\lambda J(x)}\, dx \ < \ \infty,$$

and then we have a probability density function

$$f(x; \lambda) \ = \ Z_\lambda h(x) e^{-\lambda J(x)} \tag{3.1}$$

and corresponding probability measure

$$P_\lambda(B) \ = \ \int_B f(x; \lambda)\, dx,$$

defined for Borel measurable subsets B of $\mathbb{R}^n$. We investigate the behavior of P_λ as $\lambda \to \infty$.

We will use *weak convergence of probability measures.* Suppose $\{P_m : m \in \mathbb{Z}\}$ is a sequence of probability measures and P is a fixed probability measure, all defined on the Borel sets $\mathcal{E}$ of a given Polish space E. *Weak convergence of* P_m *to* P, denoted $P_m \Rightarrow P$, means that

$$\int \phi dP_m \to \int \phi dP \qquad \text{for all bounded continuous real valued functions } \phi. \tag{3.2}$$

We note that $P_m \Rightarrow P$ is equivalent to the following (cf. Theorem 2.1 of [3]):

$$P_m(B) \to P(B) \qquad \text{for all Borel sets of } E \text{ such that } P(\partial B) = 0. \tag{3.3}$$

Recall that the probability distributions in the MH algorithm converge in total variation norm to the stationary distribution. However, convergence in total variation is not possible here since the limiting measure is mutually singular with each of the Gibbs measures. Convergence in total variation means $\sup_B |P_n(B) - P(B)| \to 0$.

For a function ψ with domain and range in some Euclidean spaces, we say $\psi \in C^{k,\alpha}(\Omega)$ (where $k \geq 0$ is an integer and $\alpha \in (0,1]$) if $D^k\psi$ exists, is continuous in Ω, and satisfies a uniform Hölder condition

$$\left|D^k\psi(x) - D^k\psi(y)\right| \;\leq\; C\,|x-y|^\alpha \quad \text{for} \quad x,\, y \,\in\, \Omega.$$

We say that an n-dimensional manifold M is $C^{k,\alpha}$ if it has a $C^{k,\alpha}$-atlas, i.e., a partition into finitely many open sets (in the relative topology of M), and each open set in the partition is homeomorphic with $\mathbb{R}^n$ by a $C^{k,\alpha}$ homeomorphism whose inverse is also $C^{k,\alpha}$.

Now we state our first result.

THEOREM 3.1. *Assume the following*:

(A1) $h \in C^0(\mathbb{R}^n)$ *is bounded, and* $J \in C^{3,\alpha}(\mathbb{R}^n)$.

(A2) $h \geq 0$, *and* $J \geq 0$.

(A3) *For some* $p > 0$, $J(x) \geq \|x\|^p$ *for all* $\|x\|$ *sufficiently large.*

(A4) $M = \{x \in \mathbb{R}^n : J(x) = 0\}$ *is nonempty and bounded.*

(A5) *There exist a bounded disjoint open set* $U_1, U_2, \ldots, U_j$ *and integers* $0 \leq m_1 < m_2 < \cdots < m_j \leq n$ *satisfying the following*:

- (a) $M \subset U = U_1 \cup U_2 \cup \cdots \cup U_j$.
- (b) *Each set* $M_i = M \cap U_i$ *is a* $C^{2,\alpha}$ *smooth* m_i*-dimensional manifold.*
- (c) *On each* M_i, *the Hessian* D^2J *is nonnegative definite and has constant rank* $n - m_i$ *for* $i = 1, \ldots, j$.
- (d) *For some* $a \in M_1$, $h(a) > 0$.

Let $\mathcal{H}^{n-k_1}$ *be the* $(n-k_1)$*-dimensional Hausdorff measure on* M_1, *and let* $\Lambda(a)$ *be the product of the* k_1 *positive eigenvalues of* $D^2J(a)$ *for* $a \in M_1$. *Then*

$$P_\lambda \;\Rightarrow\; P \text{ as } \lambda \to \infty,$$

where, for any Borel set $B \subset \mathbb{R}^n$,

$$P(B) = Z \int_{M_1 \cap B} h(a)\,\Lambda(a)^{-1/2}\, d\mathcal{H}^{n-k_1}(a)$$

with

$$Z^{-1} = \int_{M_1} h(a)\,\Lambda(a)^{-1/2}\, d\mathcal{H}^{n-k_1}(a)\,.$$

Observe that, as $\lambda \to \infty$, the probability measures P_λ concentrate only on the highest dimensional stratum M_1 of $M = J^{-1}\{0\}$ and do not produce any lower dimensional measures on $M_2 \cup \cdots \cup M_j$.

Sections 6–8 gather material necessary for the proof of Theorem 3.1 in section 9. We note that special cases of this result are certainly well known; in particular, if M is zero-dimensional, i.e., a finite set of points $\{x_1, \ldots, x_m\}$, then by Laplace's method of asymptotic expansion [4], it is easy to see that the Gibbs measures will converge weakly to the probability measure on the discrete set of points with probabilities proportional to $h(x_i) \det D^2 J(x_i)$, $1 \le i \le m$.

Next we discuss a generalization of the main result to treat multiple limits. First, the ambient space $\mathbb{R}^n$ is replaced by a compact n-dimensional Riemannian manifold N. Consider probability measures of the form

$$dP_{\lambda_1\lambda_2\cdots\lambda_j} \;=\; Z_{\lambda_1\lambda_2\cdots\lambda_j} h(x) e^{-\lambda_1 J_1(x) - \lambda_2 J_2(x) - \cdots - \lambda_j J_j(x)}\, d\mathcal{H}^n(x). \tag{3.4}$$

Here the Hausdorff measure $\mathcal{H}^n$ is the standard Riemannian volume measure of N. We consider limits of the form

$$\lim_{\lambda_1\to\infty} \lim_{\lambda_2\to\infty} \cdots \lim_{\lambda_j\to\infty} P_{\lambda_1\lambda_2\cdots\lambda_j}.$$

If $J : N \longrightarrow \mathbb{R}$ is a smooth function, then the Riemannian Hessian of J at $a \in N$ is the bilinear form on the tangent space $T_a N$ whose value at (v, v) for $v \in T_a N$ is simply the initial second derivative of J along a geodesic starting at a with initial velocity v. In case the x_i are geodesic normal coordinates at a, this Hessian is represented at a by the usual matrix $\frac{\partial^2}{\partial x_i \partial x_j} J(a)$. We will also need the properties of the wedge product [25], specifically that if $v_1, \ldots, v_k$ are n-dimensional vectors, then $|v_1 \wedge \cdots \wedge v_k|$ is the k-dimensional volume of the k-dimensional parallelopiped $\{\sum_{i=1}^k t_i v_i : 0 \le t_i \le 1\}$.

THEOREM 3.2. *Suppose that $p > 0$ and that, for $i = 0, \ldots, j$, J_i is a nonnegative $C^{3,\alpha}$ function on N. Suppose also that each set $M_i = J_i^{-1}\{0\}$ is a $C^{3,\alpha}$ smooth compact $(n - k_i)$-dimensional submanifold on which the Riemannian Hessian $D^2 J_i$ is a nonnegative definite operator with range E_i of constant dimension k_i. With $k = k_1 + \cdots + k_j$, we also assume that the M_i intersect transversally (i.e., the E_i are linearly independent), giving the $(n - k)$-dimensional manifold*

$$M \;=\; \{x \in N : J_1(x) = J_2(x) = \cdots = J_j(x) = 0\}$$

on which the total Hessian $D^2(J_1 + \cdots + J_j)$ has range E of constant dimension k. Assume $h(a) > 0$ for some $a \in M$.

For $\lambda_1, \ldots, \lambda_j \ge 1$, let $P_{\lambda_1\cdots\lambda_j}$ be the probability measure on N given in (3.4), *where $h(x)$ is a bounded continuous function. Then*

$$P_{\lambda_1\cdots\lambda_j} \Rightarrow P \quad \text{as} \quad \lambda_1, \ldots, \lambda_j \to \infty, \ \textit{regardless of the order},$$

where, for any Borel set $B \subset \mathbb{R}^n$,

$$P(B) \;=\; Z \int_{M \wedge B} e^{-J_0(a)}\, \Lambda_1(a)^{-1/2} \cdots \Lambda_j(a)^{-1/2} \Theta(a)\, d\mathcal{H}^{n-k}(a).$$

Here

$$Z^{-1} \;= \int_M e^{-J_0}\, \Lambda_1^{-1/2} \cdots \Lambda_j^{-1/2} \Theta\, d\mathcal{H}^{n-k},$$

$\Lambda_i(a)$ *is the product of the* k_i *positive eigenvalues of* $D^2 J_i(a)$, *and*

$$\Theta(a) \;=\; |v_1^1 \wedge \cdots \wedge v_{k_1}^1 \wedge \cdots \wedge v_1^j \wedge \cdots \wedge v_{k_j}^j|,$$

where $\{v_1^i, \ldots, v_{k_i}^i\}$ *is an orthonormal basis for* $E_i(a)$ *for* $i = 1, \ldots, j$.

The proof of this result is given in section 9.2.

4. Examples. We will consider in depth three examples. The first is an application of Theorem 3.1, and the second of Theorem 3.2. The third example is similar to the first, but the hypotheses of Theorem 3.1 are not met, and using ad hoc calculations we show that the limiting measure is not the desired one. This example also shows that the order of taking the limits in a setting like Theorem 3.2 can matter when the hypotheses are violated.

4.1. Example 1. Consider the following two functions on $\mathbb{R}^2$:

$$J_1(x) \stackrel{\text{def}}{=} \left(|x|^2 - 1\right)^2, \tag{4.1}$$

$$J_2(x) \stackrel{\text{def}}{=} \begin{cases} x_1^2 x_2^2 & \text{if } |x|^2 < 2, \\ x_1^2 x_2^2 + \left(|x|^2 - 2\right)^4 & \text{if } |x|^2 \geq 2. \end{cases} \tag{4.2}$$

Let

$$g(x; \lambda_1, \lambda_2) \stackrel{\text{def}}{=} e^{-\lambda_1 J_1(x) - \lambda_2 J_2(x)},$$

$$Z_{\lambda_1,\lambda_2}^{-1} \stackrel{\text{def}}{=} \int_{\mathbb{R}^2} g(x; \lambda_1, \lambda_2) dx,$$

and let us define the normalized densities and the corresponding Gibbs measures as follows:

$$f(x; \lambda_1, \lambda_2) \stackrel{\text{def}}{=} Z_{\lambda_1,\lambda_2}\, g(x; \lambda_1, \lambda_2), \tag{4.3}$$

$$R(B; \lambda_1, \lambda_2) \stackrel{\text{def}}{=} \int_B g(x; \lambda_1, \lambda_2) dx, \tag{4.4}$$

$$P(B; \lambda_1, \lambda_2) \stackrel{\text{def}}{=} \int_B f(x; \lambda_1, \lambda_2) dx. \tag{4.5}$$

In the above definitions, B denotes a Borel set in $\mathbb{R}^2$.

Consider the limit $\lambda_1 \to \infty$ with λ_2 fixed. This could arise if we consider simulated annealing for the constrained optimization problem

$$\operatorname{Argmin}\{J_2(x) \,|\, x \in M\}, \qquad \text{where } M = \{x \in \mathbb{R}^2 \,|\, J_1(x) = 0\}.$$

We can apply Theorem 3.1 to find the limiting Gibbs measure in this case. Assumptions (A1)–(A3) and (A5) clearly hold. The set M where $J_1 = 0$ is the unit circle centered at the origin, S_1; hence (A4) is satisfied.

The Hessian of J_1 is given by

$$D^2 J_1(x_1, x_2) = \begin{pmatrix} 4\left(3x_1^2 + x_2^2 - 1\right) & 8x_1 x_2 \\ 8x_1 x_2 & 4\left(x_1^2 + 3x_2^2 - 1\right) \end{pmatrix}.$$

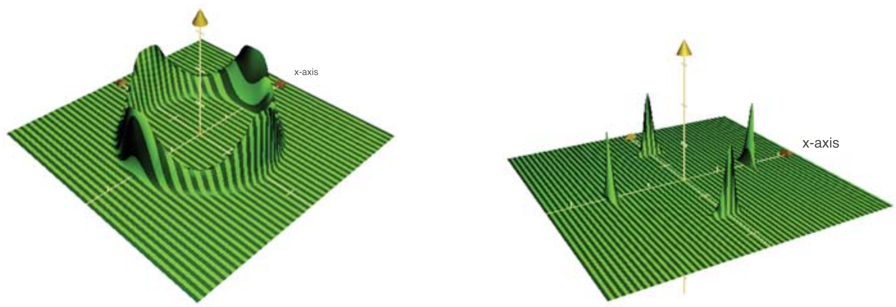


FIG. 1. *Plot of the unnormalized Gibbs density* $e^{-(\lambda_1 J_1(x)+\lambda_2 J_2(x))}$ *in Example* 1. *The picture on the left is a plot with* $\lambda_1 = 60$ *and* $\lambda_2 = 5$. *The picture on the right is a plot with* $\lambda_1 = 460$ *and* $\lambda_2 = 50$.

In particular,

$$D^2 J_1(x_1, 0) = \begin{pmatrix} 4\left(3x_1^2 - 1\right) & 0 \\ 0 & 4\left(x_1^2 - 1\right) \end{pmatrix}.$$

Thus, along the x_1-axis the eigenvalues are given by the diagonal entries, and the eigenvectors are the corresponding coordinate vectors. Noting that J_1 is rotationally symmetric, we see that

$$\operatorname{rank} D^2 J_1(x) = \begin{cases} 2 & \text{if } |x|^2 \neq 1 \text{ or } \frac{1}{3}, \\ 1 & \text{otherwise,} \end{cases}$$

and $D^2 J_1(x)$ is nonnegative definite for $|x|^2 \geq 1/3$. Thus in assumption (A5), $j = 1$ and we may take $U = U_1 = \{x : 1/\sqrt{3} < |x| < \sqrt{2}\}$, and the rank of the Hessian on $M = M_1$, which is the unit circle, is $k_1 = 1$. Consequently, it follows from Theorem 3.1 that

$$P(\cdot; \lambda_1, \lambda_2) \Rightarrow P_1(\cdot; \lambda_2) \qquad \text{as } \lambda_1 \to +\infty,$$

where $P_1(\cdot; \lambda_2)$ has a density with respect to $\mathcal{H}^1$ on S_1 which is proportional to $e^{-\lambda_2 J_2(x)}$. Here, $\mathcal{H}^1$ may be thought of as "arc-length measure."

It is easy to check that

$$P_1(\cdot; \lambda_2) \Rightarrow \frac{1}{4} \sum_{x \in \{(0,\pm 1),(\pm 1,0)\}} \delta_x(\cdot) \qquad \text{as } \lambda_2 \to +\infty, \tag{4.6}$$

which is the uniform probability distribution on the four points that minimize J_2 subject to $J_1 = 0$. This is clearly the desirable outcome in this case. Figure 1 illustrates this convergence.

4.2. Example 2. Here we give an example illustrating Theorem 3.2. Letting M be some compact two-dimensional Riemannian manifold which contains a subset isometric to the square $[-3, 3] \times [-3, 3] \subset R^2$, we will simply use the coordinate system of the latter set. Let J_1 be as in (4.1) and let

$$J_3(x) \ \stackrel{\text{def}}{=}\ \left((x_1 - 1)^2 + 2x_2^2 - 3x_2\right)^2.$$

Using the linear isomorphism

$$L(x_1, x_2) = (x_1 - 1, 2(x_2 + 3/4)),$$

we see that $J_3(x) = J_4(L(x))$, where $J_4(y) = (|y|^2 - 3/2)^2$. Since the rank of the Hessian is invariant under a linear transformation, we see as in Example 1 in section 4.1 that the rank of the Hessian $D^2 J_3$ equals 1 on the ellipse $J_3^{-1}\{0\}$. This ellipse crosses the unit circle $M_1 = J_1^{-1}\{0\}$ orthogonally at $a = (1, 0)$ with slope 0, and nonorthogonally at $a = (0, 1)$ with slope 2. In fact, for a smooth function on a planar region, the slope of $u^{-1}\{0\}$ at a point with $\frac{\partial u}{\partial x_2} \neq 0$ is given by the formula

$$-\frac{\partial u/\partial x_1}{\partial u/\partial x_2},$$

and we may apply this with $u(x_1, x_2) = (x_1 - 1)^2 + 2x_2^2 - 3x_2$. So we may compute

$$\Theta(a) \ = \ |(0,1) \wedge (1,0)| \ = \ 1, \qquad \Theta(b) \ = \ |(1,0) \wedge (2/\sqrt{5}, 1/\sqrt{5})| \ = \ 1/\sqrt{5}\ .$$

Thus, if P_{λ_1,λ_3} is the corresponding Gibbs measure as in Theorem 3.2, then

$$P_{\lambda_1,\lambda_3} \ \Rightarrow \ \frac{\sqrt{5}}{1+\sqrt{5}}\delta_a \ + \ \frac{1}{1+\sqrt{5}}\delta_b \ \ \textit{as}\ \lambda_1, \lambda_3 \to \infty.$$

Note that, unlike in the previous example, we do not obtain a uniform distribution on the set of constrained minima. If SA is applied in this case, we are more likely to find the point a than b. The nonuniform distribution here comes from the factor in Theorem 3.2 that comes from the factor Θ, but one could also obtain such nonuniformity from the factors Λ_i. While this does not affect the convergence of SA to a minimum, one may wish to explore the set of all minima, and a nonuniform distribution would then not be desirable.

4.3. Example 3. We use the same functions (4.1)–(4.3), but now we let $\lambda_2 \to \infty$ holding λ_1 fixed. This would arise from applying SA to a constrained optimization problem in which the objective from the previous example becomes a constraint and vice versa, i.e., if we attempt to solve

$$\mathrm{Argmin}\{J_1(x) \,|\, x \in M\}, \qquad \text{where } M = \{x \in \mathbb{R}^2 \,|\, J_2(x) = 0\}.$$

We cannot apply Theorem 3.1 to this situation because assumption (A5) is not valid. (Note that our complicated expression for J_2 is constructed so that $\int \exp[-\lambda J_2] < \infty$ for all $\lambda > 0$ and J_2 is $C^{3,\alpha}$.) The Hessian of J_2 is given by

$$D^2 J_2(x) = \begin{pmatrix} 2x_2^2 & 4x_1x_2 \\ 4x_1x_2 & 2x_1^2 \end{pmatrix} \qquad \text{if } |x|^2 \leq 2.$$

Its eigenvalues are

$$x^2 + y^2 \pm \sqrt{x^4 + 14y^2x^2 + y^4}.$$

Consequently, for $|x|^2 < 2$,

$$\operatorname{rank} D^2 J_2(x_1, x_2) = \begin{cases} 2 & \text{if } x_1 \neq 0 \text{ and } x_2 \neq 0, \\ 1 & \text{if } x_1 = 0 \text{ or } x_2 = 0 \text{ and } x_1 \neq x_2, \\ 0 & \text{if } x_1 = x_2 = 0. \end{cases}$$

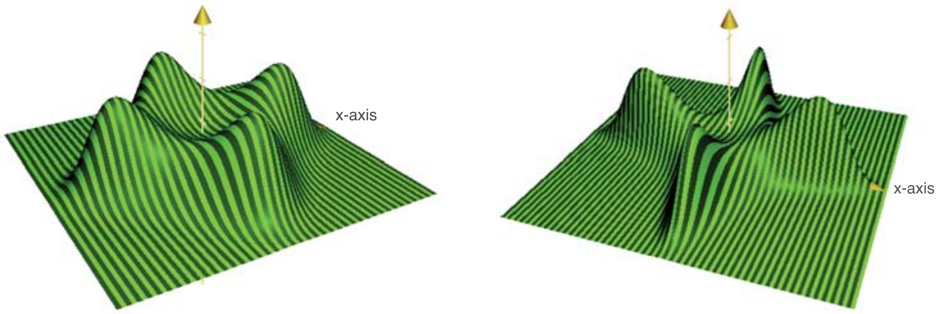


FIG. 2. *Plot of the unnormalzied Gibbs density* $e^{-(\lambda_1 J_1(x_1,x_2)+\lambda_2 J_2(x_1,x_2))}$ *in Example* 3. *The picture on the left is a plot with* $\lambda_1 = 1$ *and* $\lambda_2 = 5$, *and the one on the right is with* $\lambda_1 = 1$ *and* $\lambda_2 = 30$.

Since the set $M = J_2^{-1}(0)$ consists of the segments of the x and y axes for which $|x| \le 2$, the rank of the Hessian drops from being 1 on $M \setminus \{0\}$ to 0 at the origin. Figure 2 suggests that the limiting measure is concentrated along the axes near the desired constrained minima. However, this picture is misleading, as we shall prove.

Let B be a neighborhood of $0 \in \mathbb{R}^2$ of the form

$$B = (-\epsilon, \epsilon) \times (-\epsilon, \epsilon).$$

We will show that

$$R(B^c; \lambda_1, \lambda_2) = o(R(B; \lambda_1, \lambda_2)) \qquad \text{as } \lambda_2 \to \infty, \tag{4.7}$$

where R is the unnormalized Gibbs measure in (4.4). From this it easily follows that

$$P(\cdot; \lambda_1, \lambda_2) \Rightarrow \delta_0 \qquad \text{as } \lambda_2 \to \infty.$$

To estimate the left-hand side of (4.7), define

$$Q(\lambda_2) = \int_0^\epsilon \int_0^\epsilon e^{-\lambda_2 x_1^2 x_2^2}\, dx_1\, dx_2.$$

Then clearly

$$R(B; \lambda_1, \lambda_2) \asymp Q(\lambda_2),$$

where we use $\asymp$ to indicate that the left-hand side is bounded above and below by finite positive multiples of the right-hand side, where the constants do not depend on λ_2. By direct calculation

$$Q(\lambda_2) = \tfrac{\sqrt{\pi}}{2\sqrt{\lambda_2}} \int_0^{\sqrt{\lambda_2}\epsilon^2} \tfrac{1}{v}\, \mathrm{Erf}(v) dv, \qquad \text{where} \qquad \mathrm{Erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-x^2} dx,$$

is the Error function [1]. Using the properties that $\mathrm{Erf}(z) \sim \frac{2}{\sqrt{\pi}} z$ as $z \to 0$ and $\mathrm{Erf}(z) \to 1$ as $z \to \infty$, it follows that for $\lambda_2 > \epsilon^{-4}$,

$$R(B; \lambda_1, \lambda_2) \asymp \lambda_2^{-1/2} \log \lambda_2. \tag{4.8}$$

It is clear from the proof of Theorem 3.1 (see (9.2)–(9.5)) that $R(B^c;\lambda_1,\lambda_2)$ is asymptotically dominated by the $R(\cdot;\lambda_1,\lambda_2)$ measure of the tubular neighborhoods of the intersection of the coordinate axes with B^c and the disk $|x|^2 \le 2$, and that

$$R(B^c;\lambda_1,\lambda_2) \asymp \lambda_2^{-1/2}. \tag{4.9}$$

Combining (4.8) and (4.9) gives the result (4.7).

As seen in the first example, we can take $\lambda_1 \to \infty$ and then $\lambda_2 \to \infty$ and get the "right answer," but the calculation here shows that we cannot take the limits in the other order and get the desired result. Note that Theorem 3.2 does not apply here either for the same reason as for Theorem 3.1: the appropriate Hessian does not have constant rank on the required set.

5. Further applications and discussion.

5.1. Applications to materials science. Differential inclusions are a modeling methodology that has found application in material science. Unlike a differential equation where an exact relationship for the derivative is given, in a differential inclusion the derivative is required to lie in a specified set. In the materials science applications, the set corresponds to microscale crystalline configurations and typically consists of a finite number of allowed gradients (first derivatives) and all rotations of them. For more material on differential inclusions, including theory, methods, and application, see [2], [19], [8], [18], [24], [23], [9], and the references therein.

In this section, we consider a simple example which has almost all of the features of the models used in practice. Let $\Omega = [0,1]^2$ be the unit square in $\mathbb{R}^2$. Suppose that we wish to find a weakly differentiable function $u : \Omega \to \mathbb{R}$ satisfying the following conditions:

(i) u is continuous (continuity condition);
(ii) $u(x) = g(x)$ for $x \in \partial\Omega$, where g is given (boundary condition);
(iii) $|\nabla u| = 1$ in Ω (gradient constraint).

If a solution exists, it can be found by solving the constrained optimization problem

$$\min \int_{\partial\Omega} (g(s) - u(s))^2\, dS \quad \text{subject to } |\nabla u| = 1 \ \text{ a.e. in } \Omega \ \text{ and } \ u \in C^0(\Omega).$$

This is a nonconvex minimization problem with numerous local minima (see, e.g., [7]); thus SA is a reasonable approach. We consider a finite dimensional approximation u_h which is piecewise affine with different parameters in each element of a triangular partition of Ω. For simplicity, we first subdivide Ω into squares, then subdivide the squares by a diagonal (cf. Figure 4). We will suppose that the continuity condition is enforced, but the gradient constraint is relaxed. With continuity the function u_h in one of the subdivision squares is determined by its values at the four corners of the square; denote them by $v = (v_1, v_2, v_3, v_4)$, as indicated in Figure 4. An SA approach of the type discussed in section 2 could be based on first using the modified objective $\lambda_1 J_1(u) + \lambda_2 J_2(u)$, where J_1 and J_2 correspond to the boundary and gradient constraints, respectively, e.g.,

$$J_1(u_h) = \int_{\partial\Omega} (g(s) - u_h(s))^2\, dS, \tag{5.1}$$

$$J_2(u_h) = \int_{\Omega} \left(|\nabla u_h(x)|^2 - 1\right)^2 dx. \tag{5.2}$$

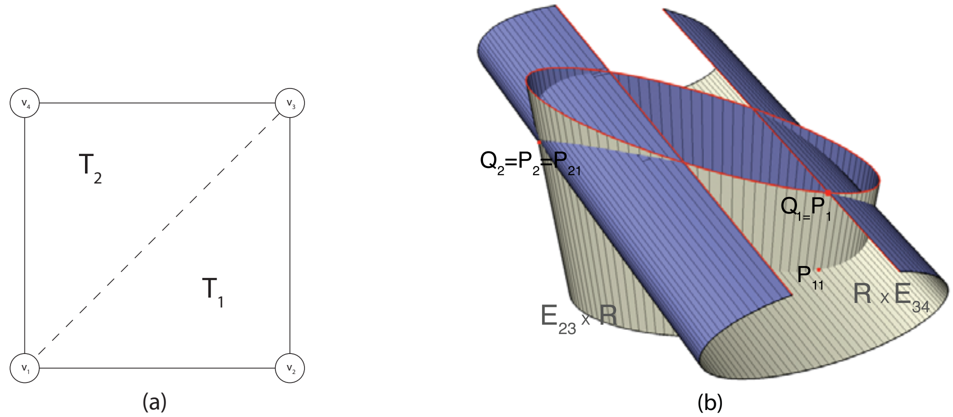


FIG. 3. (a) *The left plot indicates the labeling for a square element in the subdivision of* Ω. (b) *The plot on the right shows* $(E_{23} \times \mathbb{R}) \cap (\mathbb{R} \times E_{34})$ *corresponding to the gradient constraints in each of the two triangular subregions. At the points* Q_1 *and* Q_2 *the Hessian of the gradient constraint functional drops in rank.*

Now we show that this Lagrangian formulation has the same problems as the example in section 4.3, namely, that assumption (A5) of Theorem 3.1 is violated. Consider a single square with its two triangular elements in the partition of Ω. For simplicity, we assume that the side length is 1. If we set $v_1 = 0$, then we imbed the problem in a visualizable three-dimensional space. Consider the set of values (v_2, v_3, v_4) satisfying the gradient constraint in the lower triangular element, namely,

$$|\nabla u_h|^2 = |(v_2, v_3 - v_2)|^2 = 1.$$

Define an ellipse in the (v_2, v_3) coordinate plane

$$E_{23} = \{(v_2, v_3) : |(v_2, v_3 - v_2)|^2 = 1\}.$$

Then the set of values (v_2, v_3, v_4) satisfying the gradient constraint in the lower triangular element is given by the cylindrical surface $E_{23} \times \mathbb{R}$. We can parameterize the surface with $(\alpha, z) \in [0, 2\pi) \times \mathbb{R}$; i.e., the surface $E_{23} \times \mathbb{R}$ is the graph of the map

$$(\alpha, z) \mapsto (\cos(\alpha), \sin(\alpha) + \cos(\alpha), z), \qquad \alpha \in [0, 2\pi), \quad z \in \mathbb{R}.$$

Similarly, we can define an ellipse

$$E_{34} = \{(v_3, v_4) \mid |(v_3 - v_4, v_4)|^2 = 1\},$$

and then the set of values (v_2, v_3, v_4) satisfying the gradient constraint in the upper triangular element is given by the cylindrical surface $\mathbb{R} \times E_{34}$. The corresponding parameterization is (β, y), where $\mathbb{R} \times E_{34}$ is the graph of the map

$$(\beta, y) \mapsto (y, \sin(\beta) + \cos(\beta), \sin(\beta)), \qquad \beta \in [0, 2\pi), \quad y \in \mathbb{R}.$$

The intersection of these two cylindrical surfaces is a pair of ellipses which intersect as depicted by the right-hand panel in Figure 3. Note that the continuity constraint means that the gradient in the upper triangular region must be identical with the

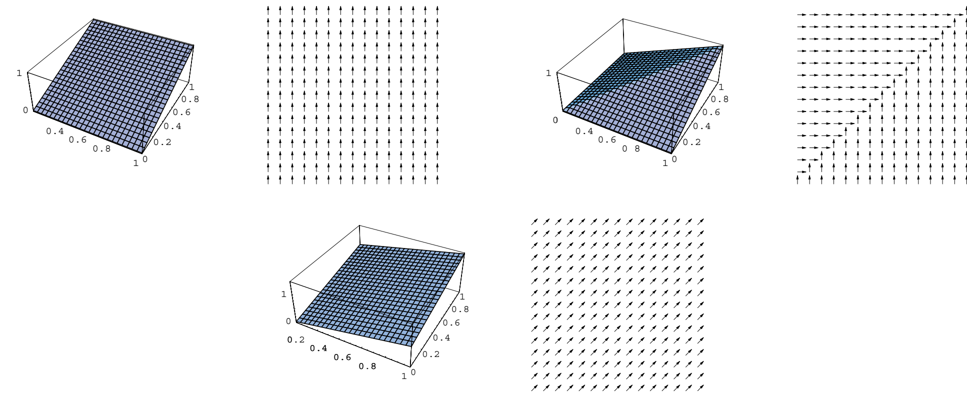


FIG. 4. *Three possible solutions for the differential inclusion $|\nabla u_h| = 1$ and $u_h \in C^0(\Omega)$ which are piecewise affine in the triangular elements.*

gradient in the lower region or else it has to be the reflection about the diagonal of the gradient in the lower region. Hence, either $\beta = \alpha$ and then we obtain a single affine function throughout the square or $\beta = \pi/2 - \alpha \pmod{2\pi}$ and then we obtain a piecewise affine solution with different parameters in the upper and lower triangular elements. For a given $\alpha \in [0, 2\pi)$ (which specifies a unit length gradient in the lower triangular element), there will typically be two values of β (corresponding to a unit length gradient in the upper triangular element), except when the gradient is parallel with the diagonal boundary, i.e., $\alpha = \pi/4$ and $\alpha = 5\pi/4$. Examples of such continuous piecewise affine functions with unit length gradients are shown in Figure 4. The points $\alpha = \pi/4$ and $\alpha = 5\pi/4$ correspond to the points Q_2 and Q_1 in Figure 3(b), respectively. The corresponding u_h with $\alpha = \pi/4$ and its gradient field are given in the last plot in Figure 4 where necessarily $\beta = \pi/4$ for the gradient in the upper triangular element.

The analogy to Example 3 in section 4.3 is as follows. The above discussion shows that there exist points in the constraint set which are the intersections of two smooth one-dimensional manifolds. Since the sum of the rank of the Hessian and the dimension of the tangent space of the manifold is bounded above by the dimension of the ambient space, we know the rank of the Hessian is ≤ 2 at all points on the constraint set, but ≤ 1 at Q_1 and Q_2. Thus, this example will clearly violate the assumptions of our theorems. Using complicated arguments based on elementary considerations (symmetries of the constraint set and properties of quartic polynomials), we can show that the Hessian has rank 1 at Q_1 and Q_2, and at some points outside of Q_1 and Q_2 (namely, the other vertices of the ellipses), it has rank 2. Based on numerical calculations, we conjecture that it has rank 2 at all points except Q_1 and Q_2, similarly to Example 3 in section 4.3. A natural conjecture would be that the same phenomenon that occurred in Example 3 will occur here: as the multiplier corresponding to the gradient constraint $\lambda_2 \to \infty$, the limiting measure will concentrate on the points $\alpha = \pi/4$ and $\alpha = 5\pi/4$. However, the same type of reasoning suggests that along boundaries between triangular elements in different squares, there will be a tendency for the gradient to align itself with the direction of the boundary. Because of all of these interactions, we do not know how this affects the behavior of an SA solution to this problem, but it does suggest that there will be difficulties with such an approach. Clearly further extension of the results given here is necessary to provide insight. The strain densities described in [19], [20], [18], [17], and [2] provide a similar implementation of the gradient constraints for the microscale modeling of crystalline materials. We believe that these will have similar problems.

5.2. Potential application to Bayesian statistics. While SA has been our prime motivation for developing the results in section 3, we believe they may have implications for other applications of MCMC, and specifically for Bayesian statistics. In the Bayesian framework, as one obtains more samples, the negative log likelihood in the neighborhood of the true value of the parameter behaves like $nJ_1(\theta)$, where n is the sample size, θ is a variable value of the parameter, and J_1 is a fixed function (see [27]) which is minimized at the true parameter value. Thus, the posterior density is approximated by

$$g_n(\theta) \;=\; Z_n \exp[-J_0(\theta) - nJ_1(\theta)],$$

where J_0 is the negative logarithm of the prior density. This approximation is valid as long as there is a unique true value of the parameter. Bayesian methods are being applied to more and more complicated models, and some models may lack identifiability; i.e., there are multiple parameter values giving the same likelihood. In regular cases, the set of possible true values will be a manifold, and Theorem 3.1 may prove useful for obtaining asymptotic approximations of the posterior.

5.3. Generalizations. Here we mention some generalizations that would be desirable. The assumption of Theorem 3.1 we have found most problematic is that the set M must be a smooth compact manifold. We believe that it should be possible to extend the results to more general sets M, possibly with restrictions on J and h. Developing a more general theory that would cover examples similar to the one in section 4.3 would also be of interest.

6. Nearest point projection for a submanifold. Recall that, for any Borel set $A \subset \mathbb{R}^n$ (or even any metric space A) and any number $k \geq 0$, the k-dimensional Hausdorff measure $\mathcal{H}^k(A)$ is defined [11, sect. 2.10.2]. It is normalized so that, for integer k in $\mathbb{R}^k$, $\mathcal{H}^k$ coincides with the k-dimensional Lebesgue measure. In a higher dimensional $\mathbb{R}^n$, the restriction of $\mathcal{H}^k$ to a k-dimensional C^1 submanifold M coincides with the Riemannian volume measure on M for the metric induced from $\mathbb{R}^n$. In particular, a k-dimensional ball of radius r in $\mathbb{R}^k$,

$$\mathbb{B}_r^k(a) \;\equiv\; \{x \in \mathbb{R}^k \,:\, |x-a| < r\}\ ,$$

has

$$\mathcal{H}^k(\mathbb{B}_r^k(a)) \;=\; \alpha_k r^k, \tag{6.1}$$

where α_k is the k-dimensional Lebesgue measure of the unit ball in $\mathbb{R}^k$.

Our notation for an integral with respect to a (lower dimensional) Hausdorff measure will have the form

$$\int_A f(a)\, d\mathcal{H}^k a \qquad \text{or} \qquad \int_A f\, d\mathcal{H}^k,$$

while our integrals with respect to the (top dimensional) Lebesgue measure will keep the standard notation

$$\int_U f(x)\, dx \qquad \text{rather than} \qquad \int_U f(x)\, d\mathcal{H}^n x.$$

In particular, we have the polar coordinate formula for a Lebesgue integrable function f on the ball $\mathbb{B}_R(0) \equiv \mathbb{B}^n_R(0)$,

$$\int_{\mathbb{B}_R(0)} f(x)\,dx \;=\; \int_{S^{n-1}} \int_0^R f(r\omega)\, r^{n-1}\, dr\, d\mathcal{H}^{n-1}\omega,$$

where S^{n-1} is the $(n-1)$-dimensional unit sphere in $\mathbb{R}^n$. One readily checks that

$$\mathcal{H}^{n-1}(S^{n-1}) \;=\; n\,\alpha_n \tag{6.2}$$

by differentiating (6.1).

For any vector subspace T of $\mathbb{R}^n$, the orthogonal projection

$$\Pi_T \;:\; \mathbb{R}^n \;\to\; T$$

is the linear map which takes any point $x \in \mathbb{R}^n$ to the unique point $\Pi_T(x)$ in T that is *nearest* to x.

Suppose that M is a compact m-dimensional $C^{2,\alpha}$ submanifold of $\mathbb{R}^n$. Then the m-dimensional *tangent space* T_aM and $(n-m)$-dimensional *normal space* $(T_aM)^\perp$ are continuously differentiable functions of $a \in M$. We can use the function

$$\operatorname{dist}(x, M) \;=\; \inf\{|x-a| \;:\; a \in M\}$$

to define a "tubular" neighborhood of M on which there is a well-defined nearest point map Π_M whose differential at a point x is close to the orthogonal projection of $\mathbb{R}^n$ onto $T_{\Pi_M(x)}M$. Specifically, we have the following lemma.

LEMMA 6.1. *There are positive constants δ and C depending only on M so that if*

$$U = \{x \in \mathbb{R}^n \;:\; \mathit{dist}(x, M) < \delta\},$$

then the following hold:

(1) *Every point $x \in U$ has a unique nearest point $\Pi_M(x)$ in M.*
(2) *The map Π_M is $C^{1,\alpha}$ smooth.*
(3) *For all $x \in U$,*

$$\|D\Pi_M(x) \;-\; \Pi_{T_{\Pi_M(x)}M}\| \;\le\; C|\Pi_M(x) - x|^\alpha.$$

(4) *The m-dimensional Jacobian*

$$\mathcal{J}_m\Pi_M \equiv \|\wedge_m D\Pi_M\| = \sqrt{\det\left((D\Pi_M)\circ(D\Pi_M)^*\right)}$$

(see [11, *sect.* 3.2.22]*) satisfies*

$$|\,\mathcal{J}_m\Pi_M(x) \;-\; 1\,| \;\le\; C|\Pi_M(x) - x|^\alpha$$

for all $x \in U$.

Proof. As discussed, for example, in [10], the nearest point neighborhood property of a compact C^2 submanifold M depends on its curvature bound. In fact (1) and (2) hold specifically by taking

$$\delta \;=\; \left(\max_{a\in m}\|A_M(a)\|\right)^{-1},$$

where A_M is the second fundamental form of M. For $a \in M$, $D\Pi_M(a) = \Pi_{T_aM}$. Since, for a compact $C^{2,\alpha}$ submanifold M, the map Π_M is $C^{1,\alpha}$ bounded in some compact neighborhood of M, we obtain (3) by slightly shrinking δ. Finally for (4) we note that the linear map $\Pi_{T_{\Pi(x)}M}$, being an orthogonal projection onto an m-dimensional space, has m-dimensional Jacobian equal to 1. Since $\sqrt{t}$ is smooth near $t = 1$, the estimate follows from (3) and the formula for $\mathcal{J}_m$. $\square$

For each $a \in M$ note that the set $\Pi_M^{-1}\{a\}$ is simply an $(n-m)$-dimensional flat disk normal to M at a; in fact,

$$\Pi_M^{-1}\{a\} \;=\; \{y + a \,:\, y \in (T_aM)^\perp \,:\, |y| < \delta\}.$$

Integrals over a tubular neighborhood may be computed using the Jacobian $\mathcal{J}_m\Pi_M$.

LEMMA 6.2. *For any bounded continuous function ψ on $\mathbb{R}^n$,*

$$\int_U \psi(x)\,dx \;=\; \int_M \left(\int_{\Pi_M^{-1}\{a\}} \psi \cdot (\mathcal{J}_m\Pi_M)^{-1}\,d\mathcal{H}^{n-m} \right) d\mathcal{H}^m a.$$

Proof. We may apply the (coarea) change of variable formula [11, sect. 3.2.22, expression (3)] for the map $\Pi_M : U \to M$,

$$\int_U \phi(x)(\mathcal{J}_m\Pi_M)(x)\,dx \;=\; \int_M \left(\int_{\Pi_M^{-1}\{a\}} \phi\,d\mathcal{H}^{n-m} \right) d\mathcal{H}^m a,$$

with $\phi(x) = \psi(x) \cdot (\mathcal{J}_m\Pi_M)^{-1}(x)$. $\square$

For use in section 9.2, we next observe that the following holds.

Lemmas 6.1 *and* 6.2 *continue to hold in the case when the ambient space $\mathbb{R}^n$ is replaced by an n-dimensional Riemannian manifold N.*

Concerning Lemma 6.2, one additional observation is required. In the general Riemannian case, each set $\Pi_M^{-1}\{a\}$, for $a \in M$, is now a uniformly smooth (but possibly curved) $(n-m)$-dimensional disk in N normal to M. It may be parameterized by the planar disk in the normal space,

$$N_\delta(a) \;\equiv\; \{v \in (T_aM)^\perp \;:\; |v| < \delta\} \;\subset\; T_aN.$$

LEMMA 6.3. *There exist positive δ and C and, for every point $a \in M$, a $C^{1,\alpha}$ function G_a mapping $N_\delta(a)$ diffeomorphically onto $\Pi_M^{-1}\{a\}$ so that $G_a(0) = a$, $DG_a(0)$ is an isometry, and, for every $y \in N_\delta(a)$,*

$$\|DG_a(y) - DG_a(0)\| \;\le\; C|y|^\alpha;$$

hence,

$$|y| \;\le\; dist(G_a(y), a) \;\le\; |y| \,+\, C|y|^{1+\alpha}, \tag{6.3}$$

and

$$|\,\mathcal{J}_{n-m}G_a(y) \;-\; 1\,| \;\le\; C|y|^\alpha, \tag{6.4}$$

where $\mathcal{J}_{n-m}G_a = \sqrt{\det\left((DG_a)^* \circ (DG_a)\right)}$.

Proof. The desired parameterizing map G_a is obtained by simply restricting the Riemannian exponential map Exp_a^N to the normal disk $N_\delta(a)$. The estimates all follow from properties of this exponential map. In particular, as a varies over the compact submanifold M, all estimates are uniform because of the $C^{0,\alpha}$ bound on the sectional curvature of M. ∎

Lemma 6.2 is now replaced by the following lemma.

LEMMA 6.4. *For any bounded continuous function ψ on N,*

$$\int_U \psi(x)\,dx = \int_M \left(\int_{\Pi_M^{-1}\{a\}} \psi\cdot(\mathcal{J}_m\Pi_M)^{-1}\,d\mathcal{H}^{n-m}\right) d\mathcal{H}^m(a)$$

$$= \int_M \left(\int_{N_\delta(a)} \psi(G_a(y))\cdot(\mathcal{J}_m\Pi_M)^{-1}(G_a(y))\cdot\mathcal{J}_{n-m}(G_a(y))\,dy\right) d\mathcal{H}^m(a)\,.$$

Proof. The first equality follows from [11, sect. 3.2.22, expression (3)] as in the proof of Lemma 6.2. For the second equality, we then apply the (area) change of variable formula [11, sect. 3.2.5] for the map $G_a : N_\delta(a) \to \Pi_M^{-1}\{a\}$,

$$\int_{N_\delta(a)} \phi(G_a(y))\cdot(\mathcal{J}_{n-m}G_a)(y)\,dy \;=\; \int_{\Pi_M^{-1}\{a\}} \phi\; d\mathcal{H}^{n-m}$$

with $\phi = \psi\cdot(\mathcal{J}_m\Pi_M)^{-1}$. ∎

7. Expansion of a nonnegative function of fixed nondegeneracy.

PROPOSITION 7.1. *Suppose that F is a nonnegative, $C^{3,\alpha}$ smooth function on an open subset of $\mathbb{R}^n$, the zero set $M = F^{-1}\{0\}$ is an embedded $C^{2,\alpha}$ submanifold, and rank$\frac{\partial^2 F}{\partial x_i\partial x_j} \equiv n-m$ on M. For each compact subset A of M there are positive constants C, δ so that for each $a \in A$,* grad$F(a)$ *vanishes and the symmetric matrix $\frac{\partial^2 F}{\partial x_i\partial x_j}(a)$ has, counting multiplicities, m zero eigenvalues and $n-m$ positive eigenvalues*

$$\mu_1(a) \;\le\; \mu_2(a) \;\le\; \cdots \;\le\; \mu_{n-m}(a),$$

which are continuous in a with a positive minimum and a finite maximum. Also there is an orthogonal rotation Γ_a of $\mathbb{R}^n$ so that

$$\Gamma_a(\{0\}\times\mathbb{R}^{n-m}) \;=\; (T_aM)^\perp$$

and

$$\left| F(a+\Gamma_a(x)) \;-\; \sum_{i=1}^{m-n} \frac{1}{2}\mu_i(a)(x_{m+i}-a_{m+i})^2 \right| \;\le\; C\left(\sup_{\mathbb{B}_\delta(a)} \|D^3F\|\right)|x-a|^3$$

for all $x \in \mathbb{B}_\delta(0)$.

Proof. Note that for each $a \in M$ and each vector $v \in \mathbb{R}^n$, the function $F_v(t) = F(a+tv)$ has a minimum at $t=0$. So $v\cdot$grad$F(a) = \frac{dF_v}{dt}\big|_{t=0} = 0$, and

$$0 \;\le\; \left.\frac{d^2F_v}{dt^2}\right|_{t=0} \;=\; \left.\frac{d}{dt}v\cdot \mathrm{grad}F(a+tv)\right|_{t=0} \;=\; \sum_{i,j} v_i \frac{\partial^2 F}{\partial x_i\partial x_j}(a)v_j.$$

Thus grad$F(a) = 0$, and all the eigenvalues of $\frac{\partial^2 F}{\partial x_i\partial x_j}(a)$ are nonnegative. In general, the full collection of eigenvalues of a square matrix, being the complex roots of the characteristic polynomial, varies continuously as the matrix varies; see, e.g., [11]. Here, by assumption, for $a \in M$, the matrix $\left(\frac{\partial^2 F}{\partial x_i \partial x_j}\right)(a)$ has, counting multiplicities, precisely m zero eigenvalues and precisely $n-m$ nonzero, hence positive, eigenvalues. So, under ordering by size, these positive eigenvalues become continuous functions on M. By the compactness of M, μ_1 has a positive minimum and μ_{n-m} a finite maximum.

For $a \in A$, we let $v_1, \dots, v_n$ be orthonormal eigenvectors of the symmetric matrix $\left(\frac{\partial^2 F}{\partial x_i \partial x_j}\right)(a)$ corresponding to the eigenvalues $0, \dots, 0, \mu_1(a), \dots, \mu_{n-m}(a)$, and choose the rotation Γ_a of $\mathbb{R}^n$ satisfying $\Gamma_a(\mathbf{e}_i) = v_i$ for $i = 1, \dots, n$. With $H_a(x) = a + \Gamma_a(x)$, we deduce that the matrix $\left(\frac{\partial^2 (F \circ H_a)}{\partial x_i \partial x_j}\right)(0)$ is diagonal with the first m eigenvalues being zero and the $(m+i)$th eigenvalue being $\mu_i(a)$ for $i = 1, \dots, n-m$. Since also $(F \circ H_a)(0) = 0$, $\operatorname{grad}(F \circ H_a)(0) = 0$, and $\|D^2 F\|$ is bounded in some neighborhood of M, the second order Taylor expansion for $F \circ G_a$ now proves the last inequality of the proposition. ◻

8. Some integrals.

LEMMA 8.1. *For* $k = 1, 2, \dots,$

$$\int_0^\infty e^{-\lambda t^2} t^{k-1}\, dt \;=\; \beta_k \lambda^{-\frac{k}{2}},$$

where

$$\beta_k \;=\; \begin{cases} 2^{-\frac{k}{2}}(k-2)(k-4)\cdots(2) & \textit{for } k \textit{ even,} \\ 2^{-\frac{k-1}{2}}(k-2)(k-4)\cdots(3)\cdot\sqrt{\pi} & \textit{for } k \textit{ odd.} \end{cases}$$

Proof. The substitution $s = \sqrt{\lambda} t$ gives the factor $\lambda^{-\frac{k}{2}}$ and reduces to the case $\lambda = 1$.

Integration by parts gives

$$\int_0^\infty e^{-t^2} t^{k-1}\, dt \;=\; \frac{-1}{2}\int_0^\infty t^{k-2}\, d(e^{-t^2}) \;=\; \frac{k-2}{2}\int_0^\infty e^{-t^2} t^{k-3}\, dt.$$

This may be applied with k replaced by $k-2, k-4, \dots$, finally giving the formula

$$\int_0^\infty e^{-t^2} t^{2j}\, dt \;=\; \begin{cases} 2^{-\frac{k-2}{2}}(k-2)(k-4)\cdots(2)\cdot\int_0^\infty e^{-t^2} t\, dt & \text{for } k \text{ even} \\ 2^{-\frac{k-1}{2}}(k-2)(k-4)\cdots(3)\cdot\int_0^\infty e^{-t^2}\, dt & \text{for } k \text{ odd.} \end{cases}$$

Of course, substituting $s = t^2$ gives $\int_0^\infty e^{-t^2} t\, dt = \frac{1}{2}$, and the last integral is found by the usual polar coordinate trick

$$\left(\int_0^\infty e^{-t^2}\, dt\right)^2 = \int_0^\infty \int_0^\infty e^{-x^2-y^2}\, dx\, dy = \int_0^{2\pi}\int_0^\infty e^{-r^2} r\, dr\, d\theta$$
$$= \frac{2\pi}{2}\int_0^\infty e^{-u}\, du = \pi. \qquad ◻$$

COROLLARY 8.2.

(8.1) $$\lim_{\lambda\to\infty} \lambda^{\frac{k}{2}} \int_0^\infty e^{-\lambda t^2} t^j\, dt \;=\; 0 \quad \textit{for any integer } j > k-1,$$

and

$$\lim_{\lambda\to\infty} \lambda^{\frac{k}{2}} \int_\delta^\infty e^{-\lambda t^2} t^{k-1}\, dt \;=\; 0 \qquad \textit{for any } \delta > 0. \tag{8.2}$$

Proof. Applying Lemma 8.1 with $k = j+1$ gives the first conclusion because $\lambda^{(-j+k-1)/2} \to 0$ as $\lambda \to \infty$. For the second, we change variables $s = \lambda^{1/2} t$ to see that

$$\lambda^{\frac{k}{2}} \int_\delta^\infty e^{-\lambda t^2} t^{k-1}\, dt \;=\; \lambda^{\frac{k}{2}} \lambda^{-\frac{k-1}{2}} \lambda^{-1/2} \int_{\lambda^{1/2}\delta}^\infty e^{-s^2} s^{k-1}\, ds \;\to\; 0 \quad \text{as } \lambda \to \infty$$

because $\int_0^\infty e^{-s^2} s^{k-1}\, ds < \infty$. ◻

COROLLARY 8.3. *For $\delta > 0$ and $0 < \mu_1 \le \mu_2 \le \cdots \le \mu_k < \infty$,*

$$\lim_{\lambda\to\infty} \lambda^{\frac{k}{2}} \int_{\mathbb{B}^k_\delta(0)} e^{-\frac{1}{2}\lambda(\mu_1 y_1^2 + \cdots + \mu_k y_k^2)}\, dy = 2^{k/2}\Lambda^{-1/2} k \alpha_k \beta_k,$$

where $\Lambda = \mu_1 \cdots \mu_k$, α_k is as in (6.1), *and β_k is as in Lemma* 8.1.

Proof. One can explicitly compute the integral over the k-dimensional elliptical region

$$E^k_\delta \;=\; \{\, y \in \mathbb{R}^k \;:\; \mu_1 y_1^2 + \cdots + \mu_1 y_k^2 \;<\; 2\delta^2 \,\},$$

because $E^k_\delta = L(\mathbb{B}^k_\delta(0))$, where

$$L(z_1, \ldots, z_k) \;=\; ((2/\mu_1)^{1/2} z_1, \ldots, (2/\mu_k)^{1/2} z_k) \quad \text{for } (z_1, \ldots, z_k) \in \mathbb{R}^k\ .$$

Using the change of variables $y = L(z)$ with $dy = (\mathcal{J}_k L)\, dz = 2^{k/2}\Lambda^{-1/2} dz$, as well as polar coordinates, Lemma 8.1, and (8.2), we find that

$$\begin{aligned}
&\lambda^{\frac{k}{2}} \int_{E^k_\delta} e^{-\frac{1}{2}\lambda(\mu_1 y_1^2 + \cdots + \mu_k y_k^2)}\, dy \\
&\quad = \; 2^{k/2}\Lambda^{-1/2} \lambda^{\frac{k}{2}} \int_{B^k_\delta} e^{-\lambda|z|^2}\, dz \\
&\quad = \; 2^{k/2}\Lambda^{-1/2} \lambda^{\frac{k}{2}} \int_{S^{k-1}} \int_0^\delta e^{-\lambda r^2} r^{k-1}\, dr\, d\mathcal{H}^{k-1} \\
&\quad = \; 2^{k/2}\Lambda^{-1/2} \lambda^{\frac{k}{2}} k \alpha_k \int_0^\delta e^{-\lambda r^2} r^{k-1}\, dr \\
&\quad = \; 2^{k/2}\Lambda^{-1/2} k \alpha_k \lambda^{\frac{k}{2}} \left(\int_0^\infty e^{-\lambda r^2} r^{k-1}\, dr \;-\; \int_\delta^\infty e^{-\lambda r^2} r^{k-1}\, dr \right) \\
&\quad \to 2^{k/2}\Lambda^{-1/2} k \alpha_k \beta_k \;-\; 0 \qquad\qquad \text{as } \lambda \to \infty.
\end{aligned}$$

We get precisely the same limit with E^k_δ replaced by the ball $\mathbb{B}^k_\delta(0)$ because we have the inclusions

$$\mathbb{B}^k_\gamma(0) \subset \mathbb{B}^k_\delta(0) \subset \mathbb{B}^k_\epsilon(0) \qquad \text{and} \qquad \mathbb{B}^k_\gamma(0) \subset E^k_\delta(0) \subset \mathbb{B}^k_\epsilon(0)$$

with $\gamma = \min\{\delta, (2/\mu_k)^{1/2}\delta\}$ and $\epsilon = \max\{\delta, (2/\mu_1)^{1/2}\delta\}$ and we have

$$\lim_{\lambda\to\infty} \lambda^{\frac{k}{2}} \int_{\mathbb{B}^k_\epsilon(0)\setminus\mathbb{B}^k_\gamma(0)} e^{-\frac{1}{2}\lambda(\mu_1 y_1^2+\cdots+\mu_k y_k^2)}\,dy \le \lim_{\lambda\to\infty} \lambda^{\frac{k}{2}} \int_{\mathbb{B}^k_\epsilon(0)\setminus\mathbb{B}^k_\gamma(0)} e^{-\frac{1}{2}\lambda(\mu_1|y|^2)}\,dy$$

$$\le \lim_{\lambda\to\infty} k\alpha_k\lambda^{\frac{k}{2}} \int_\gamma^\infty e^{-\frac{1}{2}\lambda(\mu_1 r^2)} r^{k-1}\,dr = 0$$

by (8.2). ◻

LEMMA 8.4. *Suppose $0 \le k \le n$, F is a nonnegative continuous function on $\mathbb{R}^n$, $p > 0$, and $F(y) \ge |y|^p$ whenever $|y|$ is sufficiently large. Then, for any bounded open neighborhood U of $F^{-1}\{0\}$,*

$$\lim_{\lambda\to\infty} \lambda^{k/2} \int_{\mathbb{R}^n\setminus U} e^{-\lambda F(y)}\,dy \;=\; 0.$$

Proof. We may assume $p < 2$. Choose $R > 0$ so that $\overline{U} \subset \mathbb{B}_R(0)$ and $F(y) \ge |y|^p$ whenever $|y| \ge R$. On the bounded region $\mathbb{B}_R(0)\setminus U$, F has a positive lower bound ε, and

$$\lambda^{k/2} e^{-\lambda F(y)} \;\le\; \lambda^{k/2} e^{-\lambda\varepsilon} \;\to\; 0$$

uniformly as $\lambda \to \infty$. Thus

$$\lim_{\lambda\to\infty} \lambda^{k/2} \int_{\mathbb{B}_R(0)\setminus U} e^{-\lambda F(y)}\,dy \;=\; 0.$$

For the remaining set $\mathbb{R}^k \setminus \mathbb{B}_R(0)$, we use polar coordinates and change variables $s = \lambda^{1/p} t$ to see that

$$\lambda^{k/2} \int_{\mathbb{R}^n\setminus\mathbb{B}_R(0)} e^{-\lambda F(y)}\,dy \le n\alpha_{n-1}\lambda^{k/2} \int_R^\infty e^{-\lambda t^p} t^{n-1}\,dt$$

$$= n\alpha_{n-1}\lambda^{\frac{k}{2}-\frac{n-1}{p}-\frac{1}{p}} \int_{\lambda^{1/p}R}^\infty e^{-s^p} s^{n-1}\,ds \;\to\; 0$$

as $\lambda \to \infty$ because $\frac{k}{2} - \frac{n}{p} \le 0$ and $\int_0^\infty e^{-s^p} s^{n-1}\,ds < \infty$. ◻

9. Proof of main theorems.

9.1. Proof of Theorem 3.1. We begin by noting that without loss of generality, we may set $h \equiv 1$. To show the dependence of the measures on h, let $P_\lambda^{(h)}$ denote the Gibbs measures in Theorem 3.1 and $P^{(h)}$ the limiting measure. For any bounded continuous ϕ, since ϕh is bounded and continuous by our assumptions, if the theorem holds when $h \equiv 1$, then we have

$$\int \phi h\, dP_\lambda^{(1)} \;\to\; \int \phi h\, dP^{(1)},$$

and noting that $\int h\, dP^{(1)} > 0$ by our assumption (A5)(d), we have

$$\int \phi\, dP_\lambda^{(h)} \;=\; \frac{\int \phi h\, dP_\lambda^{(1)}}{\int h\, dP_\lambda^{(1)}} \;\to\; \frac{\int \phi h\, dP^{(1)}}{\int h\, dP^{(1)}} \;=\; \int \phi\, dP^{(h)}.$$

First we treat the case $j = 1$ where the Hessian $D^2 J$ has constant rank $k = k$ on the submanifold $M = M_1 = J^{-1}\{0\}$. Taking $F = J$ and $M = J^{-1}\{0\}$, we choose δ and $U \subset U_1$ as in Lemma 6.1 and Proposition 7.1 with $m = n - k$. In the remainder of the proof, we will occasionally enlarge C, finitely many times, without changing the notation. Nevertheless, the constant C will always just depend on n and J.

Let ϕ be a bounded continuous function on $\mathbb{R}^n$ and ε be any positive number with

$$\varepsilon \;<\; \min\left\{\frac{1}{2}, \frac{1}{2C}\right\},$$

where C is as in Lemma 6.1.

First we may assume that the δ and corresponding tubular neighborhood U in Lemma 6.1 are small enough so that, for any point $x \in U$ and nearest point $a = \Pi_M(x) \in M$, one has

$$|x - a| \;<\; \delta \;<\; \varepsilon, \quad |\phi(x) - \phi(a)| \;<\; \varepsilon. \tag{9.1}$$

For λ sufficiently large, we have, by Lemma 8.4, that

$$\lambda^{\frac{k}{2}} \int_{\mathbb{R}^n \setminus U} \phi(x) e^{-\lambda J(x)}\, dx \;\leq\; \big(\sup|\phi|\big)\, \lambda^{\frac{k}{2}} \int_{\mathbb{R}^n \setminus U} e^{-\lambda J(x)}\, dx \;<\; \varepsilon. \tag{9.2}$$

For the integral over U, we use Lemma 6.2 with ψ replaced by $\phi e^{-\lambda J}$ to write

$$\lambda^{\frac{k}{2}} \int_U \phi(x) e^{-\lambda J(x)}\, dx = \lambda^{\frac{k}{2}} \int_M \left(\int_{\Pi_M^{-1}(a)} z_a(x)\, d\mathcal{H}^k(x) \right) d\mathcal{H}^{n-k}(a), \tag{9.3}$$

where

$$z_a(x) \overset{\text{def}}{=} \phi(x) \cdot e^{-\lambda J(x)} \cdot \mathcal{J}_k \Pi_M(x)\,.$$

We will first make upper estimates. For the first factor with $\Pi_M(x) = a$, we use (9.1) to see that

$$\phi(x) \;\leq\; \phi(a) \;+\; \varepsilon.$$

For the third factor, we use Lemma 6.1 to obtain

$$\mathcal{J}_k \Pi_M(x) \;\leq\; (1 + C\varepsilon^\alpha).$$

Combining these and changing C, we now have the upper bound

$$\begin{aligned} &\lambda^{\frac{k}{2}} \int_U \phi(x) e^{-\lambda J(x)}\, dx \\ &\quad\leq \lambda^{\frac{k}{2}} (1 + C\varepsilon^\alpha) \int_M (\phi(a) + \varepsilon) \int_{\Pi_M^{-1}\{a\}} e^{-\lambda J(x)}\, d\mathcal{H}^k(x)\, d\mathcal{H}^{n-k}(a). \end{aligned} \tag{9.4}$$

The remaining second factor in z_α is in the last integral. To estimate this, we rotate coordinates as in Proposition 7.1 with $F = J$ and use Corollary 8.3 with $k = k$, and

(8.1) with $k = k+2$ to deduce that

$$
\begin{aligned}
\lambda^{\frac{k}{2}} \int_{\Pi_M^{-1}\{a\}} e^{-\lambda J(x)}\, d\mathcal{H}^k(x) &\le \lambda^{\frac{k}{2}} \int_{\mathbb{B}_\delta(0)} e^{-\lambda(\sum_{i=1}^k \frac{1}{2}\mu_i(a) y_i^2)} e^{C\lambda|y|^3}\, dy \\
&\le \lambda^{\frac{k}{2}} \int_{\mathbb{B}_\delta(0)} e^{-\lambda(\sum_{i=1}^k \frac{1}{2}\mu_i(a) y_i^2)} (1 + 2C\lambda|y|^3)\, dy \\
&\le \lambda^{\frac{k}{2}} \int_{\mathbb{B}_\delta(0)} e^{-\lambda(\sum_{i=1}^k \frac{1}{2}\mu_i(a) y_i^2)}\, dy + 2C\lambda^{\frac{k+2}{2}}\, k\alpha_k \int_0^\infty e^{-\frac{1}{2}\lambda\mu_1(a) r^2} r^{3+k-1}\, dr \\
&\longrightarrow 2^{\frac{k}{2}} \Lambda^{-1/2}(a)\, k\alpha_k\beta_k + 0 \quad \text{as } \lambda \to \infty,
\end{aligned} \tag{9.5}
$$

where $\Lambda(a) = \mu_1(a) \ldots \mu_k(a)$. Taking the $\limsup_{\lambda\to\infty}$ in (9.4) along with (9.5), recalling (9.2), and then letting $\varepsilon \downarrow 0$, we conclude that

$$
\limsup_{\lambda\to\infty} \lambda^{\frac{k}{2}} \int_{\mathbb{R}^n} \phi(x) e^{-\lambda J(x)}\, dx \;\le\; 2^{\frac{k}{2}}\, k\alpha_k\beta_k \int_M \phi(a)\Lambda^{-1/2}(a)\, d\mathcal{H}^{n-k}(a). \tag{9.6}
$$

Next, arguing in the same manner using lower bounds gives the inequality

$$
\liminf_{\lambda\to\infty} \lambda^{\frac{k}{2}} \int_{\mathbb{R}^n} \phi(x) e^{-\lambda J(x)}\, dx \;\ge\; 2^{\frac{k}{2}}\, k\alpha_k\beta_k \int_M \phi(a)\Lambda^{-1/2}(a)\, d\mathcal{H}^{n-k}(a). \tag{9.7}
$$

This essentially finishes the proof. For the normalization, we define, for $\lambda > 1$,

$$
Y_\lambda \;=\; \left(2^{\frac{k}{2}}\, k\alpha_k\beta_k\right)^{-1} Z\lambda^{\frac{k}{2}},
$$

where

$$
Z \;=\; \left(\int_M \Lambda^{-1/2}(a)\, d\mathcal{H}^{n-k}(a) \right)^{-1}
$$

as before. We now apply (9.6) and (9.7) first with $\phi \equiv 1$ to see that

$$
\lim_{\lambda\to\infty} \frac{Y_\lambda}{Z_\lambda} \;=\; \lim_{\lambda\to\infty} Y_\lambda \int_{\mathbb{R}^n} e^{-\lambda J(x)}\, dx \;=\; 1,
$$

and second with the general bounded continuous ϕ to obtain the conclusion

$$
\begin{aligned}
\lim_{\lambda\to\infty} Z_\lambda \int_{\mathbb{R}^n} \phi(x) e^{-\lambda J(x)}\, dx = &\lim_{\lambda\to\infty} Y_\lambda \int_{\mathbb{R}^n} \phi(x) e^{-\lambda J(x)}\, dx \\
= &\; Z \int_M \phi(a)\Lambda^{-1/2}(a)\, d\mathcal{H}^{n-k}(a),
\end{aligned}
$$

which gives the desired convergence of measures $P_\lambda \Rightarrow P$.

Finally we consider the case $j > 1$ of Theorem 3.1 involving the extra disjoint compact submanifolds $M_2, \ldots, M_j$ on each of which the Hessian $D^2 J$ has constant rank strictly larger than k_1, which is its rank on M_1. Now we will repeat most of the above arguments and again use the factor $\lambda^{k_1/2}$ to try to estimate

$$
\lambda^{\frac{k_1}{2}} \int_{\mathbb{R}^n} \phi(x) e^{-\lambda J(x)}\, dx
$$

as $\lambda \to \infty$. As before we may, by Lemma 8.4, restrict our integration to any fixed neighborhood U of M. We take $U = U_1 \cup \cdots \cup U_j$, where each U_i is, as before, a sufficiently small (depending on a given ε and test function ϕ) tubular neighborhood of M_i.

For the region U_1, we find, by estimating the upper and lower bounds just as before, that

$$\lim_{\lambda\to\infty} \lambda^{\frac{k_1}{2}} \int_{U_1} \phi(x) e^{-\lambda J(x)}\, dx \;=\; 2^{\frac{k_1}{2}}\, k_1 \alpha_{k_1} \beta_{k_1} \int_{M_1} \Lambda^{-1/2}(a)\, d\mathcal{H}^{n-k_1}(a), \tag{9.8}$$

where $\Lambda(a)$ is, as before, the product of the k_1 positive eigenvalues of $D^2 J(a)$ for $a \in M_1$.

However, for any region U_i with $i = 2, \ldots, j$, one finds that, with $a \in M_i$, in place of (9.5), one has the upper estimate

(9.9)

$$\begin{aligned}
&\lambda^{\frac{k_1}{2}} \int_{\Pi_M^{-1}\{a\}} e^{-\lambda J(x)}\, d\mathcal{H}^{k_i}(x) \\
&\quad \le\; \lambda^{\frac{k_1}{2}} \int_{\mathbb{B}_\delta(0)} e^{-\frac{1}{2}\lambda\mu_1(a)|y|^2)} (1 \;+\; 2C\lambda|y|^3)\, dy \\
&\quad =\; \lambda^{\frac{k_1}{2}}\, k_i \alpha_{k_i} \int_0^\infty e^{-\frac{1}{2}\lambda\mu_1(a) r^2} r^{k_i - 1}\, dr \;+\; 2C\lambda^{\frac{k_1+2}{2}}\, k_i \alpha_{k_i} \int_0^\infty e^{-\frac{1}{2}\lambda\mu_1(a) r^2} r^{3+k_i-1}\, dr \\
&\quad \longrightarrow\; 0 + 0 \quad \text{as } \lambda \to \infty
\end{aligned}$$

by (8.1), because $k_i > k_1$ and $3 + k_i > k_1 + 2$. It follows that, for any bounded continuous ϕ on $\mathbb{R}^n$,

$$\lim_{\lambda\to\infty} \lambda^{\frac{k_1}{2}} \int_{U_i} \phi(x) e^{-\lambda J(x)}\, dx \;=\; 0 \tag{9.10}$$

for $i = 2, \ldots, j$. With

$$Y_\lambda \;=\; \left(2^{\frac{k_1}{2}}\, k_1 \alpha_{k_1} \beta_{k_1}\right)^{-1} Z \lambda^{\frac{k_1}{2}},$$

we conclude from (8.1), (9.8), and (9.10) as before that, as $\lambda \to \infty$, $Y_\lambda / Z_\lambda \to 1$ and

$$Z_\lambda \int_{\mathbb{R}^n} \phi(x) e^{-\lambda J(x)}\, dx \;\to\; Z \int_{M_1} \phi(a) \Lambda^{-1/2}(a)\, d\mathcal{H}^{n-k_1} a,$$

which completes the proof. $\square$

9.2. Multiple limit theorem. Multiple constraints generated by functions $J_1, J_2, \ldots, J_j$ lead to consideration of Gibbs measures obtained from multiple limits $\lim_{\lambda_1\to\infty} \lim_{\lambda_2\to\infty} \cdots \lim_{\lambda_j\to\infty}$. With suitable nondegeneracy assumptions on the Hessians of the J_i, one expects to obtain consecutively suitably weighted Hausdorff measures on lower and lower dimensional submanifolds. As before, it suffices to prove the case $h \equiv 1$. To inductively follow this procedure, one first needs to verify the following proposition.

PROPOSITION 9.1. *Theorem* 3.1 *remains true if* $\mathbb{R}^n$ *is replaced by an* n*-dimensional complete Riemannian manifold* N, *the term* $D^2 J$ *is replaced by the Riemannian Hessian, and the Lebesgue measure on* $\mathbb{R}^n$ *is replaced by the standard* n*-dimensional Hausdorff measure on* N.

Proof. The proof follows exactly as in section 9.1 using Riemannian normal coordinates in the expansions until we need to apply Lemma 6.4 instead of Lemma 6.2 and replace the right-hand side of (9.3) by

$$\lambda^{\frac{k}{2}} \int_M \left(\int_{N_\delta(a)} \phi(G_a(y)) \cdot e^{-\lambda J(G_a(y))} \cdot \mathcal{J}_{n-k} G_a(y) \cdot \mathcal{J}_k(G_a(y))\, dy \right) d\mathcal{H}^{n-k} a\,, \tag{9.11}$$

where G_a is given in Lemma 6.3. Note that, by our choice of U, also small enough for Lemma 6.3, and by (9.1), one has, with $x = G_a(y) \in U$ (hence $a = \Pi_M(x)$), that

$$|y| \;\leq\; \mathrm{dist}(x,a) \;<\; \delta \;<\; \varepsilon\,.$$

We estimate the first three factors of (9.11) as before in our estimate of (9.3). Although it is no longer true that $\mathcal{J}_{n-k} G_a \equiv 1$, we may use (6.4) to estimate the new factor $\mathcal{J}_{n-k} G_a(y)$. The fourth factor is also estimated using Lemma 6.3 to obtain

$$\mathcal{J}_k(G_a(y)) \;\leq\; 1 + C\varepsilon^\alpha.$$

Combining these and changing C, we now have the upper bound

$$\lambda^{\frac{k}{2}} \int_U \phi(x) e^{-\lambda J(x)}\, d\mathcal{H}^n x \leq\; \lambda^{\frac{k}{2}} (1 + C\varepsilon^\alpha) \int_M (\phi(a) + \varepsilon) \int_{N_\delta(a)} e^{-\lambda J(G_a(y))}\, dy\, d\mathcal{H}^{n-k} a,$$

which corresponds to (9.4). The remainder of the proof now follows precisely as before in section 9. ∎

9.3. Proof of Theorem 3.2. For each $a \in M$ and $1 \leq h < i \leq j$, the subspaces $E_h(a)$ and $E_i(a)$ have zero-dimensional intersection because the span of $E_1(a), \ldots, E_j(a)$ is of dimension $k = k_1 + \cdots + k_j$. In the special case they are all *mutually orthogonal*, the Hessian matrices $D^2 J_1(a), \ldots, D^2 J_j(a)$ may be simultaneously diagonalized. We now find the expansion in the fiber, $\Pi_W^{-1}\{a\}$, that $\lambda_1 J_1 + \cdots + \lambda_n J_j$ now has the form

$$\begin{aligned} &\frac{1}{2}\big[\lambda_1 \mu_{n-k+1} y_{n-k+1}^2 + \cdots + \lambda_1 \mu_{n-k+k_1} y_{n-k+k_1}^2 + \cdots \\ &\qquad + \lambda_j \mu_{n-k_j+1} y_{n-k_j+1}^2 + \cdots + \lambda_j \mu_n y_n^2\big] \;+\; \text{higher order terms.} \end{aligned}$$

We may now estimate terms as in the proofs in section 9, including that of Proposition 9.1. In taking any limit $\lambda_i \to \infty$, the remaining factors $e^{-\lambda_j J_j}$ may be treated as part of the test function in the weak convergence. The mutual orthogonality also implies that $\Theta \equiv 1$, and we thus obtain Theorem 3.2 in the orthogonal case.

In the general transverse (but not necessarily orthogonal) case, we can make a smooth change of coordinates of a full neighborhood U of each $a \in M$ so that, at each $x \in U \cap M$, the differential of the transformation preserves $E(x)^\perp$ and transforms $E(x)$ by taking the vectors chosen from some orthonormal bases of $E_1(x), \ldots, E_j(x)$ (which by assumption span $\mathbb{R}^n$) to a single orthonormal basis of $E(x)$. With this change of variables, we may compute the desired integrals over U and then use the orthogonality as above. This change of variable that makes the E_i mutually orthogonal gives rise to a new n-dimensional Jacobian term, which is precisely Θ on $U \cap M$. The desired limiting formula follows. ∎

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, 1972.

[2] K. Bhattacharya and G. Dolzmann, *Relaxation of some multi-well problems*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 279–320.

[3] P. Billingsley, *Convergence of Probability Measures*, 2nd ed., John Wiley & Sons, New York, 1999.

[4] N. Bleistein and R. A. Handelsman, *Asymptotic Expansions of Integrals*, Dover, New York, 1986.

[5] I. Bohachevsky, M. Johnson, and M. Stein, *Generalized simulated annealing for function optimization*, Technometrics, 28 (1986), pp. 209–217.

[6] A. Corana, M. Marchesi, C. Martini, and S. Ridella, *Continuous variables with the simulated annealing algorithm*, ACM Trans. Math. Software, 13 (1987), pp. 262–280.

[7] D. Cox, P. Klouček, and D. R. Reynolds, *A subgrid projection method for relaxation of non-attainable differential inclusions*, in Proceedings of ENUMATH 2001, Numerical Mathematics and Advanced Applications, F. Brezzi, A. Buffa, S. Corsaro, and A. Murli, eds., Springer-Verlag, Berlin, 2001, pp. 575–584.

[8] D. Cox, P. Klouček, and D. R. Reynolds, *On the asymptotically stochastic computational modeling of microstructures*, Future Generation Computer Systems, 20 (2004), pp. 409–424.

[9] B. Dacorogna and P. Marcellini, *Implicit Partial Differential Equations*, Birkhäuser Boston, Boston, 1999.

[10] H. Federer, *Curvatures measures*, Trans. Amer. Math. Soc., 93 (1959), pp. 418–491.

[11] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[12] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC, London, 1995.

[13] B. Hajek, *Cooling schedules for optimal simulated annealing*, Math. Oper. Res., 13 (1988), pp. 311–329.

[14] A. R. Hedar and M. Fukushima, *Derivative-free filter simulated annealing method for constrained continuous global optimization*, J. Global Optim., 35 (2006), pp. 521–549.

[15] R. A. Holley, S. Kusuoka, and D. W. Stroock, *Asymptotics of the spectral gap with applications to the theory of simulated annealing*, J. Funct. Anal., 83 (1989), pp. 333–347.

[16] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680.

[17] P. Klouček and M. Luskin, *Computational modeling of the martensitic transformation with surface energy*, Math. Comput. Modelling, 20 (1994), pp. 101–121.

[18] P. Klouček and M. Luskin, *The computation of the dynamics of the martensitic deformation*, Contin. Mech. Thermodyn., 6 (1994), pp. 209–240.

[19] P. Klouček and D. R. Reynolds, *On the modeling of nonlinear thermodynamics in SMA wires*, Comput. Methods Appl. Mech. Engrg., 196 (2006), pp. 180–191.

[20] P. Klouček, D. R. Reynolds, and T. I. Seidman, *Computational modeling of vibration damping in SMA wires*, Contin. Mech. Thermodyn., 16 (2004), pp. 495–514.

[21] M. Locatelli, *Simulated annealing algorithms for continuous global optimization: Convergence conditions*, J. Optim. Theory Appl., 104 (2000), pp. 121–133.

[22] N. Metropolis, A. W. Rosebluth, M. N. Rosebluth, A. H. Teller, and E. Teller, *Equation of state calculations by fast computing mechines*, J. Chem. Phys., 21 (1953), pp. 1087–1092.

[23] P. Pedregal, *Variational Methods in Nonlinear Elasticity*, SIAM, Philadelphia, 2000.

[24] D. Reynolds, *Vibration Damping Using Martensitic Phase Transformation*, Ph.D. thesis, Rice University, Houston, TX, 2002.

[25] M. Spivak, *Calculus on Manifolds: A Modern Approach to Classical Theorems of Advanced Calculus*, W.A. Benjamin, New York, Amsterdam, 1965.

[26] L. Tierney, *Markov chains for exploring posterior distributions*, Ann. Statist., 22 (1994), pp. 1701–1762.

[27] L. Tierney and J. B. Kadane, *Accurate approximations for posterior moments and marginal densities*, J. Amer. Statist. Assoc., 81 (1986), pp. 82–86.

[28] B. W. Wah, Y. X. Chen, and T. Wang, *Simulated annealing with asymptotic convergence for nonlinear constrained optimization*, J. Global Optim., 39 (2007), pp. 1–37.

SIAM J. MATH. ANAL.
Vol. 39, No. 5, pp. 1497–1535

SPHERICAL SOLUTIONS TO A NONLOCAL FREE BOUNDARY PROBLEM FROM DIBLOCK COPOLYMER MORPHOLOGY*

XIAOFENG REN† AND JUNCHENG WEI‡

Abstract. The Γ-limit of the Ohta–Kawasaki density functional theory of diblock copolymers is a nonlocal free boundary problem. For some values of block composition and the nonlocal interaction, an equilibrium pattern of many spheres exists in a three-dimensional domain. A subrange of the parameters is found where the multiple sphere pattern is stable. This stable pattern models the spherical phase in the diblock copolymer morphology. The spheres are approximately round. They satisfy an equation that involves their mean curvature and a quantity that depends nonlocally on the whole pattern. The locations of the spheres are determined via a Green's function of the domain.

Key words. spherical phase, diblock copolymer morphology, sphere coarsening, interface oscillation

AMS subject classifications. 35R35, 82B24, 82D60

DOI. 10.1137/070690286

1. Introduction. A diblock copolymer melt is a soft material, characterized by fluid-like disorder on the molecular scale and a high degree of order at a longer length scale. A molecule in a diblock copolymer is a linear subchain of A-monomers grafted covalently to another subchain of B-monomers. Because of the repulsion between the unlike monomers, the different type subchains tend to segregate, but as they are chemically bonded in chain molecules, segregation of subchains cannot lead to a macroscopic phase separation. Only a local microphase separation occurs: microdomains rich in A-monomers and microdomains rich in B-monomers emerge as a result. These microdomains form patterns that are known as morphology phases. Various phases, including lamellar, cylindrical, spherical, and gyroid, have been observed in experiments. See Bates and Fredrickson [1] for more on block copolymers.

This paper deals with the spherical phase of the block copolymer morphology (Figure 1.1, Plot 1). Let $a \in (0,1)$ be the block composition fraction which is the number of the A-monomers divided by the number of all the A- and B-monomers in a chain molecule. The spherical phase occurs when a is relatively close to 0 (or close to 1), and the A-monomers (or B-monomers, respectively) form small balls in space.

The model we use here is a nonlocal free boundary problem derived from the Ohta–Kawasaki density functional theory of diblock copolymers [18]. Let D be a bounded and sufficiently smooth domain in R^3 occupied by a diblock copolymer melt in the spherical phase. Let E be a subset of D where A-monomers concentrate. Then $D\backslash E$ is the subset where B-monomers concentrate. Denote the part of the boundary of E that is in D by $\partial_D E$ which is the set of the interfaces separating the A-rich microdomains from the B-rich microdomains. Denote the Lebesgue measure of E by $|E|$. Given a block composition fraction $a \in (0,1)$, one has $|E| = a|D|$.

*Received by the editors May 2, 2007; accepted for publication (in revised form) July 16, 2007; published electronically January 22, 2008.
http://www.siam.org/journals/sima/39-5/69028.html

†Department of Mathematics, The George Washington University, Washington, DC 20052 (ren@gwu.edu). This author's research was supported in part by NSF grant DMS-0509725.

‡Department of Mathematics, Chinese University of Hong Kong, Hong Kong, People's Republic of China (wei@math.cuhk.edu.hk). This author's research was supported in part by an Earmarked grant of RGC of Hong Kong.

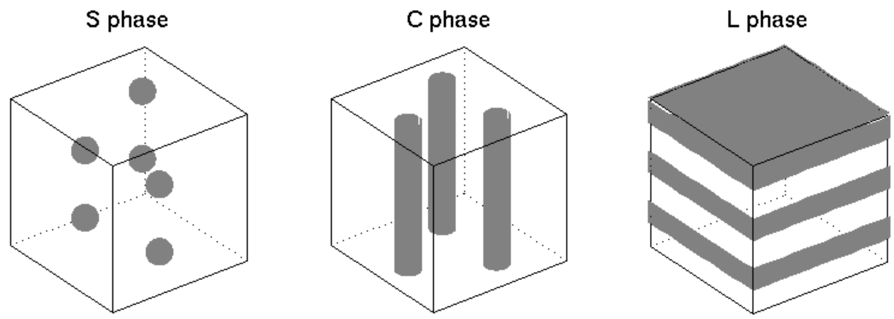


FIG. 1.1. *The spherical, cylindrical, and lamellar morphology phases commonly observed in diblock copolymer melts. The dark color indicates the concentration of type A monomers, and the white color indicates the concentration of type B monomers.*

Moreover, there exists a number λ such that at every point on $\partial_D E$

$$H(\partial_D E) + \gamma(-\Delta)^{-1}(\chi_E - a) = \lambda. \tag{1.1}$$

Here $H(\partial_D E)$ is the mean curvature of $\partial_D E$ viewed from E, γ is a positive parameter, and χ_E is the characteristic function of E, i.e. $\chi_E(x) = 1$ if $x \in E$, and $\chi_E(x) = 0$ if $x \in D \backslash E$. The expression $(-\Delta)^{-1}(\chi_E - a)$ is the solution v of the problem

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

where the bar over a function is the average of the function over its domain, i.e.,

$$\overline{v} = \frac{1}{|D|} \int_D v(x)\, dx.$$

Because $(-\Delta)^{-1}$ is a nonlocal operator defined from $\{q \in L^2(D) : \overline{q} = 0\}$ to itself, the free boundary problem (1.1) is nonlocal.

Equation (1.1) is the Euler–Lagrange equation of the free energy J of the system. The functional J is given by

$$J(E) = |D\chi_E|(D) + \frac{\gamma}{2} \int_D |(-\Delta)^{-1/2}(\chi_E - a)|^2\, dx, \quad E \in \Sigma. \tag{1.2}$$

The admissible set Σ of the functional J is the collection of all measurable subsets of D of measure $a|D|$ and of finite perimeter, i.e.,

$$\Sigma = \{E \subset D : E \text{ is Lebesgue measurable}, \ |E| = a|D|, \ \chi_E \in BV(D)\}. \tag{1.3}$$

Here $BV(D)$ is the space of functions of bounded variation on D. In (1.2), $|D\chi_E|(D)$ is the perimeter of E. When ∂E is smooth, this is merely the surface area of $\partial_D E$. For a more general E, χ_E is a BV-function and $D\chi_E$ is a vector valued finite measure. We denote the magnitude of this measure by $|D\chi_E|$ which is a positive, finite measure. The perimeter of E is defined to be the size of D under this measure. The operator $(-\Delta)^{-1/2}$ is the positive square root of $(-\Delta)^{-1}$.

The main difficulty in (1.1) stems from the nonlocal term. Without it, i.e., if $\gamma = 0$, (1.1) would just be the equation of constant mean curvature. However with the nonlocal term the curvature of a solution in general is not constant. One exception occurs in the study of the lamellar phase (Figure 1.1, Plot 3) where interfaces are parallel planes (Ren and Wei [20, 23]). The solution we are looking for in this paper is a union of a number of disconnected sets each of which is close to a small round ball. The solution is hence termed a spherical solution.

Nishiura and Ohnishi [16] formulated the Ohta–Kawasaki theory on a bounded domain as a singularly perturbed variational problem with a nonlocal term and also identified the free boundary problem (1.1). Ren and Wei [20] showed that (1.2) is a Γ-limit of the singularly perturbed variational problem. See the last section for more discussion on the Ohta–Kawasaki theory and Γ-convergence.

Since then much work has been done mathematically to these problems. The lamellar phase (Figure 1.1, Plot 3) is studied by Ren and Wei [20, 22, 23, 27, 28], Fife and Hilhorst [9], Choksi and Ren [4], Chen and Oshita [2], and Choksi and Sternberg [6]. The result obtained by Müller [15] is related to the lamellar phase in the case $a = 1/2$, as observed in [16]. Radially symmetric bubble and ring patterns are studied by Ren and Wei [21, 26, 29]. The gyroid phase is numerically studied by Teramoto and Nishiura [33]. Triblock copolymers are studied by Ren and Wei [24, 25]. A diblock copolymer/homopolymer blend is studied by Choksi and Ren [5]. Also, see Ohnishi et al. [17] and Choksi [3].

The cylindrical phase (Figure 1.1, Plot 2) is studied by Ren and Wei [31, 30], in which a variant of the Lyapunov–Schmidt reduction procedure is developed to study a cross section of Figure 1.1, Plot 2. A pattern with a number of approximate small discs is found which satisfies the two-dimensional version of (1.1). In two dimensions, $\partial_D E$ is a union of curves and $H(\partial_D E)$ is the curvature of the curves.

In this paper we adapt the Lyapunov–Schmidt reduction procedure to three dimensions to construct spherical solutions. These solutions look like Figure 1.1, Plot 1. They model the spherical phase of diblock copolymer morphology.

The main results are presented in section 2. Our strategy to prove them consists of setting up a first approximation (section 3) and through linearization (sections 4 and 5) and fixed point argument (sections 6 and 7) solving a projected version of the full problem (up to spherical harmonics of order 0 and 1 corresponding to translations and changes in volume). This reduces the infinite dimensional variational problem to a finite dimensional minimization problem in centers and radii. After finding a minimum of the finite dimensional problem, we show that it is indeed an exact solution of the full problem, using a tricky reparametrization argument (section 8).

Our construction yields, in addition, information on the spectra of linearization, interpreted as forms of stability-instability.

Compared to the two-dimensional case, the study of the linearized problem is more involved here. In two dimensions the corresponding linearized problem is analyzed by the Fourier series method. Here in three dimensions we use spherical harmonics to diagonalize the linearized operator (see Lemma 5.1). More differences between the two-dimensional and the three-dimensional cases are given in section 9.

2. Main results. The Green's function of $-\Delta$ is denoted by G. It is a sum of two parts:

$$G(x,y) = \frac{1}{4\pi|x-y|} + R(x,y). \tag{2.1}$$

The first part on the right-hand side of (2.1) is the fundamental solution in three dimensions. The second part is the regular part of $G(x,y)$, denoted by $R(x,y)$. The

Green's function satisfies
(2.2)

$$-\Delta_x G(x,y) = \delta(x-y) - \frac{1}{|D|} \text{ in } D, \quad \partial_{\nu(x)} G(x,y) = 0 \text{ on } \partial D, \quad \overline{G(\cdot,y)} = 0 \ \forall y \in D.$$

Here Δ_x is the Laplacian with respect to the x-variable of G, and $\nu(x)$ is the outward normal direction at $x \in \partial D$. We set

$$F(\xi_1, \xi_2, \ldots, \xi_K) = \sum_{k=1}^{K} R(\xi_k, \xi_k) + \sum_{k=1}^{K} \sum_{l=1, l\neq k}^{K} G(\xi_k, \xi_l), \tag{2.3}$$

for $\xi_k \in D$ and $\xi_k \neq \xi_l$ if $k \neq l$. Because $G(x,y) \to \infty$ if $|x-y| \to 0$ and $R(x,x) \to \infty$ if $x \to \partial D$, F admits at least one global minimum.

The average sphere radius is

$$\rho = \left(\frac{3a|D|}{4\pi K}\right)^{1/3}. \tag{2.4}$$

The main result of this paper is the following existence theorem.

THEOREM 2.1. *Let $K \geq 2$ be an integer.*

1. *For every $\epsilon > 0$ there exists $\delta > 0$, depending on ϵ, K, and D only, such that if*

$$\gamma\rho^3 > 3 + \epsilon, \tag{2.5}$$

$$\left|\gamma\rho^3 - \frac{3(n+2)(2n+1)}{2}\right| > \epsilon n^2 \ \forall\ n = 2,3,4,\ldots, \tag{2.6}$$

and

$$\rho < \delta, \tag{2.7}$$

then there exists a solution E of (1.1).

2. *The solution E is a union of K approximate balls. The radius of each ball is close to ρ.*
3. *Let the centers of these balls be $\zeta_1, \zeta_2, \ldots, \zeta_K$. Then $\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_K)$ is close to a global minimum of the function F.*

The precise meaning that each component of E is close to a ball of radius ρ is given in (8.18). As ρ (or a) tends to 0, ζ converges to a global minimum of F, possibly along a subsequence.

We have opted for a rather general existence theorem. The solution found in the theorem is not necessarily stable. The stability of the solution depends on how (2.6) is satisfied.

THEOREM 2.2. *If* (2.6) *is satisfied because*

$$\gamma\rho^3 - \frac{3(n+2)(2n+1)}{2} < -\epsilon n^2 \ \forall\ n \geq 2, \tag{2.8}$$

then the spherical solution is stable. Otherwise if (2.6) *is satisfied but*

$$\epsilon n^2 < \gamma\rho^3 - \frac{3(n+2)(2n+1)}{2}, \ \textit{and}\ \gamma\rho^3 - \frac{3(n+3)(2n+3)}{2} < -\epsilon(n+1)^2 \tag{2.9}$$

for some $n \geq 2$, then the spherical solution is unstable.

When we delete intervals around $\frac{3(n+2)(2n+1)}{2}$, $n = 2, 3, \ldots$, in (2.6), the width of the intervals, $2\epsilon n^2$, grows as n becomes large. At some point an interval will include nearby members in the sequence $\frac{3(n+2)(2n+1)}{2}$. When this happens, $\gamma\rho^3$ cannot be placed above such $\frac{3(n+2)(2n+1)}{2}$. This implies that there exists $C(\epsilon) > 0$ depending on ϵ such that

$$\gamma < \frac{C(\epsilon)}{\rho^3}. \tag{2.10}$$

A little computation shows that $C(\epsilon)$ is

$$C(\epsilon) = \frac{3}{2}\left(\frac{6+\sqrt{36+18\epsilon}}{2\epsilon}+2\right)\left(\frac{6+\sqrt{36+18\epsilon}}{\epsilon}+1\right).$$

Combining (2.10) with (2.5) we see that ρ and γ are in a somewhat narrow parameter range

$$\rho < \delta, \quad \frac{3+\epsilon}{\rho^3} < \gamma < \frac{C(\epsilon)}{\rho^3}, \tag{2.11}$$

and $\gamma\rho^3$ must stay away from the sequence $\frac{3(n+2)(2n+1)}{2}$, $n = 2, 3, \ldots$, in the sense of (2.6). From (2.11) one sees that ρ must be small and γ be appropriately large.

We may assign a negative gradient flow to J and consider a dynamic counterpart of (1.1) (see [16]). The condition (2.5) prevents coarsening in such a dynamic process. By coarsening we mean that some balls become larger and other balls shrink and disappear.

The gap condition (2.6) controls interface oscillation. Interface oscillation refers to a phenomenon that oscillations appear on the boundary of a ball. The gap condition also suggests bifurcations to oscillating solutions. Elsewhere gap conditions have appeared in constructing layered solutions for singularly perturbed problems. See Malchiodi and Montenegro [12], del Pino, Kowalczyk, and Wei [8], Pacard and Ritoré [19], and the references therein.

The solution found in Theorem 2.1 may be unstable because of interface oscillation. The condition (2.8) in Theorem 2.2 eliminates this possibility. Under (2.8), ϵ must be no greater than 3, and ρ and γ must satisfy a more stringent requirement

$$\rho < \delta, \quad \frac{3+\epsilon}{\rho^3} < \gamma < \frac{30-4\epsilon}{\rho^3}. \tag{2.12}$$

This means that $\gamma\rho^3$ must stay to the left of the sequence $\frac{3(n+2)(2n+1)}{2}$, $n = 2, 3, \ldots$. If (2.9) holds, then we have an unstable mode that tends to bring oscillations to the spheres.

The spheres in the solution we construct are approximately round, with the same approximate radius. Theorem 2.1, part 3 asserts that the sphere centers must minimize F approximately.

We can even determine the optimal number of balls in a spherical pattern. Because of (2.11), we write

$$\gamma = \frac{\mu}{a} = \frac{\mu}{\left(\frac{4\pi K}{3|D|}\right)\rho^3}. \tag{2.13}$$

Now a and μ are the parameters of the problem. We hold μ fixed and make a and hence ρ small.

With (2.13) and (2.4) the leading order of the free energy is calculated from the formula in Lemma 8.1,

$$4\pi\rho^2 K + \frac{\gamma}{2}\left(\frac{8\pi\rho^5 K}{15}\right) = 4\pi K^{1/3}\left(\frac{3a|D|}{4\pi}\right)^{2/3} + \frac{\mu 4\pi}{15a}\left(\frac{3a|D|}{4\pi}\right)^{5/3} K^{-2/3}. \tag{2.14}$$

With respect to K the last quantity is minimized at

$$K = \frac{|D|\mu}{10\pi}. \tag{2.15}$$

Note that the choice (2.15) of K does not violate the condition (2.12), since, with this K,

$$\gamma = \frac{\mu}{a} = \mu\frac{3|D|}{K4\pi\rho^3} = \mu\frac{3|D|}{4\pi\rho^3}\frac{10\pi}{\mu|D|} = \frac{30}{4\rho^3}. \tag{2.16}$$

Equation (2.15) gives the optimal number of spheres in a spherical pattern.

3. Approximate solutions. Throughout the rest of this paper we are given $\epsilon > 0$, and γ and ρ satisfy (2.5) and (2.6).

Let U_1 be a small neighborhood in D^K of the set $\{\eta : F(\eta) = \min_{\xi\in D^K} F(\xi)\}$, and U_2 be the set

$$U_2 = \Big\{(r_1, r_2, \dots, r_K) \in R^K : r_k \in ((1-\delta_2)\rho, (1+\delta_2)\rho), \\ k = 1, 2, \dots, K, \quad \sum_{k=1}^{K} \frac{4\pi r_k^3}{3} = a|D|\Big\}. \tag{3.1}$$

The constant δ_2 is positive, small, and depends on ϵ. It will be fixed later in the proofs of Lemmas 5.3 and 8.2. Define

$$U = U_1 \times U_2. \tag{3.2}$$

Let $\xi_1, \xi_2, \dots, \xi_K$ be K distinct points in D such that $\xi = (\xi_1, \xi_2, \dots, \xi_K)$ is in U_1, and $r = (r_1, r_2, \dots, r_K)$ is in U_2. Denote the ball centered at ξ_k of radius r_k by B_k. The union of the B_k's is B:

$$B = \bigcup_{k=1}^{K} B_k = \bigcup_{k=1}^{K} \{x \in R^3 : |x - \xi_k| < r_k\}. \tag{3.3}$$

With U_1 close to $\{\eta : F(\eta) = \min_{\kappa\in D^K} F(\kappa)\}$ and ρ sufficiently small, the B_k's are all inside D and disjoint.

LEMMA 3.1. *When E is B, the left-hand side of* (1.1) *is*

$$\frac{1}{r_k} + \gamma\left[\frac{r_k^2}{3} + \frac{4\pi r_k^3}{3}R(\xi_k, \xi_k) + \sum_{l\neq k}\frac{4\pi r_l^3}{3}G(\xi_k, \xi_l)\right] + O(\rho)$$

at each $\xi_k + r_k\theta_k$, where $\theta_k \in S^2$ and S^2 is the unit sphere.

Proof. At a boundary point $\xi_k + r_k\theta_k$ of B_k, the curvature is $\frac{1}{r_k}$.

We compute $v_k = (-\Delta)^{-1}(\chi_{B_k} - \frac{4\pi r_k^3}{3|D|})$. Define

$$P_k(x) = \begin{cases} -\frac{|x-\xi_k|^2}{6} + \frac{r_k^2}{2} & \text{if } |x-\xi_k| < r_k \\ \frac{r_k^3}{3|x-\xi_k|}, & \text{if } |x-\xi_k| \geq r_k. \end{cases}$$

Then $-\Delta P_k = \chi_{B_k}$. Write $v_k(x) = P_k(x) + Q_k(x,\xi_k)$. Clearly

$$-\Delta Q_k(x,\xi_k) = -\frac{4\pi r_k^3}{3|D|},\ \partial_{\nu(x)} Q_k(x,\xi_k) = -\partial_\nu \frac{4\pi r_k^3}{3}\frac{1}{4\pi|x-\xi_k|}$$
$$\text{on } \partial D,\ \overline{Q_k(\cdot,\xi_k)} = -\overline{P_k}.$$

From (2.2) we see that $Q_k(x,\xi_k)$ and $\frac{4\pi r_k^3}{3}R(x,\xi_k)$ satisfy the same equation and the same boundary condition, where R is the regular part of the Green's function G. Therefore they can differ only by a constant. This constant is $\overline{Q_k(\cdot,\xi_k)} - \frac{4\pi r_k^3}{3}\overline{R(\cdot,\xi_k)}$. But $\overline{v_k} = \overline{G(\cdot,\xi_k)} = 0$ implies that this constant is

$$-\overline{P_k} + \frac{4\pi r_k^3}{3}\overline{\frac{1}{4\pi|x-\xi_k|}} = \frac{4\pi r_k^5}{3}\frac{1}{10|D|}$$

by direct calculation. Hence

$$Q_k(x,\xi_k) = \frac{4\pi r_k^3}{3}R(x,\xi_k) + \frac{4\pi}{3}\frac{r_k^5}{10|D|}$$

and

$$v_k(x) = P_k(x) + \frac{4\pi r_k^3}{3}R(x,\xi_k) + \frac{4\pi}{3}\frac{r_k^5}{10|D|}. \tag{3.4}$$

Let $v = (-\Delta)^{-1}(\chi_B - a) = \sum_l v_l$. Then at $\xi_k + r_k\theta_k$

$$\begin{aligned} v(\xi_k + r_k\theta_k) &= \frac{r_k^2}{3} + \frac{4\pi r_k^3}{3}R(\xi_k + r_k\theta_k, \xi_k) + \sum_{l\neq k}\frac{4\pi r_l^3}{3}G(\xi_k + r_k\theta_k, \xi_l) \\ &\quad + \sum_{l=1}^{K}\frac{4\pi}{3}\frac{r_l^5}{10|D|} \\ &= \frac{r_k^2}{3} + \frac{4\pi r_k^3}{3}R(\xi_k,\xi_k) + \sum_{l\neq k}\frac{4\pi r_l^3}{3}G(\xi_k,\xi_l) + O(\rho^4). \end{aligned} \tag{3.5}$$

The lemma follows from (2.10). ∎

LEMMA 3.2. *The free energy of B is*

$$\begin{aligned} J(B) = \sum_{k=1}^{K} 4\pi r_k^2 + \frac{\gamma}{2}\Bigg\{ &\sum_{k=1}^{K}\left[\frac{8\pi r_k^5}{15} + \left(\frac{4\pi}{3}\right)^2 r_k^6 R(\xi_k,\xi_k)\right] \\ &+ \sum_{k=1}^{K}\sum_{l=1, l\neq k}^{K}\left(\frac{4\pi}{3}\right)^2 r_k^3 r_l^3 G(\xi_k,\xi_l) + \sum_{k=1}^{K}\sum_{l=1}^{K}\left(\frac{4\pi}{3}\right)^2\left(\frac{r_k^3 r_l^5}{10|D|} + \frac{r_k^5 r_l^3}{10|D|}\right)\Bigg\}. \end{aligned}$$

Proof. The local part of the free energy is just $\sum_{k=1}^K 4\pi r_k^2$.
The nonlocal part of the free energy is

$$\begin{aligned}
&\int_D |(-\Delta)^{-1/2}(\chi_B - a)|^2\,dx \\
&\quad = \int_D (\chi_B - a)v(x)\,dx \; = \; \sum_{l=1}^K \int_{B_l} v(x)\,dx \\
&\quad = \sum_{l=1}^K \sum_{k=1}^K \int_{B_l} v_k(x)\,dx \; = \; \sum_{l=1}^K \sum_{k=1}^K \left[\int_{B_l} P_k(x)\,dx + \int_{B_l} Q_k(x,\xi_k)\,dx\right].
\end{aligned}$$

There are two possibilities. When $l = k$, from the definition of P_k we find

$$\int_{B_k} P_k(x)\,dx = \frac{8\pi r_k^5}{15}. \tag{3.6}$$

For the integral of Q_k, we have

$$\int_{B_k} Q_k(x,\xi_k)\,dx = \frac{4\pi r_k^3}{3}\int_{B_k} R(x,\xi_k)\,dx + \left(\frac{4\pi}{3}\right)^2 \frac{r_k^8}{10|D|}.$$

Since $R(x,\xi_k) - \frac{1}{6|D|}|x-\xi_k|^2$ is harmonic in x, by the mean value theorem for harmonic functions

$$\begin{aligned}
\int_{B_k} R(x,\xi_k)\,dx &= \int_{B_k}\left(R(x,\xi_k) - \frac{1}{6|D|}|x-\xi_k|^2\right)dx + \int_{B_k}\frac{1}{6|D|}|x-\xi_k|^2\,dx \\
&= \frac{4\pi r_k^3}{3}R(\xi_k,\xi_k) + \frac{4\pi}{3}\frac{r_k^5}{10|D|}. \qquad (3.7)
\end{aligned}$$

Hence

$$\int_{B_k} v_k\,dx = \frac{8\pi r_k^5}{15} + \left(\frac{4\pi}{3}\right)^2 r_k^6 R(\xi_k,\xi_k) + \left(\frac{4\pi}{3}\right)^2 \frac{r_k^8}{5|D|}.$$

When $l \neq k$, for $x \in B_l$, since P_k is harmonic,

$$\begin{aligned}
\int_{B_l} v_k\,dx &= \int_{B_l} P_k\,dx + \frac{4\pi r_k^3}{3}\int_{B_l} R(x,\xi_k)\,dx + \left(\frac{4\pi}{3}\right)^2 \frac{r_k^5 r_l^3}{10|D|} \\
&= \frac{4\pi}{3} r_l^3 \frac{r_k^3}{3|\xi_k - \xi_l|} + \frac{4\pi r_k^3}{3}\left[\int_{B_l}\left(R(x,\xi_k) - \frac{1}{6|D|}|x-\xi_l|^2\right)dx\right. \\
&\quad \left. + \int_{B_l}\frac{1}{6|D|}|x-\xi_l|^2\,dx\right] + \left(\frac{4\pi}{3}\right)^2 \frac{r_k^5 r_l^3}{10|D|} \\
&= \left(\frac{4\pi}{3}\right)^2 \frac{r_k^3 r_l^3}{4\pi|\xi_k - \xi_l|} + \left(\frac{4\pi}{3}\right)^2 r_k^3 r_l^3 R(\xi_k,\xi_l) + \left(\frac{4\pi}{3}\right)^2 \left(\frac{r_k^3 r_l^5}{10|D|} + \frac{r_k^5 r_l^3}{10|D|}\right).
\end{aligned}$$

Finally the nonlocal part of the free energy is

$$\begin{aligned}\int_D (\chi_B - a) v\, dx = &\sum_{k=1}^K \left[\frac{8\pi r_k^5}{15} + \left(\frac{4\pi}{3}\right)^2 r_k^6 R(\xi_k, \xi_k) \right] \\ &+ \sum_{k=1}^K \sum_{l=1, l\neq k}^K \left[\left(\frac{4\pi}{3}\right)^2 \frac{r_k^3 r_l^3}{4\pi|\xi_k - \xi_l|} + \left(\frac{4\pi}{3}\right)^2 r_k^3 r_l^3 R(\xi_k, \xi_l) \right] \\ &+ \sum_{k=1}^K \sum_{l=1}^K \left(\frac{4\pi}{3}\right)^2 \left(\frac{r_k^3 r_l^5}{10|D|} + \frac{r_k^5 r_l^3}{10|D|} \right). \end{aligned} \tag{3.8}$$

The lemma now follows. □

4. Perturbed spheres. We perturb each ball B_k considered in the last section. A perturbed ball denoted by E_{ϕ_k} is described by a function $\phi_k = \phi_k(\theta_k)$, $\theta_k \in S^2$:

$$E_{\phi_k} = \{\xi_k + t\theta_k : \ \theta_k \in S^2, \ t \in [0, (r_k^3 + \phi_k(\theta_k))^{1/3}]\}. \tag{4.1}$$

Each ϕ_k is small compared to r_k^3 so that $r_k^3 + \phi_k(\theta_k)$ is positive. Each θ_k is identified by its longitude and latitude $(\theta_{k,1}, \theta_{k,2})$, namely

$$\theta_k = (\cos\theta_{k,1} \sin\theta_{k,2}, \ \sin\theta_{k,1} \sin\theta_{k,2}, \ \cos\theta_{k,2}). \tag{4.2}$$

The ϕ_k's satisfy

$$\sum_{k=1}^K \int_{S^2} \phi_k(\theta_k)\, d\theta_k = 0. \tag{4.3}$$

Here the integral is a surface integral over S^2 and

$$d\theta_k = \sin\theta_{k,2}\, d\theta_{k,1} d\theta_{k,2} \tag{4.4}$$

is the surface element on S^2. Hence the total volume inside the perturbed spheres remains fixed:

$$\begin{aligned}\sum_{k=1}^K |E_{\phi_k}| &= \sum_k \int_{S^2} \int_0^{(r_k^3 + \phi_k(\theta_k))^{1/3}} t^2\, dt d\theta_k \\ &= \sum_k \int_{S^2} \left(\frac{r_k^3}{3} + \frac{\phi_k(\theta_k)}{3} \right) d\theta_k = \sum_k \frac{4\pi r_k^3}{3} = a|D|. \end{aligned}$$

The union of the E_{ϕ_k}'s is E_ϕ:

$$E_\phi = \bigcup_{k=1}^K E_{\phi_k}. \tag{4.5}$$

With these notations $B = E_0$.

We let $\theta = (\theta_1, \theta_2, \dots, \theta_K)$ and $\phi(\theta) = (\phi_1(\theta_1), \phi_2(\theta_2), \dots \phi_K(\theta_K))$. To express surface area in terms of ϕ_k, first define

$$L(s, p, q, \beta) = s^{-1/3} \sqrt{\frac{p^2}{9\sin^2\beta} + \frac{q^2}{9} + s^2}, \tag{4.6}$$

and then define

$$L_k\left(\phi_k, \frac{\partial \phi_k}{\partial \theta_{k,1}},\ \frac{\partial \phi_k}{\partial \theta_{k,2}},\ \theta_{k,2}\right) = r_k^2 L\left(1+\frac{\phi_k}{r_k^3},\ \frac{1}{r_k^3}\frac{\partial \phi_k}{\partial \theta_{k,1}},\ \frac{1}{r_k^3}\frac{\partial \phi_k}{\partial \theta_{k,2}},\ \theta_{k,2}\right). \tag{4.7}$$

The surface area of $\partial_D E_\phi$ can be expressed as

$$\sum_{k=1}^K |D\chi_{E_{\phi_k}}|(D) = \sum_{k=1}^K \int_{S^2} L_k\left(\phi_k, \frac{\partial \phi_k}{\partial \theta_{k,1}},\ \frac{\partial \phi_k}{\partial \theta_{k,2}},\ \theta_{k,2}\right) d\theta_k. \tag{4.8}$$

The nonlocal part of J in (1.2) may be written in terms of ϕ as

$$\frac{\gamma}{2}\int_D |(-\Delta)^{-1/2}(\chi_{E_\phi}-a)|^2\,dx = \frac{\gamma}{2}\int_{E_\phi}\int_{E_\phi} G(x,y)\,dxdy. \tag{4.9}$$

The first variation of J can now be written as

$$J'(E_\phi)(w) = \sum_{k=1}^K \int_{S^2}\left[\frac{\partial L_k}{\partial \phi_k}w_k + \frac{\partial L_k}{\partial \phi_{k,1}}w_{k,1} + \frac{\partial L_k}{\partial \phi_{k,2}}w_{k,2}\right] d\theta_k \tag{4.10}$$

$$+ \sum_{k=1}^K \int_{S^2} w_k(\theta_k)\left[\sum_{l=1}^K \frac{\gamma}{3}\int_{E_{\phi_l}} G(\xi_k+(r_k^3+\phi_k(\theta_k))^{1/3}\theta_k, y)\,dy\right] d\theta_k. \tag{4.11}$$

Here we have used shorthand notations $\phi_{k,1} = \frac{\partial \phi_k}{\partial \theta_{k,1}}$ and $\phi_{k,2} = \frac{\partial \phi_k}{\partial \theta_{k,2}}$, and so on. From (4.10) we define a second order, quasilinear elliptic operator

$$\mathcal{H}_k(\phi_k)(\theta_k) = \frac{1}{\sin\theta_{k,2}}\left[\frac{\partial L_k}{\partial \phi_k}\sin\theta_{k,2} - \frac{\partial}{\partial \theta_{k,1}}\left(\frac{\partial L_k}{\partial \phi_{k,1}}\sin\theta_{k,2}\right) - \frac{\partial}{\partial \theta_{k,2}}\left(\frac{\partial L_k}{\partial \phi_{k,2}}\sin\theta_{k,2}\right)\right]. \tag{4.12}$$

This is just the mean curvature of the perturbed sphere ∂E_{ϕ_k} at $\xi_k+(r_k^3+\phi_k(\theta_k))^{1/3}\theta_k$, multiplied by $\frac{1}{3}$. The second part (4.11) of the first variation of J gives rise to a nonlocal operator

$$\phi \to \sum_{l=1}^K \frac{\gamma}{3}\int_{E_{\phi_l}} G(\xi_k+(r_k^3+\phi_k(\theta_k))^{1/3}\theta_k, y)\,dy. \tag{4.13}$$

This is just

$$\frac{\gamma}{3}(-\Delta)^{-1}(\chi_{E_\phi}-a)(\xi_k+(r_k^3+\theta_k)^{1/3}\theta_k),$$

the nonlocal part of (1.1) at $\xi_k+(r_k^3+\phi_k(\theta_k))^{1/3}\theta_k$ multiplied by $\frac{1}{3}$.

There are two cases in the sum over l in (4.13); when $l=k$ we write

$$\frac{\gamma}{3}\int_{E_{\phi_k}} G(\xi_k+(r_k^3+\phi_k(\theta_k))^{1/3}\theta_k, y)\,dy$$
$$= \frac{\gamma}{3}\int_{E_{\phi_k}} \frac{dy}{4\pi|\xi_k+(r_k^3+\phi_k(\theta_k))^{1/3}\theta_k - y|} + \frac{\gamma}{3}\int_{E_{\phi_k}} R(\xi_k+(r_k^3+\phi_k(\theta_k))^{1/3}\theta_k, y)\,dy.$$

We denote the last two terms by

$$\mathcal{A}_k(\phi_k)(\theta_k) = \frac{\gamma}{3}\int_{E_{\phi_k}} \frac{dy}{4\pi|\xi_k + (r_k^3 + \phi_k(\theta_k))^{1/3}\theta_k - y|}, \tag{4.14}$$

$$\mathcal{B}_k(\phi_k)(\theta_k) = \frac{\gamma}{3}\int_{E_{\phi_k}} R(\xi_k + (r_k^3 + \phi_k(\theta_k))^{1/3}\theta_k, y)\, dy. \tag{4.15}$$

When $l \neq k$ in (4.13) we let

$$\mathcal{C}_{kl}(\phi_k, \phi_l)(\theta_k) = \frac{\gamma}{3}\int_{E_{\phi_l}} G(\xi_k + (r_k^3 + \phi_k(\theta_k))^{1/3}\theta_k, y)\, dy. \tag{4.16}$$

The left-hand side of (1.1) (multiplied by $\frac{1}{3}$) now becomes

$$\mathcal{H}_k(\phi_k)(\theta_k) + \mathcal{A}_k(\phi_k)(\theta_k) + \mathcal{B}_k(\phi_k)(\theta_k) + \sum_{l\neq k}\mathcal{C}_{kl}(\phi_k, \phi_l)(\theta_k)$$

at $\xi_k + (r_k^3 + \phi_k(\theta_k))^{1/3}\theta_k$. Let us define

$$\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_K), \tag{4.17}$$

where

$$\mathcal{S}_k(\phi)(\theta_k) = \mathcal{H}_k(\phi_k)(\theta_k) + \mathcal{A}_k(\phi_k)(\theta_k) + \mathcal{B}_k(\phi_k)(\theta_k) + \sum_{l\neq k}\mathcal{C}_{kl}(\phi_k, \phi_l)(\theta_k) + \lambda(\phi). \tag{4.18}$$

Here $\lambda(\phi)$ is a number, independent of k. It is given by

$$\lambda(\phi) = -\frac{1}{K}\sum_{k=1}^{K}\overline{\left[\mathcal{H}_k(\phi_k) + \mathcal{A}_k(\phi_k) + \mathcal{B}_k(\phi_k) + \sum_{l\neq k}\mathcal{C}_{kl}(\phi_k, \phi_l)\right]}. \tag{4.19}$$

The bar over the quantity here stands for the average of the quantity over S^2. With this definition of λ,

$$\sum_{k=1}^{K}\overline{\mathcal{S}_k(\phi_k)} = 0. \tag{4.20}$$

The operator $\mathcal{S}$ maps from

$$\mathcal{X} = \left\{\phi = (\phi_1, \phi_2, \ldots, \phi_K):\ \phi_k \in W^{2,p}(S^2),\ k = 1, 2, \ldots, K,\ \ \sum_{k=1}^{K}\overline{\phi_k} = 0\right\} \tag{4.21}$$

to

$$\mathcal{Y} = \left\{q = (q_1, q_2, \ldots, q_K):\ q_k \in L^p(S^2),\ k = 1, 2, \ldots, K,\ \ \sum_{k=1}^{K}\overline{q_k} = 0\right\}. \tag{4.22}$$

For technical reasons p is assumed to be in the range

$$2 < p < \infty. \tag{4.23}$$

This guarantees that $D\phi_k$ is continuous, a fact needed in the proof of Lemma 6.1. Equation (1.1) now becomes

$$\mathcal{S}(\phi) = 0. \tag{4.24}$$

By defining

$$\mathcal{C} = (\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K), \text{ where } \mathcal{C}_k(\phi_1, \phi_2, \dots, \phi_K) = \sum_{l \neq k} \mathcal{C}_{kl}(\phi_k, \phi_l), \tag{4.25}$$

we write

$$\mathcal{S} = \mathcal{H} + \mathcal{A} + \mathcal{B} + \mathcal{C} + \lambda. \tag{4.26}$$

In the map $\mathcal{S}$ the inputs $\phi_1, \phi_2, \dots, \phi_k$ interact only in $\mathcal{C}$ and λ. The other operators can be written in the block matrix form

$$\mathcal{H} = \begin{bmatrix} \mathcal{H}_1 & 0 & \dots & 0 \\ 0 & \mathcal{H}_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mathcal{H}_K \end{bmatrix}, \ \mathcal{A} = \begin{bmatrix} \mathcal{A}_1 & 0 & \dots & 0 \\ 0 & \mathcal{A}_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mathcal{A}_K \end{bmatrix}, \ \mathcal{B} = \begin{bmatrix} \mathcal{B}_1 & 0 & \dots & 0 \\ 0 & \mathcal{B}_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mathcal{B}_K \end{bmatrix}, \tag{4.27}$$

where each entry in a matrix is an operator from $W^{2,p}(S^2)$ to $L^p(S^2)$. The scalar operator λ gives the projection $-(\lambda(\phi), \lambda(\phi), \dots, \lambda(\phi))$ of $\mathcal{H}(\phi) + \mathcal{A}(\phi) + \mathcal{B}(\phi) + \mathcal{C}(\phi)$ to the one-dimensional space spanned by $(1, 1, \dots, 1)$.

Let us write the first Fréchet derivatives of these operators. For simplicity we write

$$\phi_{k,i} = \frac{\partial \phi_k}{\partial \theta_{k,i}}, \ \phi_{k,ij} = \frac{\partial^2 \phi_k}{\partial \theta_{k,ij}}, \ u_{k,i} = \frac{\partial u_k}{\partial \theta_{k,i}}, \ u_{k,ij} = \frac{\partial^2 u_k}{\partial \theta_{k,ij}}.$$

Calculations show that

$$\mathcal{H}_k'(\phi_k)(u_k) = \frac{\partial \mathcal{H}_k}{\partial \phi_k} u_k + \sum_{i=1}^{2} \frac{\partial \mathcal{H}_k}{\partial \phi_{k,i}} u_{k,i} + \sum_{i,j=1}^{2} \frac{\mathcal{H}_k}{\partial \phi_{k,ij}} u_{k,ij}, \tag{4.28}$$

$$\begin{aligned} \mathcal{A}_k'(\phi_k)(u_k)(\theta_k) &= \frac{\gamma}{9} \int_{S^2} \frac{u_k(\omega_k)\, d\omega_k}{4\pi |(r_k^3 + \phi_k(\theta_k))^{1/3}\theta_k - (r_k^3 + \phi_k(\omega_k))^{1/3}\omega_k|} \\ &\quad - \frac{\gamma u_k(\theta_k)}{9(r_k^3 + \phi_k(\theta_k))^{2/3}} \int_{\tilde{E}_{\phi_k}} \frac{((r_k^3 + \phi_k(\theta_k))^{1/3}\theta_k - y) \cdot \theta_k}{4\pi |(r_k^3 + \phi_k(\theta_k))^{1/3}\theta_k - y|^3}\, dy. \end{aligned} \tag{4.29}$$

$$\begin{aligned} &\mathcal{B}_k'(\phi_k)(u_k)(\theta_k) \\ &\quad = \frac{\gamma}{9} \int_{S^2} u_k(\omega_k) R(\xi_k + (r_k^3 + \phi_k(\theta_k))^{1/3}\theta_k, \xi_k + (r_k^3 + \phi_k(\omega_k))^{1/3}\omega_k)\, d\omega_k \\ &\quad + \frac{\gamma u_k(\theta_k)}{9(r_k^3 + \phi_k(\theta_k))^{2/3}} \int_{E_{\phi_k}} \nabla R(\xi_k + (r_k^3 + \phi_k(\theta_k))^{1/3}\theta_k, y) \cdot \theta_k\, dy. \end{aligned} \tag{4.30}$$

$$\begin{aligned} &\mathcal{C}_{kl}'(\phi_k, \phi_l)(u_k, u_l)(\theta_k) \\ &\quad = \frac{\gamma}{9} \int_{S^2} u_l(\omega_l) G(\xi_k + (r_k^3 + \phi_k(\theta_k))^{1/3}\theta_k, \xi_l + (r_l^3 + \phi_l(\omega_l))^{1/3}\omega_l)\, d\omega_l \\ &\quad + \frac{\gamma u_k(\theta_k)}{9(r_k^3 + \phi_k(\theta_k))^{2/3}} \int_{E_{\phi_l}} \nabla G(\xi_k + (r_k^3 + \phi_k(\theta_k))^{1/3}\theta_k, y) \cdot \theta_k\, dy. \end{aligned} \tag{4.31}$$

In $\mathcal{A}'_k$, $\tilde{E}_{\phi_k} = E_{\phi_k} - \xi_k$ is a shift of E_{ϕ_k}. The center of $\tilde{E}_{\phi_k}$ is 0. The derivative

$$\lambda'(\phi_1, \phi_2, \ldots, \phi_K)(u_1, u_2, \ldots, u_K) \tag{4.32}$$

is so chosen that

$$\sum_{k=1}^{K} \overline{\mathcal{S}'_k(u)} = 0. \tag{4.33}$$

5. A linear operator. Let $\mathcal{L}$ be the linearized operator of $\mathcal{S}$ at $\phi = 0$, i.e.,

$$\mathcal{L} = \mathcal{S}'(0). \tag{5.1}$$

Going back to (4.28), (4.29), (4.30), and (4.31) we find that

$$\begin{aligned}
\mathcal{H}'_k(0)(u_k) &= -\frac{1}{9r_k^4}\left[\frac{1}{\sin^2\theta_{k,2}}\frac{\partial^2 u_k}{\partial\theta_{k,1}^2} + \frac{\partial^2 u_k}{\partial\theta_{k,2}^2} + \cot\theta_{k,2}\frac{\partial u_k}{\partial\theta_{k,2}}\right] - \frac{2}{9r_k^4}u,\\
\mathcal{A}'_k(0)(u_k)(\theta_k) &= \frac{\gamma}{9r_k}\int_{S^2}\frac{u_k(\omega_k)\,d\omega_k}{4\pi|\theta_k-\omega_k|} - \frac{\gamma u_k(\theta_k)}{27r_k},\\
\mathcal{B}'_k(0)(u_k)(\theta_k) &= \frac{\gamma}{9}\int_{S^2} u_k(\omega_k)R(\xi_k + r_k\theta_k, \xi_k + r_k\omega_k)\,d\omega_k\\
&\quad + \frac{\gamma u_k(\theta_k)}{9r_k^2}\int_{B_k}\nabla R(\xi_k + r_k\theta_k, y)\cdot\theta_k\,dy,\\
\mathcal{C}'_{kl}(0,0)(u_k,u_l)(\theta_k) &= \frac{\gamma}{9}\int_{S^2} u_l(\omega_l)G(\xi_k + r_k\theta_k, \xi_l + r_l\omega_l)\,d\omega_l\\
&\quad + \frac{\gamma u_k(\theta_k)}{9r_k^2}\int_{B_l}\nabla G(\xi_k + r_k\theta_k, y)\cdot\theta_k\,dy.
\end{aligned}$$

The derivation of $\mathcal{A}'_k(0)$ is explained in more detail in Appendix A.

Let us separate $\mathcal{L}$ to a dominant part $\mathcal{L}_1$ and a minor part $\mathcal{L}_2$. We define $\mathcal{L}_{1,k}$, the kth component of $\mathcal{L}_1$, to be

$$\mathcal{L}_{1,k}(u)(\theta_k) = \mathcal{H}'_k(0)(u_k)(\theta_k) + \mathcal{A}'_k(0)(u_k)(\theta_k) + l_1(u).$$

The real valued linear operator l_1 is independent of k. It is so chosen that $\mathcal{L}_1$ maps from $\mathcal{X}$ to $\mathcal{Y}$. The rest of $\mathcal{L}$ is denoted by $\mathcal{L}_2$.

We are more interested in the operators $\Pi\mathcal{L}$ and $\Pi\mathcal{L}_1$, where Π is the orthogonal projection operator from $\mathcal{Y}$ to

$$\mathcal{Y}_* = \{q = (q_1, \ldots, q_K) \in \mathcal{Y} :\ q_k \perp H_1,\ q_k \perp 1,\ k = 1, \ldots, K\}. \tag{5.2}$$

Here H_1 is the space of spherical harmonics of degree 1. See, for instance, [10] for more on spherical harmonics. The operator $\Pi\mathcal{L}$ is defined on

$$\mathcal{X}_* = \{\phi = (\phi_1, \ldots, \phi_K) \in \mathcal{X} :\ \phi_k \perp H_1,\ \phi_k \perp 1,\ k = 1, \ldots, K\}. \tag{5.3}$$

We use the same Π to denote the orthogonal projection from

$$L^2(S^2) \text{ to } \{q_k \in L^2(S^2) : q_k \perp H_1,\ q_k \perp 1\}. \tag{5.4}$$

LEMMA 5.1. *Consider $\Pi\mathcal{L}_1$ as an operator from $\mathcal{X}_*$ to $\mathcal{Y}_*$. The eigenvalues of $\Pi\mathcal{L}_1$ are*

$$\lambda_{k,n} = \frac{(n-1)(n+2)}{9r_k^4} - \frac{\gamma}{9r_k}\left[\frac{2(n-1)}{3(2n+1)}\right], \quad k = 1,2,\dots,K, \ n = 2,3,4,\dots \tag{5.5}$$

whose multiplicity is $2n+1$. The corresponding eigenvectors are the spherical harmonics of degree n; i.e., H_n is the eigenspace associated with $\lambda_{k,n}$.

Proof. In $\mathcal{X}_*$, $\mathcal{L}_1$ is simplified to

$$\begin{aligned}\mathcal{L}_{1,k}(u) = &-\frac{1}{9r_k^4}\left[\frac{1}{\sin^2\theta_{k,2}}\frac{\partial^2 u_k}{\partial\theta_{k,1}^2} + \frac{\partial^2 u_k}{\partial\theta_{k,2}^2} + \cot\theta_{k,2}\frac{\partial u_k}{\partial\theta_{k,2}}\right] - \frac{2u_k}{9r_k^4}\\ &+ \frac{\gamma}{9r_k}\int_{S^2}\frac{u_k(\omega_k)\,d\omega_k}{4\pi|\theta_k-\omega_k|} - \frac{\gamma u_k(\theta_k)}{27r_k},\end{aligned}$$

for each k. This is a diagonalized operator. Note that in $\mathcal{X}_*$, $\Pi\mathcal{L}_1 = \mathcal{L}_1$. To find the spectrum of $\mathcal{L}_1$ in $\mathcal{X}_*$ we consider the effect of $\mathcal{L}_1$ on the spherical harmonics $h \in H_n$ of degree n. Since

$$\frac{1}{\sin^2\theta_{k,2}}\frac{\partial^2}{\partial\theta_{k,1}^2} + \frac{\partial^2}{\partial\theta_{k,2}^2} + \cot\theta_2\frac{\partial}{\partial\theta_{k,2}} := \Delta_{S^2} \tag{5.6}$$

is the Laplacian–Beltrami operator on the unit sphere,

$$-\left[\frac{1}{\sin^2\theta_{k,2}}\frac{\partial^2 h}{\partial\theta_{k,1}^2} + \frac{\partial^2 h}{\partial\theta_{k,2}^2} + \cot\theta_2\frac{\partial h}{\partial\theta_{k,2}}\right] = n(n+1)h. \tag{5.7}$$

In Appendix B we find that

$$\int_{S^2}\frac{h(\omega)\,d\omega}{4\pi|\theta-\omega|} = \frac{h(\theta)}{2n+1}. \tag{5.8}$$

Following (5.7) and (5.8) one deduces that

$$\mathcal{L}_{1,k}(h) = \left[\frac{n(n+1)-2}{9r_k^4} + \frac{\gamma}{9r_k}\left(\frac{1}{2n+1} - \frac{1}{3}\right)\right]h. \tag{5.9}$$

This proves the lemma. □

The second part of $\mathcal{L}$ is minor.

LEMMA 5.2. *There exists $C > 0$ independent of ξ, r, ρ, and γ such that*

$$\|\mathcal{L}_2(u)\|_{L^p} \le \frac{C}{\rho^2}\|u\|_{L^p}$$

for all $u \in \mathcal{Y}_*$. A similar estimate holds if the two p's above are replaced by 2.

Proof. Let $\mathcal{L}_{2,k}$ be the kth component of $\mathcal{L}_2$. Then

$$\begin{aligned}\mathcal{L}_{2,k}(u)(\theta_k) = &\frac{\gamma}{9}\int_{S^2} u_k(\omega_k)(R(\xi_k + r_k\theta_k, \xi_k + r_k\omega_k) - R(\xi_k,\xi_k))\,d\omega_k\\ &+ \frac{\gamma u_k(\theta_k)}{9r_k^2}\int_{B_k}\nabla R(\xi_k + r_k\theta_k, y)\cdot\theta_k\,dy\\ &+ \sum_{l\neq k}\frac{\gamma}{9}\int_{S^2} u_l(\omega_l)(G(\xi_k + r_k\theta_k, \xi_l + r_l\omega_l) - G(\xi_k,\xi_l))\,d\omega_l\\ &+ \sum_{l\neq k}\frac{\gamma u_k(\theta_k)}{9r_k^2}\int_{B_l}\nabla G(\xi_k + r_k\theta_k, y)\cdot\theta_k\,dy\\ &+ l_2(u),\end{aligned}$$

where $l_2(u)$ is real valued and independent of k. It is included so that $\mathcal{L}_2(u)$ is in $\mathcal{Y}$.

Because

$$\begin{aligned}R(\xi_k + r_k\theta_k, \xi_k + r_k\omega_k) - R(\xi_k,\xi_k) &= O(\rho),\\ G(\xi_k + r_k\theta_k, \xi_l + r_l\omega_l) - G(\xi_k,\xi_l) &= O(\rho),\end{aligned}$$

we obtain that

$$\left\|\frac{\gamma}{9}\int_{S^2} u_k(\omega_k)(R(\xi_k + r_k\theta_k, \xi_k + r_k\omega_k) - R(\xi_k,\xi_k))\,d\omega_k\right\|_{L^p} \le C\gamma\rho\|u\|_{L^p}$$
$$\left\|\frac{\gamma}{9}\int_{S^2} u_l(\omega_l)(G(\xi_k + r_k\theta_k, \xi_l + r_l\omega_l) - G(\xi_k,\xi_l))\,d\omega_l\right\|_{L^p} \le C\gamma\rho\|u_k\|_{L^p}.$$

Since the volume of B_k is $\frac{4\pi r_k^3}{3}$,

$$\left\|\frac{\gamma u_k(\theta_k)}{9r_k^2}\int_{B_k}\nabla R(\xi_k + r_k\theta_k, y)\cdot\theta_k\,dy\right\|_{L^p} \le C\gamma\rho\|u_k\|_{L^p}$$
$$\left\|\frac{\gamma u_k(\theta_k)}{9r_k^3}\int_{B_l}\nabla G(\xi_k + r_k\theta_k, y)\cdot\theta_k\,dy\right\|_{L^p} \le C\gamma\rho\|u_k\|_{L^p}.$$

The condition

$$\sum_{k=1}^{K}\overline{\mathcal{L}_{2,k}(u)(\theta_k)} = 0$$

implies that

$$|l_2(u)| \le C\gamma\rho\|u\|_{L^p}.$$

The lemma then follows, with the help of (2.10). $\square$

LEMMA 5.3.

1. *For* $u \in \mathcal{X}_*$

$$\|u\|_{W^{2,p}} \le C\rho^4\|\Pi\mathcal{L}u\|_{L^p}.$$

2. *The operator* $\Pi\mathcal{L}$ *is invertible from* $\mathcal{X}_*$ *to* $\mathcal{Y}_*$.
3. *If* (2.8) *holds,*

$$\|u\|_{W^{1,2}}^2 \le C\rho^4\langle\Pi\mathcal{L}u, u\rangle.$$

Proof. From Lemma 5.1 we have

$$\frac{|\lambda_{k,n}|}{n^2} = \frac{n-1}{9r_k^4 n}\left|\frac{n+2}{n} - \frac{2\gamma r_k^3}{3(2n+1)n}\right| > \frac{n-1}{18r_k^4 n}\left|\frac{n+2}{n} - \frac{2\gamma\rho^3}{3(2n+1)n}\right|$$

if δ_2 in the definition (3.1) of U_2 is small enough. Then (2.6) implies that

$$\frac{|\lambda_{k,n}|}{n^2} > \frac{(n-1)}{18r_k^4 n}\frac{2\epsilon n}{3(2n+1)} \geq \frac{C}{\rho^4}, \quad n = 2, 3, \ldots.$$

If we expand u_k by spherical harmonics

$$u_k = \sum_{n=2}^{\infty}\sum_{l=1}^{2n+1} c_{n,l}h_{n,l},$$

where $h_{n,l}$, $l = 1, \ldots, 2n+1$, form an orthonormal basis in H_n, then

$$-\Delta_{S^2}u_k = \sum_{n=2}^{\infty}\sum_{l=1}^{2n+1} n(n+1)c_{n,l}h_{n,l}, \ \mathcal{L}_{1,k}u_k = \sum_{n=2}^{\infty}\sum_{l=1}^{2n+1} \lambda_{k,n}c_{n,l}h_{n,l}.$$

Our estimate on $|\lambda_{k,n}|$ shows that

$$\|\Delta_{S^2}u_k\|_{L^2}^2 = \sum_{n=2}^{\infty}\sum_{l=1}^{2n+1} n^2(n+1)^2c_{n,l}^2 \leq C\rho^8\sum_{n=2}^{\infty}\sum_{l=1}^{2n+1}\lambda_{k,n}^2c_{n,l}^2 = C\rho^8\|\mathcal{L}_{1,k}u_k\|_{L^2}^2.$$

The standard elliptic theory implies that

$$\|u\|_{W^{2,2}} \leq C\|\Delta_{S^2}u\|_{L^2} \leq C\rho^4\|\Pi\mathcal{L}_1(u)\|_{L^2}. \tag{5.10}$$

To prove part 1 of Lemma 5.3, we divide $\Pi\mathcal{L}_1$ into

$$\Pi\mathcal{L}_{1,k} = -\frac{1}{9r_k^4}\Delta_{S^2} + \mathcal{M}_k, \tag{5.11}$$

where Δ_{S^2} is defined in (5.6), and $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_K)$ is defined by (5.11). The standard elliptic estimate asserts that

$$\|u_k\|_{W^{2,p}} \leq C\|\Delta_{S^2}u_k\|_{L^p},$$

which by (5.11) is turned to

$$\begin{aligned}\|u_k\|_{W^{2,p}} &\leq C\|9r_k^4\mathcal{M}_k u - 9r_k^4\Pi\mathcal{L}_{1,k}u\|_{L^p}\\ &\leq C\rho^4(\|\mathcal{M}_k u\|_{L^p} + \|\Pi\mathcal{L}_{1,k}u\|_{L^p}).\end{aligned}$$

One observes that

$$\|\mathcal{M}u\|_{L^p} \leq \frac{C}{\rho^4}\|u\|_{L^p} \leq \frac{C}{\rho^4}\|u\|_{W^{2,2}},$$

where the last inequality comes from the Sobolev embedding $W^{2,2}(S^2) \to W^{1,p}(S^2) \subset L^p(S^2)$ for any $p \geq 1$. Hence when $p > 2$, by (5.10) we deduce that

$$\begin{aligned}\|u_k\|_{W^{2,p}} &\leq C\rho^4(\rho^{-4}\|u\|_{W^{2,2}} + \|\Pi\mathcal{L}_{1,k}u\|_{L^p})\\ &\leq C\rho^4(\|\Pi\mathcal{L}_{1,k}u\|_{L^2} + \|\Pi\mathcal{L}_{1,k}u\|_{L^p})\\ &\leq C\rho^4\|\Pi\mathcal{L}_{1,k}u\|_{L^p}.\end{aligned}$$

Lemma 5.2 implies that

$$\|\Pi\mathcal{L}u\|_{L^p} \geq \|\Pi\mathcal{L}_1 u\|_{L^p} - \|\Pi\mathcal{L}_2 u\|_{L^p} \geq \frac{C}{\rho^4}\|u\|_{W^{2,p}} - \frac{C}{\rho^2}\|u\|_{L^p} \geq \frac{C}{\rho^4}\|u\|_{W^{2,p}}$$

for small ρ. This proves part 1 of Lemma 5.3.

Part 2 of Lemma 5.3 follows from the Fredholm alternative.

When (2.8) holds,

$$\frac{\lambda_{k,n}}{n^2} = \frac{n-1}{9r_k^4 n}\left(\frac{n+2}{n} - \frac{2\gamma r_k^3}{3(2n+1)n}\right) > \frac{n-1}{18r_k^4 n}\frac{2\epsilon n}{3(2n+1)} \geq \frac{C}{\rho^4}, \quad n = 2, 3, \ldots,$$

if δ_2 in (3.1) is small. This implies that, with the help of expansion by spherical harmonics,

$$\begin{aligned}\langle \Pi\mathcal{L}_{1,k}(u_k), u_k\rangle &= \sum_{n=2}^{\infty}\sum_{l=1}^{2n+1}\lambda_{k,n}c_{n,l}^2 \;\geq\; \frac{C}{\rho^4}\sum_{n=2}^{\infty}\sum_{l=1}^{2n+1} n(n+1)c_{n,l}^2 \\ &= \frac{C}{\rho^4}\langle -\Delta_{S^2}u_k, u_k\rangle \;=\; \frac{C}{\rho^4}\langle \nabla u_k, \nabla u_k\rangle \;\geq\; \frac{C}{\rho^4}\|u\|_{W^{1,2}}^2.\end{aligned}$$

Using the estimate of Lemma 5.2 with p replaced by 2, we find that

$$\langle \Pi\mathcal{L}(u), u\rangle = \langle \Pi\mathcal{L}_1(u), u\rangle + \langle \Pi\mathcal{L}_2(u), u\rangle \geq \frac{C}{\rho^4}\|u\|_{W^{1,2}}^2 - \frac{C}{\rho^2}\|u\|_{L^2}^2 \geq \frac{C}{\rho^4}\|u\|_{W^{1,2}}^2.$$

This proves part 3 of Lemma 5.3. □

6. The second Fréchet derivative.

LEMMA 6.1. *Suppose that $\|\phi\|_{W^{2,p}} \leq c\rho^3$, where c is sufficiently small. The following estimates hold:*

1. $\|\mathcal{H}_k''(\phi_k)(u_k, v_k)\|_{L^p} \leq \frac{C}{\rho^7}\|u_k\|_{W^{2,p}}\|v_k\|_{W^{2,p}}$.
2. $\|\mathcal{A}_k''(\phi_k)(u_k, v_k)\|_{L^p} \leq \frac{C}{\rho^7}\|u_k\|_{W^{1,p}}\|v_k\|_{W^{1,p}}$.
3. $\|\mathcal{B}_k''(\phi_k)(u_k, v_k)\|_{L^p} \leq \frac{C}{\rho^5}\|u_k\|_{W^{1,p}}\|v_k\|_{W^{1,p}}$.
4. $\|\mathcal{C}_{kl}''(\phi_k, \phi_l)(u_k, u_l)(v_k, v_l)\|_{L^p} \leq \frac{C}{\rho^5}(\|u_k\|_{W^{1,p}} + \|u_l\|_{W^{1,p}})(\|v_k\|_{W^{1,p}} + \|v_l\|_{W^{1,p}})$.
5. $|\lambda''(\phi)(u, v)| \leq \frac{C}{\rho^7}\|u\|_{W^{2,p}}\|v\|_{W^{2,p}}$.

Proof. Note that by taking c small, we keep $r_k^3 + \phi_k$ positive, so ∂E_{ϕ_k} is a perturbed sphere.

The mean curvature operator $\mathcal{H}_k$ is elliptic and quasilinear. Its second Fréchet derivative is calculated from (4.28):

$$\begin{aligned}&\mathcal{H}_k''(\phi_k, D\phi_k, D^2\phi_k)(u_k, v_k) \\ &= \frac{\partial^2\mathcal{H}_k}{\partial\phi_k^2}u_k v_k + \sum_{i=1}^{2}\frac{\partial^2\mathcal{H}_k}{\partial\phi_k\partial\phi_{k,i}}(u_k v_{k,i} + u_{k,i}v_k) + \sum_{i,j=1}^{2}\frac{\partial^2\mathcal{H}_k}{\partial\phi_{k,i}\partial\phi_{k,j}}(u_{k,i}v_{k,j} + u_{k,j}v_{k,i}) \\ &\quad + \sum_{l,m=1}^{2}\frac{\partial^2\mathcal{H}_k}{\partial\phi_k\partial\phi_{k,lm}}(u_k v_{k,lm} + u_{k,lm}v_k) + \sum_{i,l,m=1}^{2}\frac{\partial^2\mathcal{H}_k}{\partial\phi_{k,i}\partial\phi_{k,lm}}(u_{k,i}v_{k,lm} + u_{k,lm}v_{k,i}).\end{aligned}$$

It is important to note that because $\mathcal{H}_k$ is quasilinear, i.e., it is linear in $D^2\phi_k$, the term

$$\sum_{i,j,l,m=1}^{2} \frac{\partial^2 \mathcal{H}_k}{\partial \phi_{k,ij} \partial \phi_{k,lm}} (u_{k,ij} v_{k,lm} + u_{k,lm} v_{k,ij})$$

is 0 and hence absent in $\mathcal{H}_k''$. The Sobolev embedding $W^{1,p} \to L^\infty$ and $\|\phi_k\|_{W^{2,p}} \le c\rho^3$ for a small c implies that $|\phi_k| \le C\rho^3$ and $|D\phi_k| \le C\rho^3$. From the definition (4.12) of $\mathcal{H}_k$ we have the pointwise estimate

$$\begin{aligned} &|\mathcal{H}_k''(\phi_k, D\phi_k, D^2\phi_k)(u_k, v_k)| \\ &\le \frac{C}{\rho^7} \left(\left| \frac{D^2\phi_k}{r_k^3} \right| |u_k|\,|v_k| + \left| \frac{D^2\phi_k}{r_k^3} \right| |u_k|\,|Dv_k| + \left| \frac{D^2\phi_k}{r_k^3} \right| |Du_k|\,|v_k| \right. \\ &\quad \left. + \left| \frac{D^2\phi_k}{r_k^3} \right| |Du_k|\,|Dv_k| + |u_k|\,|D^2 v_k| + |D^2 u_k|\,|v_k| + |Du_k|\,|D^2 v_k| + |D^2 u_k|\,|Dv_k| \right), \end{aligned}$$

when θ_k is some distance away from the two poles (where $\theta_{k,2} = 0$ or π) of S^2. Near the two poles one can use a different parametrization of S^2 so that the same pointwise estimate holds. The same Sobolev embedding implies that

$$\|\mathcal{H}_k''(\phi)(u_k, v_k)\|_{L^p} \le \frac{C}{\rho^7} \|u_k\|_{W^{2,p}} \|v_k\|_{W^{2,p}}. \tag{6.1}$$

This proves part 1 of Lemma 6.1.

We now turn to part 2 of Lemma 6.1. In our estimation of $\mathcal{A}_k''$ and $\mathcal{B}_k''$ we drop the subscript k in most quantities. The second Fréchet derivative of $\mathcal{A}_k$ is calculated from (4.29):

$$\mathcal{A}_k''(\phi)(u,v) = A_1(\phi)(u,v) + A_2(\phi)(u,v) + A_3(\phi)(u,v) + A_4(\phi)(u,v) + A_5(\phi)(u,v), \tag{6.2}$$

where

$$\begin{aligned} A_1(\phi)(u,v) &= -\frac{\gamma v(\theta)\theta}{108\pi(r^3+\phi(\theta))^{2/3}} \cdot \int_{S^2} K(\theta,\omega) u(\omega)\, d\omega, \\ A_2(\phi)(u,v) &= -\frac{\gamma u(\theta)\theta}{108\pi(r^3+\phi(\theta))^{2/3}} \cdot \int_{S^2} K(\theta,\omega) v(\omega)\, d\omega, \\ A_3(\phi)(u,v) &= \frac{\gamma}{108\pi} \int_{S^2} K(\theta,\omega) \cdot \omega \frac{u(\omega)v(\omega)}{(r^3+\phi(\omega))^{2/3}}\, d\omega, \\ A_4(\phi)(u,v) &= -\frac{\gamma u(\theta) v(\theta)}{108\pi(r^3+\phi(\theta))^{4/3}} \\ &\quad \times \int_{\tilde{E}_{\phi_k}} \frac{|(r^3+\phi(\theta))^{1/3}\theta - y|^2 - 3((r^3+\phi(\theta))^{1/3} - \theta\cdot y)^2}{|(r^3+\phi(\theta))^{1/3}\theta - y|^5}\, dy, \\ A_5(\phi)(u,v) &= \frac{2\gamma u(\theta)v(\theta)}{108\pi(r^3+\phi(\theta))^{5/3}} \int_{\tilde{E}_{\phi_k}} \frac{((r^3+\phi(\theta))^{1/3}\theta - y)\cdot\theta}{|(r^3+\phi(\theta))^{1/3}\theta - y|^3} dy. \end{aligned}$$

Recall that $\tilde{E}_{\phi_k}$ in A_4 and A_5 is $E_{\phi_k} - \xi_k$. The kernel K is

$$K(\theta,\omega) = \frac{(r^3+\phi(\theta))^{1/3}\theta - (r^3+\phi(\omega))^{1/3}\omega}{|(r^3+\phi(\theta))^{1/3}\theta - (r^3+\phi(\omega))^{1/3}\omega|^3}. \tag{6.3}$$

Here we encounter a singular integral operator

$$\mathcal{K}(u)(\theta) = \int_{S^2} K(\theta,\omega)u(\omega)\,d\omega. \tag{6.4}$$

A variant of the Calderon–Zygmund estimate [32, Theorem 1] is applicable to this operator:

$$\|\mathcal{K}(u)\|_q \le \frac{C}{\rho^2}\|u\|_{L^q}$$

for any $q \in (1,\infty)$. In [32] the kernel takes the form $K(x-y)$. To meet this requirement, we can transform (6.4) to an integral on the perturbed sphere ∂E_{ϕ_k}, then $K(\theta,\omega)$ becomes $\frac{x-y}{|x-y|^3}$, where $x, y \in \partial E_{\phi_k}$.

For $\|\phi\|_{W^{2,p}} \le c\rho^3$ with a small c, we consider

$$\|\mathcal{A}_k''(\phi)(u,v)\|_{L^p} \le \sum_{i=1}^{5}\|A_i(\phi)(u,v)\|_{L^p}.$$

For sufficiently large q

$$\|A_1(\phi)(u,v)\|_{L^p} \le \frac{C}{\rho^7}\|v_k\|_{L^q}\|\mathcal{K}(u_k)\|_{L^q} \le \frac{C}{\rho^7}\|v_k\|_{L^q}\|u_k\|_{L^q} \le \frac{C}{\rho^7}\|u_k\|_{W^{1,p}}\|v_k\|_{W^{1,p}}.$$

Similarly

$$\|A_2(\phi)(u,v)\|_{L^p} \le \frac{C}{\rho^7}\|u\|_{W^{1,p}}\|v_k\|_{W^{1,p}}.$$

Regarding A_3 we have, using the Calderon–Zygmund estimate in L^p and the Sobolev embedding theory,

$$\|A_3(\phi)(u,v)\|_{L^p} \le \frac{C}{\rho^7}\|uv\|_{L^p} \le \frac{C}{\rho^7}\|u\|_{W^{1,p}}\|v\|_{W^{1,p}}.$$

For A_4, the integral

$$\int_{\tilde{E}_{\phi_k}} \frac{|(r^3+\phi(\theta))^{1/3}\theta - y|^2 - 3((r^3+\phi(\theta))^{1/3} - \theta\cdot y)^2}{|(r^3+\phi(\theta))^{1/3}\theta - y|^5}\,dy$$

is a convergent improper integral defined by its principal part. It is of order 1 and uniformly bounded with respect to θ. In the case of ϕ equal to 0, it may be explicitly computed. (See Appendix C.) Therefore

$$\|A_4(\phi)(u,v)\|_{L^p} \le \frac{C}{\rho^7}\|uv\|_{L^p} \le \frac{C}{\rho^7}\|u\|_{W^{1,p}}\|v\|_{W^{1,p}}.$$

For A_5, because of the mild singularity, we easily find that

$$\|A_5(\phi)(u,v)\|_{L^p} \le \frac{C}{\rho^7}\|u\|_{W^{1,p}}\|v\|_{W^{1,p}}.$$

Now we have

$$\|\mathcal{A}_k''(\phi)(u_k,v_k)\|_{L^p} \le \frac{C}{\rho^7}\|u_k\|_{W^{1,p}}\|v_k\|_{W^{1,p}}.$$

This proves part 2 of Lemma 6.1.

The kernel R in $\mathcal{B}_k$ is a smooth function. Calculations from (4.30) show that

$$\begin{aligned}
&\mathcal{B}_k''(\phi)(u,v)(\theta)\\
&\quad = \frac{\gamma v(\theta)}{27(r^3+\phi(\theta))^{2/3}} \int_{S^2} u(\omega) D_1 R(\xi + (r^3+\phi(\theta))^{1/3}\theta, \xi + (r^3+\phi(\omega))^{1/3}\omega)\cdot\theta\, d\omega\\
&\qquad \frac{\gamma u(\theta)}{27(r^3+\phi(\theta))^{2/3}} \int_{S^2} v(\omega) D_1 R(\xi + (r^3+\phi(\theta))^{1/3}\theta, \xi + (r^3+\phi(\omega))^{1/3}\omega)\cdot\theta\, d\omega\\
&\qquad + \frac{\gamma}{27}\int_{S^2} \frac{u(\omega)v(\omega)}{(r^3+\phi(\omega))^{2/3}} D_2 R(\xi + (r^3+\phi(\theta))^{1/3}\theta, \xi + (r^3+\phi(\omega))^{1/3}\omega)\cdot\theta\, d\omega\\
&\qquad + \frac{\gamma u(\theta)v(\theta)}{27(r^3+\phi(\theta))^{4/3}} \int_{E_{\phi_k}} D_1^2 R(\xi + (r^3+\phi(\theta))^{1/3}\theta, y)\theta\cdot\theta\, dy\\
&\qquad - \frac{2\gamma u(\theta)v(\theta)}{27(r^3+\phi(\theta))^{5/3}} \int_{E_{\phi_k}} D_1 R(\xi + (r^3+\phi(\theta))^{1/3}\theta, y)\cdot\theta\, dy,
\end{aligned}$$

where D_1 and D_2 refer to the derivatives of R with respect to its first and second arguments, respectively. $D_1^2 R$ is the second derivative matrix of R with respect to the first argument of R. Part 3 of Lemma 6.1 is now proved easily.

The function G is also smooth in $\mathcal{C}$. We restore subscripts in the rest of this section. Similar to $\mathcal{B}_k''$, we find from (4.31) that

$$\begin{aligned}
&\mathcal{C}_{kl}''(\phi_k,\phi_l)(u_k,u_l)(v_k,v_l)(\theta_k)\\
&\quad = \frac{\gamma v_k(\theta_k)}{27(r_k^3+\phi_k(\theta_k))^{2/3}} \int_{S^2} u_l(\omega_l) D_1 G(\xi_k + (r_k^3+\phi_k(\theta_k))^{1/3}\theta_k, \xi_l\\
&\qquad + (r_l^3+\phi_l(\omega_l))^{1/3}\omega_l)\cdot\theta_k\, d\omega_l\\
&\qquad + \frac{\gamma u_k(\theta_k)}{27(r_k^3+\phi_k(\theta_k))^{2/3}} \int_{S^2} v_l(\omega_l) D_1 G(\xi_k + (r_k^3+\phi_k(\theta_k))^{1/3}\theta_k, \xi_l\\
&\qquad + (r_l^3+\phi_l(\omega_l))^{1/3}\omega_l)\cdot\theta_k\, d\omega_l\\
&\qquad + \frac{\gamma}{27}\int_{S^2} \frac{u_l(\omega_l)v_l(\omega_l)}{(r_l^3+\phi_l(\omega_l))^{2/3}} D_2 G(\xi_k + (r_k^3+\phi_k(\theta_k))^{1/3}\theta_k, \xi_l\\
&\qquad + (r_l^3+\phi_l(\omega_l))^{1/3}\omega_l)\cdot\omega_l\, d\omega_l\\
&\qquad + \frac{\gamma u_k(\theta_k)v_k(\theta_k)}{27(r_k^3+\phi_k(\theta_k))^{4/3}} \int_{E_{\phi_l}} D_1^2 G(\xi_k + (r_k^3+\phi_k(\theta_k))^{1/3}\theta_k, y)\theta_k\cdot\theta_k\, dy\\
&\qquad - \frac{2\gamma u_k(\theta_k)v_k(\theta_k)}{27(r_k^3+\phi_k(\theta_k))^{5/2}} \int_{E_{\phi_l}} D_1 G(\xi_k + (r_k^3+\phi_k(\theta_k))^{1/3}\theta_k, y)\cdot\theta_k\, dy.
\end{aligned}$$

Part 4 of Lemma 6.1 then follows.

Part 5 of Lemma 6.1 follows from parts 1–4 and the fact that

$$\begin{aligned}
0 &= \sum_k \overline{S_k''(\phi)(u,v)}\\
&= \sum_k \overline{\mathcal{H}_k''(\phi_k)(u_k,v_k)} + \sum_k \overline{\mathcal{A}_k''(\phi_k)(u_k,v_k)} + \sum_k \overline{\mathcal{B}_k''(\phi_k)(u_k,v_k)}\\
&\quad + \sum_k \overline{\mathcal{C}_k''(\phi)(u)} + K\lambda''(\phi)(u,v). \qquad \square
\end{aligned}$$

7. Reduction to $4K-1$ dimensions. We view $\mathcal{S}$ as a nonlinear operator from $\mathcal{X}$ to $\mathcal{Y}$. In this section it will be proved that, for each $(\xi, r) \in U$, a $\varphi(\cdot, \xi, r)$ exists such that $\varphi(\cdot, \xi, r) \in \mathcal{X}_*$ and

(7.1)
$$\mathcal{S}_k(\varphi)(\theta_k) = A_{k,1}\cos\theta_{k,1}\sin\theta_{k,2} + A_{k,2}\sin\theta_k\sin\theta_{k,2} + A_{k,3}\cos\theta_{k,2} + A_k, \quad k = 1, 2, \dots, K$$

for some numbers $A_{k,1}, A_{k,2}, A_{k,3}, A_k$. Note that φ is sought in $\mathcal{X}_*$. Each $\phi \in \mathcal{X}_*$ satisfies

$$\int_{S^2} \phi_k(\theta_k)\, d\theta_k = 0, \quad k = 1, 2, \dots, K, \tag{7.2}$$

$$\int_{S^2} \phi_k(\theta_k)\cos\theta_{k,1}\sin\theta_{k,2}\, d\theta_k = 0, \quad k = 1, 2, \dots, K, \tag{7.3}$$

$$\int_{S^2} \phi_k(\theta_k)\sin\theta_{k,1}\sin\theta_{k,2}\, d\theta_k = 0, \quad k = 1, 2, \dots, K, \tag{7.4}$$

$$\int_{S^2} \phi_k(\theta_k)\cos\theta_{k,2}\, d\theta_k = 0, \quad k = 1, 2, \dots, K. \tag{7.5}$$

The condition (7.2) means that $\phi_k \perp H_0$, the space of spherical harmonics of degree 0, and the conditions (7.3–7.5) state that $\phi_k \perp H_1$.

Write (7.1) as

$$\Pi\mathcal{S}(\varphi) = 0, \tag{7.6}$$

where Π is the orthogonal projection operator from $\mathcal{Y}$ to $\mathcal{Y}_*$. In the next section we will find a particular (ξ, r), say (ζ, s) at which $A_{k,1} = A_{k,2} = A_{k,3} = A_k = 0$, i.e., $\mathcal{S}(\varphi(\cdot, \zeta, s)) = 0$. This means that by finding φ we reduce the original problem (1.1) to a problem of finding a (ζ, s) in a $4K-1$ dimensional set U.

Recall $\mathcal{L}$, the linearized operator of $\mathcal{S}$ at $\phi = 0$. Expand $\mathcal{S}(\phi)$ as

$$\mathcal{S}(\phi) = \mathcal{S}(0) + \mathcal{L}(\phi) + \mathcal{N}(\phi), \tag{7.7}$$

where $\mathcal{N}$ is a higher order term defined by (7.7). Turn (7.6) to a fixed point form

$$\phi = -(\Pi\mathcal{L})^{-1}(\Pi\mathcal{S}(0) + \Pi\mathcal{N}(\phi)). \tag{7.8}$$

LEMMA 7.1. *There exists $\varphi = \varphi(\theta, \xi, r)$ such that for every $(\xi, r) \in U$, $\varphi(\cdot, \xi, r) \in \mathcal{X}_*$ solves* (7.8) *and $\|\varphi\|_{W^{2,p}} \le c\rho^5$, where c is a sufficiently large constant independent of ξ, r, ρ, and γ.*

Proof. To use the contraction mapping principle, let

$$\mathcal{T}(\phi) = -(\Pi\mathcal{L})^{-1}(\Pi\mathcal{S}(0) + \Pi\mathcal{N}(\phi)) \tag{7.9}$$

be an operator defined on

$$D(\mathcal{T}) = \{\phi \in \mathcal{X}_* : \|\phi\|_{W^{2,p}} \le c\rho^5\}, \tag{7.10}$$

where the constant c is sufficiently large and will be determined shortly.

Lemma 3.1 shows that

$$\mathcal{S}_k(0)(\theta_k) - \lambda(0) = \frac{1}{3r_k} + \frac{\gamma}{3}\left[\frac{r_k^2}{3} + \frac{4\pi r_k^3}{3}R(\xi_k, \xi_k) + \sum_{l \ne k}\frac{4\pi r_l^3}{3}G(\xi_k, \xi_l)\right] + O(\rho).$$

Each $\mathcal{S}_k(0)$ is sum of a number independent of θ_k and a quantity of order $O(\rho)$. After we apply the projection operator Π the number vanishes and

$$\|\Pi\mathcal{S}(0)\|_{L^p} = O(\rho). \tag{7.11}$$

By Lemma 5.3 we find

$$\|(\Pi\mathcal{L})^{-1}\Pi\mathcal{S}(0)\|_{W^{2,p}} \le C\rho^5. \tag{7.12}$$

For $\mathcal{N}(\phi)$ we decompose it into three parts. The first is $\mathcal{N}_1$ whose kth component is

$$\mathcal{N}_{1,k}(\phi_k) = \mathcal{H}_k(\phi_k) - \mathcal{H}_k(0) - \mathcal{H}_k'(0)(\phi_k) \tag{7.13}$$

which is $\mathcal{H}_k(\phi)$ minus its linear approximation at 0. Lemma 6.1, part 1, shows that

$$\|\mathcal{N}_1(\phi)\|_{L^p} \le \frac{C}{\rho^7}\|\phi\|^2_{W^{2,p}}. \tag{7.14}$$

The second part of $\mathcal{N}$, denoted by $\mathcal{N}_2$, is $\mathcal{A}(\phi)+\mathcal{B}(\phi)+\mathcal{C}(\phi)$ minus its linear approximation, i.e.,

$$\mathcal{N}_2(\phi) = \mathcal{A}(\phi)-\mathcal{A}(0)-\mathcal{A}'(0)(\phi)+\mathcal{B}(\phi)-\mathcal{B}(0)-\mathcal{B}'(0)(\phi)+\mathcal{C}(\phi)-\mathcal{C}(0)-\mathcal{C}'(0)(\phi). \tag{7.15}$$

Lemma 6.1, parts 2, 3, and 4, implies that

$$\|\mathcal{N}_2(\phi)\|_{L^p} \le \frac{C}{\rho^7}\|\phi\|^2_{W^{1,p}}. \tag{7.16}$$

The third part of $\mathcal{N}$, which is denoted by $\mathcal{N}_3$, merely gives a constant so that

$$\sum_k \overline{\mathcal{N}_k(\phi)} = \sum_k \overline{\mathcal{N}_{1,k}(\phi)} + \sum_k \overline{\mathcal{N}_{2,k}(\phi)} + K\mathcal{N}_3(\phi) = 0.$$

It follows that

$$|\mathcal{N}_3(\phi)| \le \frac{C}{\rho^7}\|\phi\|^2_{W^{2,p}}. \tag{7.17}$$

Therefore we deduce, from (7.14), (7.16), (7.17), and with the help of Lemma 5.3, that

$$\|\mathcal{N}(\phi)\|_{L^p} \le \frac{C}{\rho^7}\|\phi\|^2_{W^{2,p}}, \tag{7.18}$$

$$\|(\Pi\mathcal{L})^{-1}\Pi\mathcal{N}(\phi)\|_{W^{2,p}} \le \frac{C}{\rho^3}\|\phi\|^2_{W^{2,p}}. \tag{7.19}$$

Using (2.10), (7.12), (7.10), and (7.19) we find

$$\|\mathcal{T}(\phi)\|_{W^{2,p}} \le C\rho^5 + Cc^2\rho^7 \le c\rho^5$$

if c is sufficiently large and ρ sufficiently small. Therefore $\mathcal{T}$ is a map from $D(\mathcal{T})$ into itself.

Next we show that $\mathcal{T}$ is a contraction. For $\mathcal{N}_1$ we note that

$$\mathcal{N}_1(\phi_1) - \mathcal{N}_1(\phi_2) = \mathcal{H}(\phi_1) - \mathcal{H}(\phi_2) - \mathcal{H}'(0)(\phi_1 - \phi_2).$$

Therefore using Lemma 6.1, part 1, we obtain

$$\begin{aligned}
&\|\mathcal{H}(\phi_1)-\mathcal{H}(\phi_2)-\mathcal{H}'(0)(\phi_1-\phi_2)\|_{L^p}\\
&\quad\le \|\mathcal{H}'(\phi_2)(\phi_1-\phi_2)-\mathcal{H}'(0)(\phi_1-\phi_2)\|_{L^p}+\frac{C}{\rho^7}\|\phi_1-\phi_2\|^2_{W^{2,p}}\\
&\quad\le \frac{C}{\rho^7}\|\phi_2\|_{W^{2,p}}\|\phi_1-\phi_2\|_{W^{2,p}}+\frac{C}{\rho^7}\|\phi_1-\phi_2\|^2_{W^{2,p}}\\
&\quad\le \frac{C}{\rho^7}(\|\phi_1\|_{W^{2,p}}+\|\phi_2\|_{W^{2,p}})\|\phi_1-\phi_2\|_{W^{2,p}}.
\end{aligned}$$

This shows that

$$\|\mathcal{N}_1(\phi_1)-\mathcal{N}_2(\phi_2)\|_{L^p}\le \frac{C}{\rho^2}\|\phi_1-\phi_2\|_{W^{2,p}}. \tag{7.20}$$

For $\mathcal{N}_2$ we note that

$$\begin{aligned}
\mathcal{N}_2(\phi_1)-\mathcal{N}_2(\phi_2)&=\mathcal{A}(\phi_1)-\mathcal{A}(\phi_2)-\mathcal{A}'(0)(\phi_1-\phi_2)+\mathcal{B}(\phi_1)-\mathcal{B}(\phi_2)\\
&\quad-\mathcal{B}'(0)(\phi_1-\phi_2)+\mathcal{C}(\phi_1)-\mathcal{C}(\phi_2)-\mathcal{C}'(0)(\phi_1-\phi_2).
\end{aligned}\tag{7.21}$$

Therefore using Lemma 6.1, part 2, we obtain

$$\begin{aligned}
&\|\mathcal{A}(\phi_1)-\mathcal{A}(\phi_2)-\mathcal{A}'(0)(\phi_1-\phi_2)\|_{L^p}\\
&\quad\le \|\mathcal{A}'(\phi_2)(\phi_1-\phi_2)-\mathcal{A}'(0)(\phi_1-\phi_2)\|_{L^p}+\frac{C}{\rho^7}\|\phi_1-\phi_2\|^2_{W^{1,p}}\\
&\quad\le \frac{C}{\rho^7}\|\phi_2\|_{W^{1,p}}\|\phi_1-\phi_2\|_{W^{1,p}}+\frac{C}{\rho^7}\|\phi_1-\phi_2\|^2_{W^{1,p}}\\
&\quad\le \frac{C}{\rho^7}(\|\phi_1\|_{W^{1,p}}+\|\phi_2\|_{W^{1,p}})\|\phi_1-\phi_2\|_{W^{1,p}}.
\end{aligned}$$

Similarly using Lemma 6.1, parts 3 and 4, we deduce

$$\begin{aligned}
\|\mathcal{B}(\phi_1)-\mathcal{B}(\phi_2)-\mathcal{B}'(0)(\phi_1-\phi_2)\|_{L^p}&\le \frac{C}{\rho^5}(\|\phi_1\|_{W^{1,p}}+\|\phi_2\|_{W^{1,p}})\|\phi_1-\phi_2\|_{W^{1,p}}\\
\|\mathcal{C}(\phi_1)-\mathcal{C}(\phi_2)-\mathcal{C}'(0)(\phi_1-\phi_2)\|_{L^p}&\le \frac{C}{\rho^5}(\|\phi_1\|_{W^{1,p}}+\|\phi_2\|_{W^{1,p}})\|\phi_1-\phi_2\|_{W^{1,p}}.
\end{aligned}$$

From (7.21) we conclude that

$$\begin{aligned}
\|\mathcal{N}_2(\phi_1)-\mathcal{N}_2(\phi_2)\|_{L^p}&\le \frac{C}{\rho^7}(\|\phi_1\|_{W^{1,p}}+\|\phi_2\|_{W^{1,p}})\|\phi_1-\phi_2\|_{W^{1,p}}\\
&\le \frac{C}{\rho^2}\|\phi_1-\phi_2\|_{W^{1,p}}.
\end{aligned}\tag{7.22}$$

We also have

$$\|\mathcal{N}_3(\phi_1)-\mathcal{N}_3(\phi_2)\|_{L^p}\le \frac{C}{\rho^2}\|\phi_1-\phi_2\|_{W^{2,p}}. \tag{7.23}$$

Hence, following (7.20), (7.22), and (7.23), we find that

$$\begin{aligned}
&\|\mathcal{T}(\phi_1)-\mathcal{T}(\phi_2)\|_{W^{2,p}}\\
&\quad=\|(\Pi\mathcal{L})^{-1}\Pi\mathcal{N}(\phi_1)-(\Pi\mathcal{L})^{-1}\Pi\mathcal{N}(\phi_2)\|_{W^{2,p}}\le C\rho^2\|\phi_1-\phi_2\|_{W^{2,p}},
\end{aligned}\tag{7.24}$$

i.e., that $\mathcal{T}$ is a contraction map if ρ is sufficiently small. A fixed point φ exists. □

Since φ satisfies $\|\phi\|_{W^{2,p}} \le c\rho^5$, by taking ρ small we see that $r_k^3 + \varphi_k$ remains positive. ∂E_{φ_k} is a perturbed sphere.

Denote $\mathcal{S}'(\varphi)$ by $\tilde{\mathcal{L}}$. We derive a lemma for $\tilde{\mathcal{L}}$ similar to Lemma 5.3.

LEMMA 7.2. *Let Π be the same projection operator from $\mathcal{X}$ to $\mathcal{X}_*$.*

1. *There exists $C > 0$ such that for all $u \in \mathcal{X}_*$*

$$\|u\|_{W^{2,p}} \le C\rho^4 \|\Pi\tilde{\mathcal{L}}(u)\|_{L^p}.$$

2. *If (2.8) holds, then*

$$\|u\|^2_{W^{1,2}} \le C\rho^4 \langle \Pi\tilde{\mathcal{L}}(u), u\rangle.$$

Proof. By Lemma 5.3, part 1, Lemma 6.1, and the fact $\|\varphi\|_{W^{2,p}} = O(\rho^5)$, we deduce that

$$\begin{aligned}
\|\Pi\tilde{\mathcal{L}}(u)\|_{L^p} &\ge \|\Pi\mathcal{L}(u)\|_{L^p} - \|\Pi(\tilde{\mathcal{L}} - \mathcal{L})(u)\|_{L^p}\\
&\ge \frac{C}{\rho^4}\|u\|_{W^{2,p}} - \frac{C}{\rho^7}\|\varphi\|_{W^{2,p}}\|u\|_{W^{2,p}}\\
&\ge \frac{C}{\rho^4}\|u\|_{W^{2,p}} - \frac{C}{\rho^2}\|u\|_{W^{2,p}} \ \ge\ \frac{C}{\rho^4}\|u\|_{W^{2,p}}
\end{aligned}$$

when ρ is small. This proves part 1 of Lemma 7.2.

Write $\tilde{\mathcal{L}} = \mathcal{H}'(\varphi) + \mathcal{A}'(\varphi) + \mathcal{B}'(\varphi) + \mathcal{C}'(\varphi) + \lambda'(\varphi)$. Then, according to (4.7),

$$\langle \mathcal{H}_k'(\varphi_k)(u_k), u_k\rangle = \int_{S^2}\left[\frac{\partial^2 L_k}{\partial \phi_k^2}u_k^2 + 2\sum_{i=1}^2 \frac{\partial^2 L_k}{\partial\phi_k\partial\phi_{k,i}}u_k u_{k,i} + \sum_{i,j=1}^2 \frac{\partial^2 L_k}{\partial\phi_{k,i}\partial\phi_{k,j}}u_{k,i}u_{k,j}\right]d\theta_k,$$

and a similar expression holds if we replace φ_k and $\varphi_{k,i}$ by 0 in the last formula.

With $\|\varphi\|_{W^{2,p}} = O(\rho^5)$, calculations show that

$$\begin{aligned}
&|\langle(\mathcal{H}_k'(\varphi_k) - \mathcal{H}_k'(0))u_k, u_k\rangle|\\
&\quad\le \left|\int_{S^2}\left(\frac{\partial^2 L_k(\varphi_k)}{\partial\phi_k^2} - \frac{\partial^2 L_k(0)}{\partial\phi_k^2}\right)u_k^2\,d\theta_k\right|\\
&\qquad + 2\sum_{i=1}^2\left|\int_{S^2}\left(\frac{\partial^2 L_k(\varphi_k)}{\partial\phi_k\partial\phi_{k,i}} - \frac{\partial^2 L_k(0)}{\partial\phi_k\partial\phi_{k,i}}\right)u_k u_{k,i}\,d\theta_k\right|\\
&\qquad + \sum_{i,j=1}^2\left|\int_{S^2}\left(\frac{\partial^2 L_k(\varphi_k)}{\partial\phi_{k,i}\partial\phi_{k,j}} - \frac{\partial^2 L_k(0)}{\partial\phi_{k,i}\partial\phi_{k,j}}\right)u_{k,i}u_{k,j}\,d\theta_k\right|\\
&\quad\le \frac{C}{\rho^2}\|u\|^2_{L^2} + \frac{C}{\rho^2}\|u\|_{L^2}\|Du\|_{L^2} + \frac{C}{\rho^2}\|Du\|^2_{L^2} \ \le\ \frac{C}{\rho^2}\|u\|^2_{W^{1,2}}.
\end{aligned} \tag{7.25}$$

Next we estimate $\|(\mathcal{A}_k'(\varphi_k) - \mathcal{A}_k'(0))u_k\|_{L^2}$ and revisit $\mathcal{A}_k''$. Arguing as in the proof of Lemma 6.1, part 2, we deduce that

$$\|\mathcal{A}_k''(\phi)(u_k, v_k)\|_{L^2} \le \frac{C}{\rho^7}\|u_k\|_{W^{1,2}}\|v_k\|_{W^{1,2}}.$$

This implies that in this lemma

$$\|(\mathcal{A}_k'(\varphi) - \mathcal{A}_k'(0))u_k\|_{L^2} \le \frac{C}{\rho^7}C\rho^5\|u_k\|_{W^{1,2}} \le \frac{C}{\rho^2}\|u_k\|_{W^{1,2}}.$$

Simpler arguments show that

$$\|(\mathcal{B}_k'(\varphi)-\mathcal{B}_k'(0))u_k\|_{L^2}\le \frac{C}{\rho^2}\|u_k\|_{W^{1,2}},\quad \|(\mathcal{C}'(\varphi)-\mathcal{C}'(0))u\|_{L^2}\le \frac{C}{\rho^2}\|u\|_{W^{1,2}}.$$

We obtain that

$$\|(\mathcal{A}'(\varphi)+\mathcal{B}'(\varphi)+\mathcal{C}'(\varphi)-\mathcal{A}'(0)-\mathcal{B}'(0)-\mathcal{C}'(0))u\|_{L^2}\le \frac{C}{\rho^2}\|u\|_{W^{1,2}}. \tag{7.26}$$

If (2.8) holds, we combine Lemma 5.3, part 3, (7.25), and (7.26) to deduce that

$$\langle \Pi\tilde{\mathcal{L}}(u),u\rangle=\langle \Pi\mathcal{L}(u),u\rangle+\langle \Pi(\tilde{\mathcal{L}}-\mathcal{L})u,u\rangle\ge \frac{C}{\rho^4}\|u\|^2_{W^{1,2}}-\frac{C}{\rho^2}\|u\|^2_{W^{1,2}}\ge \frac{C}{\rho^4}\|u\|^2_{W^{1,2}},$$

proving the second part. ∎

One consequence of Lemma 7.2, part 1, is an estimate of $\frac{\partial\varphi}{\partial\xi_{l,j}}$.

LEMMA 7.3. *The fixed point φ satisfies $\left\|\frac{\partial\varphi}{\partial\xi_{l,j}}\right\|_{W^{2,p}}=O(\rho^4)$, $l=1,2,\dots,K$, $j=1,2,3$.*

Proof. We prove this lemma by the implicit function theorem. Fix $l\in\{1,2,\dots,K\}$ and $j\in\{1,2,3\}$. Differentiating $\Pi\mathcal{S}(\varphi)$ with respect to $\xi_{l,j}$ we find that, for $k=1,2,\dots,K$, if $k=l$, then

$$\begin{aligned}\frac{\partial\Pi\mathcal{S}_l(\varphi)}{\partial\xi_{l,j}}&=\Pi\tilde{\mathcal{L}}_l\left(\frac{\partial\varphi}{\partial\xi_{l,j}}\right)+\Pi\frac{\gamma}{3}\\&\quad\times\int_{E_{\varphi_l}}\left[\frac{\partial R(\xi_l+(r_l^3+\varphi_l(\theta_l))^{1/3}\theta_l,y)}{\partial x_j}+\frac{\partial R(\xi_l+(r_l^3+\varphi_l(\theta_l))^{1/3}\theta_l,y)}{\partial y_j}\right]dy\\&\quad+\sum_{m\ne l}\Pi\frac{\gamma}{3}\int_{E_{\varphi_m}}\frac{\partial G(\xi_l+(r_l^3+\varphi_l(\theta_l))^{1/3}\theta_l,y)}{\partial x_j}\,dy,\end{aligned}$$

and if $k\ne l$,

$$\frac{\partial\Pi\mathcal{S}_k(\varphi)}{\partial\xi_{l,j}}=\Pi\tilde{\mathcal{L}}_k\left(\frac{\partial\varphi}{\partial\xi_{l,j}}\right)+\Pi\frac{\gamma}{3}\int_{E_{\varphi_l}}\frac{\partial G(\xi_k+(r_k^3+\varphi_k(\theta_k))^{1/3}\theta_k,y)}{\partial y_j}\,dy.$$

Here $R=R(x,y)$ and $G=G(x,y)$. It is clear that

$$\begin{aligned}\Bigg\|\frac{\gamma}{3}\int_{E_{\varphi_l}}\Bigg[&\frac{\partial R(\xi_l+(r_l^3+\varphi_l(\theta_l))^{1/3}\theta_l,y)}{\partial x_j}\\&+\frac{\partial R(\xi_l+(r_l^3+\varphi_l(\theta_l))^{1/3}\theta_l,y)}{\partial y_j}\Bigg]dy\Bigg\|_{L^p}=O(\gamma\rho^3),\end{aligned}$$

$$\left\|\frac{\gamma}{3}\int_{E_{\varphi_m}}\frac{\partial G(\xi_l+(r_l^3+\varphi_l(\theta_l))^{1/3}\theta_l,y)}{\partial x_j}\,dy\right\|_{L^p}=O(\gamma\rho^3),$$

$$\left\|\frac{\gamma}{3}\int_{E_{\varphi_l}}\frac{\partial G(\xi_k+(r_k^3+\varphi_k(\theta_k))^{1/3}\theta_k,y)}{\partial y_j}\,dy\right\|_{L^p}=O(\gamma\rho^3).$$

Therefore

$$\frac{\partial \Pi \mathcal{S}(\varphi)}{\partial \xi_{l,j}} = \Pi \tilde{\mathcal{L}} \left(\frac{\partial \varphi}{\partial \xi_{l,j}} \right) + W,$$

where $\|W\|_{L^p} = O(\gamma \rho^3) = O(1)$.

On the other hand

$$\frac{\partial \Pi \mathcal{S}(\varphi)}{\partial \xi_{l,j}} = 0,$$

since $\Pi \mathcal{S}(\varphi) = 0$.

By Lemma 7.2 we deduce that

$$\left\| \frac{\partial \varphi}{\partial \xi_{l,j}} \right\|_{W^{2,p}} \leq C \rho^4 O(1) \leq C \rho^4. \qquad \square$$

8. Solving the reduced problem. We now turn to solve $\mathcal{S}(\phi) = 0$.

LEMMA 8.1. *$J(E_\varphi) = J(B) + O(\rho^6)$. More explicitly*

$$J(E_\varphi) = \sum_{k=1}^{K} 4\pi r_k^2 + \frac{\gamma}{2} \Bigg\{ \sum_{k=1}^{K} \left[\frac{8\pi r_k^5}{15} + \left(\frac{4\pi}{3} \right)^2 r_k^6 R(\xi_k, \xi_k) \right] + \sum_{k=1}^{K} \sum_{l=1, l \neq k}^{K} \left(\frac{4\pi}{3} \right)^2 r_k^3 r_l^3 G(\xi_k, \xi_l) \Bigg\} + O(\rho^5).$$

Here $J(E_\varphi) = J(E_{\varphi(\cdot,\xi,r)})$ can be considered as a function of (ξ, r).

Proof. Expanding $J(E_\varphi)$ yields

$$J(E_\varphi) = J(B) + \sum_k \int_{S^2} \mathcal{S}_k(0) \varphi_k \, d\theta_k + \frac{1}{2} \sum_k \int_{S^2} \mathcal{L}_k(\varphi) \varphi_k \, d\theta_k + O(\rho^8). \tag{8.1}$$

The error term $O(\rho^8)$ in (8.1) is obtained in the same way that (7.18) is derived.

On the other hand $\Pi \mathcal{S}(\varphi) = 0$ implies that

$$\Pi(\mathcal{S}_k(0) + \mathcal{L}_k(\varphi) + \mathcal{N}_k(\varphi)) = 0,$$

where $\mathcal{N}$ is given in (7.7) and estimated in (7.18). We multiply the last equation by φ_k and integrate to derive

$$\int_{S^2} \mathcal{S}_k(0) \varphi_k \, d\theta_k + \int_{S^2} \mathcal{L}(\varphi_k) \varphi_k \, d\theta_k = O(\rho^8).$$

We can now rewrite (8.1) as

$$J(E_\varphi) = J(B) + \frac{1}{2} \sum_k \int_{S^2} \mathcal{S}_k(0) \varphi_k \, d\theta_k + O(\rho^8).$$

Note that $\mathcal{S}_k(0)$ is the sum of a number independent of θ_k and a quantity of order ρ by Lemma 3.1. Since φ_k satisfies (7.2), the inner product of the number and φ_k is zero, and hence

$$\int_{S^2} \mathcal{S}_k(0) \varphi_k \, d\theta = O(\rho^6).$$

Therefore

$$J(E_\varphi) = J(B) + O(\rho^6).$$

Lemma 3.2 implies that

$$\begin{aligned} J(E_\varphi) &= \sum_{k=1}^K 4\pi r_k^2 + \frac{\gamma}{2}\Bigg\{ \sum_{k=1}^K \left[\frac{8\pi r_k^5}{15} + \left(\frac{4\pi}{3}\right)^2 r_k^6 R(\xi_k,\xi_k)\right] \\ &\quad + \sum_{k=1}^K \sum_{l=1,l\neq k}^K \left(\frac{4\pi}{3}\right)^2 r_k^3 r_l^3 G(\xi_k,\xi_l) \\ &\quad + \sum_{k=1}^K \sum_{l=1}^K \left(\frac{4\pi}{3}\right)^2 \left(\frac{r_k^3 r_l^5}{10|D|} + \frac{r_k^5 r_l^3}{10|D|}\right)\Bigg\} + O(\rho^6) \\ &= \sum_{k=1}^K 4\pi r_k^2 + \frac{\gamma}{2}\Bigg\{ \sum_{k=1}^K \left[\frac{8\pi r_k^5}{15} + \left(\frac{4\pi}{3}\right)^2 r_k^6 R(\xi_k,\xi_k)\right] \\ &\quad + \sum_{k=1}^K \sum_{l=1,l\neq k}^K \left(\frac{4\pi}{3}\right)^2 r_k^3 r_l^3 G(\xi_k,\xi_l)\Bigg\} + O(\rho^5). \end{aligned}$$

This proves the lemma. ∎

LEMMA 8.2. *When ρ is sufficiently small, $J(E_{\varphi(\cdot,\xi,r)})$ is minimized at some $(\xi, r) = (\zeta, s) \in U$. As $\rho \to 0$, $\frac{s}{\rho} \to (1, 1, \ldots, 1)$, and $\zeta \to \zeta_0$ along a subsequence where $\zeta_0 \in U_1$ is a global minimum of F.*

Proof. Let us rescale the problem with

$$R = \frac{r}{\rho}, \quad \tilde{J}(\xi, R) = \frac{2}{\gamma\rho^5} J(E_{\varphi(\cdot,\xi,r)}), \quad (\xi, R) \in U_1 \times \tilde{U}_2,$$

where

$$\tilde{U}_2 = \left\{(R_1, R_2, \ldots, R_K) : \ 1 - \delta_2 < R_k < 1 + \delta_2, \ \sum_{k=1}^K R_k^3 = K\right\}$$

is a scaled version of U_2. Note that by (2.5) and Lemma 8.1 that

$$\begin{aligned} \tilde{J}(\xi, R) &= \frac{8\pi}{\gamma\rho^3} \sum_{k=1}^K R_k^2 + \sum_{k=1}^K \frac{8\pi R_k^5}{15} \\ &\quad + \rho\left(\frac{4\pi}{3}\right)^2 \left[\sum_{k=1}^K (R_k^6 R(\xi_k,\xi_k)) + \sum_{k=1}^K \sum_{l\neq k} R_k^3 R_l^3 G(\xi_k,\xi_l)\right] + O(\rho^3). \end{aligned}$$

Again by (2.5) we may assume that along a subsequence

$$\frac{8\pi}{\gamma\rho^3} \to b_0 \le \frac{8\pi}{(3+\epsilon)\pi}, \text{ as } \rho \to 0. \tag{8.2}$$

Let (ζ, S) be the global minimum of $\tilde{J}$ on the closure of $U_1 \times \tilde{U}_2$. Here $S = \frac{s}{\rho}$. Let $(\zeta, S) \to (\zeta_0, S_0)$ along a subsequence as ρ tends to 0. First we claim that

$S_0 = (1, 1, \ldots, 1)$. Suppose this is false, i.e., $S_0 \neq (1, 1, \ldots, 1)$. Then as ρ tends to 0,

$$\begin{aligned}\tilde{J}(\zeta, (1, \ldots, 1)) - \tilde{J}(\zeta, S) &= \sum_k \frac{8\pi}{\gamma\rho^3} + \sum_k \frac{8\pi}{15} - \sum_k \frac{8\pi S_k^2}{\gamma\rho^3} - \sum_k \frac{8\pi S_k^5}{15} + O(\rho) \\ &\to \sum_k b_0 + \sum_k \frac{8\pi}{15} - \sum_k b_0 S_{0,k}^2 - \sum_k \frac{8\pi S_{0,k}^5}{15}.\end{aligned}$$

Because of (8.2) and the constraint $\sum_k S_{0,k}^3 = K$, it is easy to show that the last line is negative if δ_2 in (3.1) is small enough, depending on ϵ. For, under (8.2), the function

$$x \to b_0 x^{2/3} + \frac{8\pi}{15} x^{5/3}$$

is convex when x is near 1. The last assertion then follows from the Jensen's inequality, when x takes values $S_{0,k}^3$. This is a contradiction to that (ζ, S) is a minimum of $\tilde{J}$.

Next we claim that ζ_0 minimizes F in U_1. Suppose this is false. Let η be a minimum of F in U_1. Then $F(\eta) < F(\zeta_0)$. Consider

$$\begin{aligned}\frac{1}{\rho}\left(\frac{3}{4\pi}\right)^2 (\tilde{J}(\eta, S) - \tilde{J}(\zeta, S)) &= \sum_{k=1}^K S_k^6 R(\eta_k, \eta_k) + \sum_{k=1}^K \sum_{l \neq k} S_k^3 S_l^3 G(\eta_k, \eta_l) \\ &\quad - \sum_{k=1}^K S_k^6 R(\zeta_k, \zeta_k) - \sum_{k=1}^K \sum_{l \neq k} S_k^3 S_l^3 G(\zeta_k, \zeta_l) + O(\rho^2) \\ &\to F(\eta) - F(\zeta_0) \ < \ 0, \text{ as } \rho \to 0,\end{aligned}$$

another contradiction to that (ζ, S) minimizes $\tilde{J}$. Note that $(\zeta, S) \in U_1 \times \tilde{U}_2$ when ρ is small, since $(\zeta_0, S_0) \in U_1 \times \tilde{U}_2$. $\square$

We show that $\varphi(\cdot, \zeta, s)$ is an exact solution of (1.1) in the next two lemmas. The first shows that $A_k = 0$ in (7.1) at $\xi = \zeta$ and $r = s$.

Lemma 8.3. *At $\xi = \zeta$ and $r = s$, $\mathcal{S}_k(\varphi(\cdot, \zeta, s))(\theta_k) = A_{k,1} \cos\theta_{k,1} \sin\theta_{k,2} + A_{k,2} \sin\theta_{k,1} \sin\theta_{k,2} + A_{k,3} \cos\theta_{k,2}$.*

Proof. At each $(\xi, r) \in U$, let

$$p_k = r_k^3, \quad q_k = s_k^3. \tag{8.3}$$

Calculations show that

$$\begin{aligned}\frac{\partial J(E_\varphi)}{\partial p_k} &= \sum_{l=1}^K \int_{S^2} [\mathcal{S}_l(\varphi) - \lambda(\varphi)] \frac{\partial(p_l + \varphi_l)}{\partial p_k}\, d\theta_l \\ &= \int_{S^2} [\mathcal{S}_k(\varphi) - \lambda(\varphi)] \left(1 + \frac{\partial \varphi_k}{\partial p_k}\right) d\theta_k + \sum_{l \neq k} \int_{S^2} [\mathcal{S}_l(\varphi) - \lambda(\varphi)] \frac{\partial \varphi_l}{\partial p_k}\, d\theta_l \\ &= \int_{S^2} (A_{k,1} \cos\theta_{k,1} \sin\theta_{k,2} + A_{k,2} \sin\theta_{k,1} \sin\theta_{k,2} + A_{k,3} \cos\theta_{k,2} \\ &\qquad + A_k - \lambda(\varphi)) \left(1 + \frac{\partial \varphi_k}{\partial p_k}\right) d\theta_k\end{aligned}$$

$$+\sum_{l\neq k}\int_{S^2}(A_{l,1}\cos\theta_{l,1}\sin\theta_{l,2}+A_{l,2}\sin\theta_{l,1}\sin\theta_{l,2}+A_{l,3}\cos\theta_{l,2}$$
$$+A_l-\lambda(\varphi))\frac{\partial\varphi_l}{\partial p_k}\,d\theta_l$$
$$=4\pi A_k-4\pi\lambda(\varphi).$$

Here we have used the facts that

$$\frac{\partial\varphi_l}{\partial p_k}\perp\cos\theta_{l,1}\sin\theta_{l,2},\ \sin\theta_{l,1}\sin\theta_{l,2},\ \cos\theta_{l,2},\ 1$$

which follow from $\varphi\in\mathcal{X}_*$.

On the other hand at the minimum $p=q$ and $\xi=\zeta$ with respect to p, we must have

$$\frac{\partial J(E_\varphi)}{\partial p_k}|_{\xi=\zeta,p=q}=\mu$$

for all $k=1,2,\dots,K$. Here μ is a Lagrange multiplier coming from the constraint

$$\sum_{k=1}^K p_k=\frac{3a|D|}{4\pi}.$$

Therefore we deduce that

$$A_k=\frac{\mu}{4\pi}+\lambda$$

which is independent of k. By (4.20) we derive that $\sum_{k=1}^K A_k=0$, and then we conclude that each A_k must be 0. $\square$

Next we show that $A_{k,1}$, $A_{k,2}$, and $A_{k,3}$ in (7.1) are 0 at $\xi=\zeta$ and $r=s$. The proof uses a tricky reparametrization technique.

LEMMA 8.4. *At $\xi=\zeta$ and $r=s$, $\mathcal{S}(\varphi(\cdot,\zeta,s))=0$.*

Proof. To simplify notations in this proof, we do not explicitly indicate the dependence of φ on r, i.e., we write $\varphi(\cdot,\xi)$ instead of $\varphi(\cdot,\xi,r)$. For each $\xi_k=(\xi_{k,1},\xi_{k,2},\xi_{k,3})$ near ζ_k we reparametrize $\partial_D E_{\varphi_k(\cdot,\xi)}$. Let ζ_k be the center of new polar coordinates, $r_k^3+\psi_k$ the new radius cube, and η_k the new angle. A point on $\partial_D E_{\varphi_k(\cdot,\xi)}$ is described as $\zeta_k+(r_k^3+\psi_k)^{1/3}\eta_k$. It is related to the old polar coordinates via

$$\zeta_k+(r_k^3+\psi_k)^{1/3}\eta_k=\xi_k+(r_k^3+\varphi_k)^{1/3}\theta_k. \tag{8.4}$$

In the new coordinates E_{φ_k} becomes E_{ψ_k}. It is viewed as a perturbation of the ball centered at ζ_k with radius r_k. The perturbation is described by ψ_k which is a function of η_k and ξ.

The main effect of the new coordinates is to "freeze" the center. The center of the new polar system is ζ_k which is fixed while the center of the old polar system is ξ_k which varies in D.

We now consider the derivative of $J(E_{\varphi(\cdot,\xi)})=J(E_{\psi(\cdot,\xi)})$ with respect to ξ_k. On one hand, at $\xi=\zeta$ and $r=s$,

$$\frac{\partial J(E_{\psi(\cdot,\xi)})}{\partial\xi_{k,j}}|_{\xi=\zeta}=\frac{\partial J(E_{\varphi(\cdot,\xi)})}{\partial\xi_{k,j}}|_{\xi=\zeta}=0,\quad j=1,2,3, \tag{8.5}$$

since ζ is a minimum.

On the other hand calculations show that

$$\frac{\partial J(E_{\psi(\cdot,\xi)})}{\partial \xi_{k,j}} = \sum_{l=1}^{K} \int_{S^2} \mathcal{S}_l(\psi(\cdot,\xi))(\eta_l) \frac{\partial \psi_l}{\partial \xi_{k,j}}\, d\eta_l. \tag{8.6}$$

We emphasize that (8.6) is obtained under the reparametrized coordinates, in which the dependence of $J(E_{\psi(\cdot,\xi)})$ on ξ is reflected only in the dependence of ψ on ξ. Had we calculated in the original coordinates, ξ would have appeared also in the nonlocal part of J through $R(\xi_l + \cdots, \xi_l + \cdots)$ and $G(\xi_k + \cdots, \xi_l + \cdots)$. The result would have been very different from (8.6). See the proof of Lemma 7.3 which involves differentiation with respect to ξ in the original coordinates. In the derivation of (8.6) we have used the fact that $\sum_l \int_{S^2} \psi_l \, d\eta_l = 0$ which implies that $\sum_l \int_{S^2} \frac{\partial \psi_l}{\partial \xi_{k,j}}\, d\eta_l = 0$, so that $\sum_l \int_{S^2} \lambda(\psi) \frac{\partial \psi_l}{\partial \xi_{k,j}}\, d\eta_l = 0$, where $\lambda(\psi)$ is part of

$$\mathcal{S}_l(\psi) = \mathcal{H}_l(\psi) + \mathcal{A}_l(\psi) + \mathcal{B}_l(\psi) + \mathcal{C}_l(\psi) + \lambda(\psi),$$

and we can reach the right-hand side of (8.6).

The expression $\mathcal{S}(\phi)$ is invariant under reparametrization, i.e.,

$$\mathcal{S}_l(\varphi(\cdot,\xi))(\theta_l) = \mathcal{S}_l(\psi(\cdot,\xi))(\eta_l). \tag{8.7}$$

Now we return to the original coordinate system and integrate with respect to θ_l in (8.6). Then

$$\frac{\partial J(E_{\psi(\cdot,\xi)})}{\partial \xi_{k,j}} = \sum_{l=1}^{K} \int_{S^2} \mathcal{S}_l(\varphi(\cdot,\xi))(\theta_l) \frac{\partial \psi_l(\eta_l(\theta_l,\xi),\xi)}{\partial \xi_{k,j}} \left| \frac{\partial(\eta_{l,1},\eta_{l,2})}{\partial(\theta_{l,1},\theta_{l,2})} \right| \frac{\sin \eta_{l,2}}{\sin \theta_{l,2}}\, d\theta_l. \tag{8.8}$$

There are two cases: $l = k$ and $l \neq k$. We start with the first case. Recall that ψ_k and η_k are defined implicitly as functions of θ_k and ξ by (8.4). Let us agree that $\psi_k = \psi_k(\eta_k, \xi)$ is a function of η_k and ξ. Set $\Psi_k(\theta_k,\xi) = \psi_k(\eta_k(\theta_k,\xi),\xi)$. To simplify notations let us set

$$g = (r_k^3 + \Psi_k)^{1/3}, \quad \tilde{g} = (r_k^3 + \varphi_k)^{1/3}. \tag{8.9}$$

Implicit differentiation shows that, with the help of Lemmas 7.1 and 7.3,

$$\begin{bmatrix} \frac{\partial \eta_{k,1}}{\partial \theta_{k,1}} & \frac{\partial \eta_{k,1}}{\partial \theta_{k,2}} & \frac{\partial \eta_{k,1}}{\partial \xi_{k,1}} & \frac{\partial \eta_{k,1}}{\partial \xi_{k,2}} & \frac{\partial \eta_{k,1}}{\partial \xi_{k,3}} \\ \frac{\partial \eta_{k,2}}{\partial \theta_{k,1}} & \frac{\partial \eta_{k,2}}{\partial \theta_{k,2}} & \frac{\partial \eta_{k,2}}{\partial \xi_{k,1}} & \frac{\partial \eta_{k,2}}{\partial \xi_{k,2}} & \frac{\partial \eta_{k,2}}{\partial \xi_{k,3}} \\ \frac{\partial \Psi_k}{\partial \theta_{k,1}} & \frac{\partial \Psi_k}{\partial \theta_{k,2}} & \frac{\partial \Psi_k}{\partial \xi_{k,1}} & \frac{\partial \Psi_k}{\partial \xi_{k,2}} & \frac{\partial \Psi_k}{\partial \xi_{k,3}} \end{bmatrix} = -M^{-1}N, \tag{8.10}$$

where

$$M^{-1} = \begin{bmatrix} g\sin\eta_{k,1}\sin\eta_{k,2} & -g\cos\eta_{k,1}\cos\eta_{k,2} & -\frac{\cos\eta_{k,1}\sin\eta_{k,2}}{3g^2} \\ -g\cos\eta_{k,1}\sin\eta_{k,2} & -g\sin\eta_{k,1}\cos\eta_{k,2} & -\frac{\sin\eta_{k,1}\sin\eta_{k,2}}{3g^2} \\ 0 & g\sin\eta_{k,2} & -\frac{\cos\eta_{k,2}}{3g^2} \end{bmatrix}^{-1}$$

$$= \frac{1}{\sin\eta_{k,2}}\begin{bmatrix} \frac{\sin\eta_{k,1}}{g} & -\frac{\cos\eta_{k,1}}{g} & 0 \\ -\frac{\cos\eta_{k,1}\cos\eta_{k,2}\sin\eta_{k,2}}{g} & -\frac{\sin\eta_{k,1}\cos\eta_{k,2}\sin\eta_{k,2}}{g} & \frac{\sin^2\eta_{k,2}}{g} \\ -3g^2\cos\eta_{k,1}\sin^2\eta_{k,2} & -3g^2\sin\eta_{k,1}\sin^2\eta_{k,2} & -3g^2\cos\eta_{k,2}\sin\eta_{k,2} \end{bmatrix}$$

and $N = [N_{ij}]$ is a 3 by 5 matrix given by

$$N_{11} = \frac{\cos\theta_{k,1}\sin\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\theta_{k,1}} - \tilde{g}\sin\theta_{k,1}\sin\theta_{k,2},$$
$$N_{12} = \frac{\cos\theta_{k,1}\sin\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\theta_{k,2}} + \tilde{g}\cos\theta_{k,1}\cos\theta_{k,2}, \quad N_{13} = 1 + \frac{\cos\theta_{k,1}\sin\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\xi_{k,1}},$$
$$N_{14} = \frac{\cos\theta_{k,1}\sin\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\xi_{k,2}}, \quad N_{15} = \frac{\cos\theta_{k,1}\sin\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\xi_{k,3}}$$
$$N_{21} = \frac{\sin\theta_{k,1}\sin\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\theta_{k,1}} + \tilde{g}\cos\theta_{k,1}\sin\theta_{k,2},$$
$$N_{22} = \frac{\sin\theta_{k,1}\sin\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\theta_{k,2}} + \tilde{g}\sin\theta_{k,1}\cos\theta_{k,2},$$
$$N_{23} = \frac{\sin\theta_{k,1}\sin\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\xi_{k,1}}, \quad N_{24} = 1 + \frac{\sin\theta_{k,1}\sin\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\xi_{k,2}},$$
$$N_{25} = \frac{\sin\theta_{k,1}\sin\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\xi_{k,3}}, \quad N_{31} = \frac{\cos\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\theta_{k,1}},$$
$$N_{32} = \frac{\cos\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\theta_{k,2}} - \tilde{g}\sin\theta_{k,2}, \quad N_{33} = \frac{\cos\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\xi_{k,1}},$$
$$N_{34} = \frac{\cos\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\xi_{k,2}}, \quad N_{35} = 1 + \frac{\cos\theta_{k,2}}{3\tilde{g}^2}\frac{\partial\varphi_k}{\partial\xi_{k,3}}.$$

We write N as

$$\sin\theta_{k,2}\begin{bmatrix} \begin{matrix}-\tilde{g}\sin\theta_{k,1}\\ +O(\rho^3)\end{matrix} & \begin{matrix}\frac{\tilde{g}\cos\theta_{k,1}\cos\theta_{k,2}}{\sin\theta_{k,2}}\\ +O(\rho^3)\end{matrix} & \begin{matrix}\frac{1}{\sin\theta_{k,2}}\\ +O(\rho^2)\end{matrix} & O(\rho^2) & O(\rho^2) \\ \begin{matrix}\tilde{g}\cos\theta_{k,1}\\ +O(\rho^3)\end{matrix} & \begin{matrix}\frac{\tilde{g}\sin\theta_{k,1}\cos\theta_{k,2}}{\sin\theta_{k,2}}\\ +O(\rho^3)\end{matrix} & O(\rho^2) & \begin{matrix}\frac{1}{\sin\theta_{k,2}}\\ +O(\rho^2)\end{matrix} & O(\rho^2) \\ \frac{O(\rho^3)}{\sin\theta_{k,2}} & -\tilde{g} + \frac{O(\rho^3)}{\sin\theta_{k,2}} & \frac{O(\rho^2)}{\sin\theta_{k,2}} & \frac{O(\rho^2)}{\sin\theta_{k,2}} & \frac{1}{\sin\theta_{k,2}} + \frac{O(\rho^2)}{\sin\theta_{k,2}} \end{bmatrix}.$$

At $\xi = \zeta$, we have $\eta = \theta$ and $\Psi = \varphi$. Multiplying M^{-1} and N we deduce that (8.10) becomes

(8.11)

$$\begin{bmatrix} 1+O(\rho^2) & O(\rho^2) & -\frac{\sin\theta_{k,1}}{\sin\theta_{k,2}g}+O(\rho) & \frac{\cos\theta_{k,1}}{\sin\theta_{k,2}g}+O(\rho) & O(\rho) \\ O(\rho^2) & 1+O(\rho^2) & \frac{\cos\theta_{k,1}\cos\theta_{k,2}}{g}+O(\rho) & \frac{\sin\theta_{k,1}\cos\theta_{k,2}}{g}+O(\rho) & -\frac{\sin\theta_{k,2}}{g}+O(\rho) \\ O(\rho^5) & O(\rho^5) & 3g^2\cos\theta_{k,1}\sin\theta_{k,2}+O(\rho^4) & 3g^2\sin\theta_{k,1}\sin\theta_{k,2}+O(\rho^4) & 3g^2\cos\theta_{k,2}+O(\rho^4) \end{bmatrix}$$

when $\xi = \zeta$.

We have found from (8.11) that at $\xi = \zeta$,

$$\left(\frac{\partial\Psi_k}{\partial\xi_{k,1}}, \frac{\partial\Psi_k}{\partial\xi_{k,2}}, \frac{\partial\Psi_k}{\partial\xi_{k,3}}\right)|_{\xi=\zeta} = 3r_k^2\theta_k + O(\rho^4). \tag{8.12}$$

To compute $\frac{\partial\psi_k}{\partial\xi_{k,j}}$, we invert $\eta_k = \eta_k(\xi, \theta_k)$ to express $\theta_k = \Theta_k(\eta_k, \xi)$. Then

$$\frac{\partial\psi_k}{\partial\xi_{k,j}} = \frac{\partial\Psi_k}{\partial\xi_{k,j}} + \frac{\partial\Psi_k}{\partial\theta_{k,1}}\frac{\partial\Theta_{k,1}}{\partial\xi_{k,j}} + \frac{\partial\Psi_k}{\partial\theta_{k,2}}\frac{\partial\Theta_{k,2}}{\partial\xi_{k,j}}.$$

At $\xi = \zeta$, since, by (8.11),

$$\frac{\partial\Psi_k}{\partial\theta_{k,m}}|_{\xi=\zeta} = O(\rho^5) \tag{8.13}$$

and

$$\begin{aligned} \begin{bmatrix} \frac{\partial\Theta_{k,1}}{\partial\xi_{k,1}} & \frac{\partial\Theta_{k,1}}{\partial\xi_{k,2}} & \frac{\partial\Theta_{k,1}}{\partial\xi_{k,3}} \\ \frac{\partial\Theta_{k,2}}{\partial\xi_{k,1}} & \frac{\partial\Theta_{k,2}}{\partial\xi_{k,2}} & \frac{\partial\Theta_{k,2}}{\partial\xi_{k,3}} \end{bmatrix}_{\xi=\zeta} &= -\begin{bmatrix} \frac{\partial\eta_{k,1}}{\partial\theta_{k,1}} & \frac{\partial\eta_{k,1}}{\partial\theta_{k,2}} \\ \frac{\partial\eta_{k,2}}{\partial\theta_{k,1}} & \frac{\partial\eta_{k,2}}{\partial\theta_{k,2}} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial\eta_{k,1}}{\partial\xi_{k,1}} & \frac{\partial\eta_{k,1}}{\partial\xi_{k,2}} & \frac{\partial\eta_{k,1}}{\partial\xi_{k,3}} \\ \frac{\partial\eta_{k,2}}{\partial\xi_{k,1}} & \frac{\partial\eta_{k,2}}{\partial\xi_{k,2}} & \frac{\partial\eta_{k,2}}{\partial\xi_{k,3}} \end{bmatrix} \\ &= \frac{O\left(\frac{1}{\rho}\right)}{\sin\theta_{k,2}}, \end{aligned} \tag{8.14}$$

we deduce that

$$\left(\frac{\partial\psi_k}{\partial\xi_{k,1}}, \frac{\partial\psi_k}{\partial\xi_{k,2}}, \frac{\partial\psi_k}{\partial\xi_{k,3}}\right)|_{\xi=\zeta} = 3r_k^2\theta_k + \frac{O(\rho^4)}{\sin\theta_{k,2}}(1,1,1). \tag{8.15}$$

The second case $l \neq k$ is similar, for which we omit the details of our computation. At $\xi = \zeta$, we have

$$\left(\frac{\partial\psi_l}{\partial\xi_{k,1}}, \frac{\partial\psi_l}{\partial\xi_{k,2}}, \frac{\partial\psi_l}{\partial\xi_{k,3}}\right)|_{\xi=\zeta} = \frac{O(\rho^4)}{\sin\theta_{l,2}}(1,1,1). \tag{8.16}$$

Following (8.15), (8.16), and the fact that $|\frac{\partial(\eta_{l,1},\eta_{l,2})}{\partial(\theta_{l,1},\theta_{l,2})}|_{\xi=\zeta} = 1 + O(\rho^2)$, we find that (8.8) becomes

$$\begin{aligned}
\frac{\partial J(E_{\psi(\cdot,\xi)})}{\partial \xi_{k,1}}|_{\xi=\zeta} &= \int_{S^2} \mathcal{S}_k(\varphi) \Bigg(3r_k^2 \cos\theta_{k,1} \sin\theta_{k,2} \\
&\quad + \frac{O(\rho^4)}{\sin\theta_{k,2}} \Bigg) \, d\theta_k + \sum_{l\neq k} \int_{S^2} \mathcal{S}_l(\varphi) \frac{O(\rho^4)}{\sin\theta_{l,2}} \, d\theta_l, \\
\frac{\partial J(E_{\psi(\cdot,\xi)})}{\partial \xi_{k,2}}|_{\xi=\zeta} &= \int_{S^2} \mathcal{S}_k(\varphi) \Bigg(3r_k^2 \sin\theta_{k,1} \sin\theta_{k,2} \\
&\quad + \frac{O(\rho^4)}{\sin\theta_{k,2}} \Bigg) \, d\theta_k + \sum_{l\neq k} \int_{S^2} \mathcal{S}_l(\varphi) \frac{O(\rho^4)}{\sin\theta_{l,2}} \, d\theta_l, \\
\frac{\partial J(E_{\psi(\cdot,\xi)})}{\partial \xi_{k,3}}|_{\xi=\zeta} &= \int_{S^2} \mathcal{S}_k(\varphi) \left(3r_k^2 \cos\theta_{k,2} + \frac{O(\rho^4)}{\sin\theta_{k,2}} \right) \, d\theta_k \\
&\quad + \sum_{l\neq k} \int_{S^2} \mathcal{S}_l(\varphi) \frac{O(\rho^4)}{\sin\theta_{l,2}} \, d\theta_l.
\end{aligned}$$

Now we combine (7.1), (8.5), and the above to derive that at $\xi = \zeta$ and $r = s$,

$$\begin{aligned}
&A_{k,1} \int_{S^2} \cos\theta_{k,1} \sin\theta_{k,2} \left(3r_k^2 \cos\theta_{k,1} \sin\theta_{k,2} + \frac{O(\rho^4)}{\sin\theta_{k,2}} \right) d\theta_k + A_{k,2} O(\rho^4) \\
&\quad + A_{k,3} O(\rho^4) + \sum_{l\neq k} A_{l,1} O(\rho^4) + \sum_{l\neq k} A_{l,2} O(\rho^4) + \sum_{l\neq k} A_{l,3} O(\rho^4) = 0, \\
&A_{k,1} O(\rho^4) + A_{k,2} \int_{S^2} \sin\theta_{k,1} \sin\theta_{k,2} \left(3r_k^2 \sin\theta_{k,1} \sin\theta_{k,2} + \frac{O(\rho^4)}{\sin\theta_{k,2}} \right) d\theta_k \\
&\quad + A_{k,3} O(\rho^4) + \sum_{l\neq k} A_{l,1} O(\rho^4) + \sum_{l\neq k} A_{l,2} O(\rho^4) + \sum_{l\neq k} A_{l,3} O(\rho^4) = 0, \\
&A_{k,1} O(\rho^4) + A_{k,2} O(\rho^4) + A_{k,3} \int_{S^2} \cos\theta_{k,2} \left(3r_k^2 \cos\theta_{k,2} + \frac{O(\rho^4)}{\sin\theta_{k,2}} \right) d\theta_k \\
&\quad + \sum_{l\neq k} A_{l,1} O(\rho^4) + \sum_{l\neq k} A_{l,2} O(\rho^4) + \sum_{l\neq k} A_{l,3} O(\rho^4) = 0.
\end{aligned}$$

Writing the system in matrix form

$$\left(\begin{bmatrix} 4\pi r_1^2 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 4\pi r_1^2 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 4\pi r_1^2 & 0 & \dots & 0 & 0 \\ .. & & & & & & \\ .. & & & & & & \\ 0 & 0 & 0 & 0 & \dots & 4\pi r_K^2 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 4\pi r_K^2 \end{bmatrix} + O(\rho^4) \right) \begin{bmatrix} A_{1,1} \\ A_{1,2} \\ A_{1,3} \\ \dots \\ \dots \\ A_{K,2} \\ A_{K,3} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \dots \\ \dots \\ 0 \\ 0 \end{bmatrix} \tag{8.17}$$

we deduce, since (8.17) is nonsingular when ρ is small, that $A_{k,1} = A_{k,2} = A_{k,3} = 0$. ☐

The existence part of Theorem 2.1 follows from Lemma 8.4. The centers ζ_k and radii s_k of the spheres are found in Lemma 8.2. In Lemma 7.1 we see that

$\|\varphi\|_{W^{2,p}} \le c\rho^5$, which implies that the radius of a sphere is approximately

$$(s_k^3 + \varphi_k(\theta_k))^{1/3} = s_k + \frac{O(|\varphi_k(\theta_k)|)}{\rho^2} = s_k + O(\rho^3). \tag{8.18}$$

By Lemma 8.2, ζ is close to a minimum of F and s_k is close to ρ. The formula in Lemma 8.1 gives the free energy of our solution.

In Theorem 2.2, a solution is termed stable if it is a local minimizer of J in the space

$$U \times \left\{ \phi = (\phi_1, \dots, \phi_K) : \ |\rho^3 + \phi_k| \ge \frac{\rho^3}{2}, \ \phi_k \in W^{1,2}(S^2), \right. \\ \left. \phi_k \perp 1, \ \phi_k \perp H_1, \ k = 1, 2, \dots, K \right\}. \tag{8.19}$$

The condition $|\rho^3 + \phi_k| \ge \frac{\rho^3}{2}$ ensures that J is well defined in this space. Under the condition (2.8), Lemma 7.2, part 2, shows that each $\varphi(\cdot, \xi, r)$ we found in Lemma 7.1 locally minimizes J, with fixed $(\xi, r) \in U$, in $\{\phi : \ |\rho^3 + \phi_k| \ge \frac{\rho^3}{2}, \ \phi_k \in W^{1,2}(S^2), \ \phi_k \perp 1, \ \phi_k \perp H_1\}$. On the other hand $\varphi(\cdot, \zeta, s)$ minimizes $J(E_{\varphi(\cdot,\xi,r)})$ with respect to ξ and r. Hence $\varphi(\cdot, \zeta, s)$ is a local minimizer of J in (8.19).

If (2.9) holds, then we can find one eigenvalue $\lambda_{k,n}$ of $\mathcal{L}_1$, Lemma 5.1, for some $n \in \{2, 3, \dots\}$ such that

$$\lambda_{k,n} < -\frac{C}{\rho^4}, \quad \langle \mathcal{L}_1(e_{k,n}), e_{k,n} \rangle < -\frac{C}{\rho^4} \|e_{k,n}\|^2_{W^{1,2}},$$

where $e_{k,n}$ is an eigenvector corresponding to $\lambda_{k,n}$. By Lemma 5.2, the last inequality implies that

$$\langle \mathcal{L}(e_{k,n}), e_{k,n} \rangle < -\frac{C}{\rho^4} \|e_{k,n}\|^2_{W^{1,2}}.$$

Then by Lemma 6.1, parts 2, 3, and 4, and (7.25) in the proof of Lemma 7.2,

$$\langle \tilde{\mathcal{L}}(e_{k,n}), e_{k,n} \rangle < -\frac{C}{\rho^4} \|e_{k,n}\|^2_{W^{1,2}}.$$

Therefore the solution is unstable.

9. Discussion. The functional (1.2) is derived as a Γ-limit of the free energy functional in the Ohta–Kawasaki theory of diblock copolymers in [20]. Ohta and Kawasaki use a function u on D to describe the density of A-monomers and $1 - u$ to describe the density of B-monomers. The free energy of a diblock copolymer is

$$I(u) = \int_D \left[\frac{\varepsilon^2}{2} |Du|^2 + W(u) + \frac{\sigma}{2} |(-\Delta)^{-1/2}(u - a)|^2 \right] dx, \tag{9.1}$$

where u is in

$$\{u \in H^1(D) : \overline{u} = a\}. \tag{9.2}$$

The ε in (9.1) is not to be confused with the ϵ that has appeared in this paper. The function W is a balanced double-well potential such as $W(u) = \frac{1}{4}u^2(1 - u)^2$. There are three positive parameters in (9.1): ε, σ, and a, where ε is small and a is in $(0, 1)$.

These three-dimensionless parameters are related to several physical parameters of a diblock copolymer system. See [29] for the precise relationships between the dimensionless parameters here and the physical parameters.

If we take σ to be of order ε, i.e., by setting

$$\sigma = \varepsilon\gamma \tag{9.3}$$

for some γ independent of ε, then as ε tends to 0, the limiting problem of $\varepsilon^{-1}I$ turns out to be

$$J(E) = \tau|D\chi_E|(D) + \frac{\gamma}{2}\int_D |(-\Delta)^{-1/2}(\chi_E - a)|^2\,dx \tag{9.4}$$

which is the same as the J in (1.2) except for the additional constant τ here. This constant is known as the surface tension and is given by

$$\tau = \int_0^1 \sqrt{2W(q)}\,dq. \tag{9.5}$$

The functional (9.4) is defined on the same admissible set Σ, (1.3). In this paper we have taken $\tau = 1$ without loss of generality.

The theory of Γ-convergence was developed by De Giorgi [7], Modica and Mortola [14], Modica [13], and Kohn and Sternberg [11]. It was proved that $\varepsilon^{-1}I$ Γ-converges to J in the following sense.

PROPOSITION 9.1 (see Ren and Wei [20]).

1. *For every family* $\{u_\varepsilon\}$ *of functions in* (9.2) *satisfying* $\lim_{\varepsilon\to 0} u_\varepsilon = \chi_E$ *in* $L^2(D)$,

$$\liminf_{\varepsilon\to 0} \varepsilon^{-1}I(u_\varepsilon) \geq J(E).$$

2. *For every* E *in* Σ, *there exists a family* $\{u_\varepsilon\}$ *of functions in* (9.2) *such that* $\lim_{\varepsilon\to 0} u_\varepsilon = \chi_E$ *in* $L^2(D)$, *and*

$$\limsup_{\varepsilon\to 0} \varepsilon^{-1}I(u_\varepsilon) \leq J(E).$$

The relationship between I and J becomes more clear when a result of Kohn and Sternberg [11] was used to show the following.

PROPOSITION 9.2 (see Ren and Wei [20]). *Let* $\delta > 0$ *and* $E \in \Sigma$ *be such that* $J(E) < J(F)$ *for all* $\chi_F \in B_\delta(\chi_E)$ *with* $F \neq E$, *where* $B_\delta(\chi_E)$ *is the open ball of radius* δ *centered at* χ_E *in* $L^2(D)$. *Then there exists* $\varepsilon_0 > 0$ *such that for all* $\varepsilon < \varepsilon_0$ *there exists* $u_\varepsilon \in B_{\delta/2}(\chi_E)$ *with* $I(u_\varepsilon) \leq I(u)$ *for all* $u \in B_{\delta/2}(\chi_E)$. *In addition* $\lim_{\varepsilon\to 0} \|u_\varepsilon - \chi_E\|_{L^2(D)} = 0$.

The existence of a stable solution $E_{\varphi(\cdot,\zeta,s)}$ to (1.1) in the sense of Theorem 2.1 does not quite imply the existence of a local minimizer, close to $\chi_{E_{\varphi(\cdot,\zeta,s)}}$ in $L^2(D)$, of I. One must show that $E_{\varphi(\cdot,\zeta,s)}$ is a strict local minimizer in the sense of Proposition 9.2. This issue requires more study.

Our work is the first mathematically rigorous confirmation of the spherical phase of diblock copolymer morphology. This phase, depicted in Figure 1.1, plot 1, has been observed in experiments for some time [1]. Our earlier work [31, 30] in two dimensions gave a mathematical proof of the existence of the cylindrical phase of diblock copolymer morphology; see Figure 1.1, plot 2. The results obtained here are analogous to the ones obtained in [30], but there are some notable differences.

In two dimensions we studied a cross section of the cylindrical phase and constructed a stable solution which is a union of many small, approximate discs under the condition that

$$\frac{1+\epsilon}{\rho^3 \log \frac{1}{\rho}} < \gamma < \frac{12-4\epsilon}{\rho^3}. \tag{9.6}$$

Here ρ is the average disc radius defined by $\rho = \sqrt{\frac{a|D|}{K\pi}}$. Note that the two bounds for γ in (9.6) are of different orders. Recall that in three dimensions we have (2.12), i.e.,

$$\frac{3+\epsilon}{\rho^3} < \gamma < \frac{30-4\epsilon}{\rho^3}, \tag{9.7}$$

where the two bounds are of the same order. In experiments it is more likely to see the cylindrical phase than the spherical phase (see [1]). The different bounds in (9.6) and (9.7) appear to offer an explanation.

In (8.18) we have proved that the perturbed "radius" is

$$(s_k^3 + \varphi_k(\theta_k))^{1/3} = s_k + O(\rho^3). \tag{9.8}$$

In other words the deviation of the "radius" of a perturbed ball from an exact ball is of the order $O(\rho^3)$. However, in two dimensions the corresponding quantity is

$$(s_k^2 + \varphi_k(\theta_k))^{1/2} = s_k + O(\rho^2), \tag{9.9}$$

a fact found after the proof of [30, Theorem 2.1]. The approximate balls in the spherical solution found here are more round than the approximate discs in the cylindrical solution found in [30].

Appendix A. We drop the subscript k in this appendix. The derivative of $\mathcal{A}$ at 0 has two terms according to (4.29). The first is

$$\frac{\gamma}{9r_k}\int_{S^2} \frac{u(\omega)}{4\pi|\theta-\omega|}\, d\omega.$$

The second is

$$-\frac{\gamma u(\theta)}{9r_k}\int_{B_1(0)} \frac{(\theta-y)\cdot\theta}{4\pi|\theta-y|^2}\, dy$$

for which we calculate the integral. Here $B_1(0)$ is the unit ball. This integral is independent of $\theta \in S^2$ so without loss of generality we assume that $\theta = (0,0,1)$. Write $y = (r\cos p, r\sin p, y_3)$ in the cylindrical coordinates. Then the integral becomes

$$\int_{B_1(0)} \frac{(\theta-y)\cdot\theta}{4\pi|\theta-y|^2}\, dy = \frac{1}{4\pi}\int_{-1}^{1}\int_0^{2\pi}\int_0^{\sqrt{1-y_3^2}} \frac{(1-y_3)r\, dr\, dp\, dy_3}{[(1-y_3)^2+r^2]^{3/2}} = \frac{1}{3}.$$

Appendix B. The integral operator

$$h(\theta) \to \int_{S^2} \frac{h(\omega)\, d\omega}{|\theta-\omega|} \tag{B.1}$$

acts on spherical harmonics $h \in H_n$ in a simple way. Here H_n is the space of spherical harmonics of degree n on S^2. In general one has

$$\int_{S^2} \Phi(\theta \cdot \omega) h(\omega)\, d\omega = \alpha_n(\Phi) h(\theta), \tag{B.2}$$

where

$$\alpha_n(\Phi) = 2\pi \int_{-1}^{1} \Phi(t) P_n(t)\, dt. \tag{B.3}$$

See, for instance, [10, Theorem 3.4.1]. Here P_n is the nth Legendre polynomial. In our case

$$\frac{1}{|\theta - \omega|} = \frac{1}{\sqrt{2 - 2\theta \cdot \omega}},$$

so we take

$$\Phi(t) = \frac{1}{\sqrt{2 - 2t}}. \tag{B.4}$$

The classical representation of Legendre polynomials in terms of generating functions [10, Formula 3.3.39]

$$\frac{1}{(1 + r^2 - 2rt)^{1/2}} = \sum_{n=0}^{\infty} P_n(t) r^n, \quad r, t \in (-1, 1) \tag{B.5}$$

shows that

$$\int_{-1}^{1} \frac{P_n(t)\, dt}{(1 + r^2 - 2rt)^{1/2}} = r^n \int_{-1}^{1} P_n^2(t)\, dt = \frac{2r^n}{2n+1},$$

where the orthogonality of the Legendre polynomials is used [10, Formula 3.3.16]:

$$\int_{-1}^{1} P_n(t) P_m(t)\, dt = \frac{2\delta_{nm}}{2n+1}.$$

By sending $r \to 1$ we find that

$$\alpha_n(\Phi) = \frac{4\pi}{2n+1}. \tag{B.6}$$

Appendix C. Here we calculate the improper integral

$$\int_{B_1(0)} \frac{|\theta - y|^2 - 3(1 - \theta \cdot y)^2}{|\theta - y|^5}\, dy, \tag{C.1}$$

where $B_1(0)$ is the unit ball centered at 0. This integral is independent of $\theta \in S^2$. We take $\theta = (0,0,1)$. Let $z = (0,0,1) - y$ and set $z = (r\cos p, r\sin p, z_3)$ in cylindrical coordinates. Then

$$\begin{aligned}
&\int_{B_1(0)} \frac{|\theta - y|^2 - 3(1 - \theta \cdot y)^2}{|\theta - y|^5}\, dy \\
&\quad = \int_{B_1(0,0,1)} \frac{|z|^2 - 3z_3^2}{|z|^5}\, dz \\
&\quad = \int_0^2 \int_0^{\sqrt{1-(1-z_3)^2}} \int_0^{2\pi} \frac{(r^2 + z_3^2) - 3z_3^2}{(r^2 + z_3^2)^{5/2}} r\, dp dr dz_3 \;=\; -\frac{8\pi}{3}.
\end{aligned}$$

REFERENCES

[1] F. S. Bates and G. H. Fredrickson, *Block copolymers–designer soft materials*, Physics Today, 52 (1999), pp. 32–38.

[2] X. Chen and Y. Oshita, *Periodicity and uniqueness of global minimizers of an energy functional containing a long-range interaction*, SIAM J. Math. Anal., 37 (2005), pp. 1299–1332.

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

[4] R. Choksi and X. Ren, *On the derivation of a density functional theory for microphase separation of diblock copolymers*, J. Statist. Phys., 113 (2003), pp. 151–176.

[5] R. Choksi and X. Ren, *Diblock copolymer/homopolymer blends: Derivation of a density functional theory*, Phys. D, 203 (2005), pp. 100–119.

[6] R. Choksi and P. Sternberg, *On the first and second variations of a nonlocal isoperimetric problem*, J. Reine Angew. Math., 611 (2007), pp. 75–108.

[7] E. De Giorgi, *Sulla convergenza di alcune successioni d'integrali del tipo dell'area*, Rend. Mat., 8 (1975), pp. 277–294.

[8] M. del Pino, M. Kowalczyk, and J.-C. Wei, *Concentration on curves for nonlinear Schrödinger equations*, Comm. Pure Appl. Math., 60 (2007), pp. 113–146.

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

[10] H. Groemer, *Geometric Applications of Fourier Series and Spherical Harmonics*, Cambridge University Press, Cambridge, UK, 1996.

[11] R. Kohn and P. Sternberg, *Local minimisers and singular perturbations*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 69–84.

[12] A. Malchiodi and M. Montenegro, *Boundary concentration phenomena for a singularly perturbed elliptic problem*, Comm. Pure Appl. Math., 55 (2002), pp. 1507–1568.

[13] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 98 (1987), pp. 123–142.

[14] L. Modica and S. Mortola, *Un esempio di Γ^--convergenza*, Boll. Un. Mat. Ital. B (5), 14 (1977), pp. 285–299.

[15] S. Müller, *Singular perturbations as a selection criterion for periodic minimizing sequences*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 169–204.

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

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

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

[19] F. Pacard and M. Ritoré, *From constant mean curvature hypersurfaces to the gradient theory of phase transitions*, J. Differential Geom., 64 (2003), pp. 359–423.

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

[21] X. Ren and J. Wei, *Concentrically layered energy equilibria of the di-block copolymer problem*, European J. Appl. Math., 13 (2002), pp. 479–496.

[22] X. Ren and J. Wei, *On energy minimizers of the diblock copolymer problem*, Interfaces Free Bound., 5 (2003), pp. 193–238.

[23] X. Ren and J. Wei, *On the spectra of three-dimensional lamellar solutions of the diblock copolymer problem*, SIAM J. Math. Anal., 35 (2003), pp. 1–32.

[24] X. Ren and J. Wei, *Triblock copolymer theory: Free energy, disordered phase and weak segregation*, Phys. D, 178 (2003), pp. 103–117.

[25] X. Ren and J. Wei, *Triblock copolymer theory: Ordered ABC lamellar phase*, J. Nonlinear Sci., 13 (2003), pp. 175–208.

[26] X. Ren and J. Wei, *Stability of spot and ring solutions of the diblock copolymer equation*, J. Math. Phys., 45 (2004), pp. 4106–4133.

[27] X. Ren and J. Wei, *Wriggled lamellar solutions and their stability in the diblock copolymer problem*, SIAM J. Math. Anal., 37 (2005), pp. 455–489.

[28] X. Ren and J. Wei, *Droplet solutions in the diblock copolymer problem with skewed monomer composition*, Calc. Var. Partial Differential Equations, 25 (2006), pp. 333–359.

[29] X. Ren and J. Wei, *Existence and stability of spherically layered solutions of the diblock copolymer equation*, SIAM J. Appl. Math., 66 (2006), pp. 1080–1099.

[30] X. Ren and J. Wei, *Many droplet pattern in the cylindrical phase of diblock copolymer morphology*, Rev. Math. Phys., 19 (2007), pp. 879–921.

[31] X. Ren and J. Wei, *Single droplet pattern in the cylindrical phase of diblock copolymer morphology*, J. Nonlinear Sci., 17 (2007), pp. 471–503.

[32] S. W. Semmes, *A criterion for the boundedness of singular integrals on hypersurfaces*, Trans. Amer. Math. Soc., 311 (1989), pp. 501–513.

[33] T. Teramoto and Y. Nishiura, *Double gyroid morphology in a gradient system with nonlocal effects*, J. Phys. Soc. Japan, 71 (2002), pp. 1611–1614.

SIAM J. MATH. ANAL.
Vol. 39, No. 5, pp. 1536–1549

ON CONVERGENCE OF SOLUTIONS OF FRACTAL BURGERS EQUATION TOWARD RAREFACTION WAVES*

GRZEGORZ KARCH†, CHANGXING MIAO‡, AND XIAOJING XU§

Abstract. In this paper, the large time behavior of solutions of the Cauchy problem for the one-dimensional fractal Burgers equation $u_t + (-\partial_x^2)^{\alpha/2}u + uu_x = 0$ with $\alpha \in (1, 2)$ is studied. It is shown that if the nondecreasing initial datum approaches the constant states $u_\pm$ ($u_- < u_+$) as $x \to \pm\infty$, respectively, then the corresponding solution converges toward the rarefaction wave, i.e., the unique entropy solution of the Riemann problem for the nonviscous Burgers equation.

Key words. fractal Burgers equation, asymptotic behavior, rarefaction wave, Riemann problem, Lévy process

AMS subject classifications. 60J60, 35B40, 35K55, 35Q53

DOI. 10.1137/070681776

1. Introduction. The goal of this work is to study asymptotic properties of solutions of the Cauchy problem for the nonlocal conservation law

$$u_t + \Lambda^\alpha u + uu_x = 0, \quad x \in \mathbb{R},\ t > 0, \tag{1.1}$$

$$u(0, x) = u_0(x), \tag{1.2}$$

where $\Lambda^\alpha = (-\partial^2/\partial x^2)^{\alpha/2}$ is the pseudodifferential operator defined via the Fourier transform

$$\widehat{(\Lambda^\alpha v)}(\xi) = |\xi|^\alpha \widehat{v}(\xi). \tag{1.3}$$

Following [3], we will call (1.1) the fractal Burgers equation. Equations of this type appear in the study of growing interfaces in the presence of self-similar hopping surface diffusion [19]. Moreover, in their recent papers, Jourdain, Méléard, and Woyczyński [13, 14] gave probabilistic motivations for studying equations with the anomalous diffusion when the Laplacian (the generator of the Wiener process) is replaced by a more general pseudodifferential operator generating the Lévy process. In particular, the authors of [14] studied problem (1.1)–(1.2), where the initial condition u_0 is assumed to be a nonconstant function with bounded variation on $\mathbb{R}$. In other words, a.e. on $\mathbb{R}$,

$$u_0(x) = c + \int_{-\infty}^{x} m(dy) = c + H * m(x) \tag{1.4}$$

*Received by the editors February 4, 2007; accepted for publication (in revised form) July 24, 2007; published electronically January 22, 2008.
http://www.siam.org/journals/sima/39-5/68177.html

†Instytut Matematyczny, Uniwersytet Wrocławski, pl. Grunwaldzki 2/4, 50-384 Wrocław, Poland (karch@math.uni.wroc.pl). This author was partially supported by the European Commission Marie Curie Host Fellowship for the Transfer of Knowledge "Harmonic Analysis, Nonlinear Analysis and Probability" MTKD-CT-2004-013389 and the grant N201 022 32 / 09 02 of MNiSW.

‡Institute of Applied Physics and Computational Mathematics, P.O. Box 8009, Beijing 100088, People's Republic of China (miao_changxing@iapcm.ac.cn). This author was partially supported by the NSF of China through grant 10725102.

§School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems, Ministry of Education, Beijing 100875, People's Republic of China (xjxu@bnu.edu.cn). This author was partially supported by the NSF of China through grants 10571016 and 10601009.

with $c \in \mathbb{R}$, m being a finite signed measure on $\mathbb{R}$, and $H(y)$ denoting the unit step function $\mathbb{1}_{\{y\geq 0\}}$. Observe that the gradient $v(x,t) = u_x(x,t)$ satisfies

$$v_t + \Lambda^\alpha v + (vH * v)_x = 0, \quad v(\cdot, 0) = m. \tag{1.5}$$

If m is a probability measure on $\mathbb{R}$, (1.5) is a nonlinear Fokker–Planck equation. In the case of an arbitrary finite signed measure, the authors of [14] associated (1.5) with a suitable nonlinear martingale problem. Next, they studied the convergence of systems of particles with jumps as the number of particles tends to $+\infty$. As a consequence, the weighted empirical cumulative distribution functions of the particles converge to the solution of the martingale problem connected to (1.5). This phenomenon is called the *propagation of chaos* for problem (1.1)–(1.2), and we refer the reader to [14] for more details and additional references.

Motivated by the results from [14], we study problem (1.1)–(1.2) under the crucial assumption $\alpha \in (1, 2)$ and with the initial condition of the form (1.4). In our main result, we assume that u_0 is a function satisfying

$$u_0 - u_- \in L^1((-\infty, 0)) \quad \text{and} \quad u_0 - u_+ \in L^1((0, +\infty)) \quad \text{with} \quad u_- < u_+, \tag{1.6}$$

where $u_- = c$ and $u_+ - u_- = \int_{\mathbb{R}} m(dx)$.

It is well known (cf. [11, 20, 10] and Lemma 2.4 below) that the asymptotic profile as $t \to \infty$ of solutions of the viscous Burgers equation

$$u_t - u_{xx} + uu_x = 0 \tag{1.7}$$

(i.e., (1.1) with $\alpha = 2$) supplemented with an initial datum satisfying (1.6) is given by the so-called rarefaction wave. This is the continuous function

$$w^R(x,t) = W^R(x/t) = \begin{cases} u_-\,, & x/t \leq u_-, \\ x/t\,, & u_- \leq x/t \leq u_+, \\ u_+\,, & x/t \geq u_+, \end{cases} \tag{1.8}$$

which is the unique entropy solution of the following Riemann problem:

$$w_t^R + w^R w_x^R = 0, \tag{1.9}$$

$$w^R(x,0) = w_0^R(x) = \begin{cases} u_-\,, & x < 0, \\ u_+\,, & x > 0. \end{cases} \tag{1.10}$$

Below, we use the solution of the Burgers equation (1.7) with the initial datum (1.10) as the smooth approximation of the rarefaction wave (1.8).

The authors of this work were inspired by the fundamental paper of Il'in and Oleinik [11] who showed the convergence toward rarefaction waves of solutions of the nonlinear equation $u_t - u_{xx} + f(u)_x = 0$ under a strict convexity assumption imposed on f. That idea was next extended in several different directions, and we refer the reader to, e.g., [10, 18, 20, 21, 23, 24] for an overview of known results and additional references.

In this work, we contribute to the existing theory by developing tools which allow us to obtain analogous results for equations with a nonlocal and anomalous diffusion. Basic properties of solutions (namely, their existence and the regularity) of quasilinear evolution equations with $(-\Delta)^{\alpha/2}$, $\alpha \in (1, 2)$, (or, more generally, with the

Lévy diffusion) were shown in [8, 9]. On the other hand, one may expect singularities in finite time of solutions of (1.1) with $\alpha \in (0,1)$; see [2] for more details. If $u_0 \in L^1(\mathbb{R})$, Biler, Karch, and Woyczyński [4] proved that the large time asymptotics of solutions of (1.1)–(1.2) is described by the self-similar fundamental solution of equation $v_t + \Lambda^\alpha v = 0$. Analogous asymptotic properties of solutions of multidimensional generalizations of problem (1.1)–(1.2) with $u_0 \in L^1(\mathbb{R}^N)$ were studied in [5, 6]. Here, we would like also to report the recent progress in the understanding of properties of solutions of the quasi-geostrophic equation with an anomalous diffusion; cf. [7, 15] and the references therein.

The purpose of the present paper is to prove the convergence of solutions of the Cauchy problem for the fractal Burgers equations (1.1)–(1.2) toward rarefaction waves. We state our main result in the following theorem.

THEOREM 1.1. *Let $\alpha \in (1,2)$. Assume that $w^R = w^R(x,t)$ is the rarefaction wave* (1.8). *Denote by $u = u(x,t)$ the unique solution of problem* (1.1)–(1.2) *corresponding to the initial datum u_0 of the form* (1.4) *and satisfying* (1.6) *(cf. Theorem* 2.1*). For every $p \in ((3-\alpha)/(\alpha-1), \infty]$ there exists $C > 0$ independent of t such that*

$$\|u(t) - w^R(t)\|_p \le Ct^{-[\alpha-1-(3-\alpha)/p]/2} \log(2+t)$$

for all $t > 0$.

Remark 1.1. The proof of Theorem 1.1 does not work for $\alpha \in (0,1]$. In fact, our preliminary computations confirm that, in this case, one should expect completely different asymptotic profiles of solutions. First of all, the initial value problem (1.1)–(1.2) has the unique global-in-time entropy solution for every $u_0 \in L^\infty(\mathbb{R})$ and $\alpha \in (0,1]$ due to the recent work of Alibaud [1]. We expect these solutions to be regular for nondecreasing initial conditions, namely, in this case when shocks do not form for the nonviscous Burgers equation $u_t + uu_x = 0$. Using the uniqueness result from [1] combined with a standard scaling technique, one can show that (1.1) with $\alpha = 1$ has self-similar solutions of the form $u(x,t) = U(x/t)$. These profiles determine the large time asymptotics of solutions to the initial value problem (1.1)–(1.2) with $\alpha = 1$. If $\alpha \in (0,1)$, the Duhamel principle applied to problem (1.1)–(1.2) (cf. (2.10)) allows us to derive asymptotic profiles of solutions in the form

$$U_\alpha(x/t^{1/\alpha}) = \int_{-\infty}^{x/t^{1/\alpha}} P_\alpha(z)\,dz,$$

where the function P_α is as defined below (see (2.11) and (2.12)). Precise statements of these results and their proofs will be published in [12].

Remark 1.2. Our result has an important probabilistic interpretation, because the rarefaction wave (1.8) with $u_- = 0$ and $u_+ = 1$ is the probability distribution function corresponding to the uniform distribution on the interval $[0,t]$. On the other hand, the results announced in the remark above say that solutions to (1.1)–(1.2) with $\alpha \in (0,1)$ converge as $t \to \infty$ toward the symmetric α-stable laws and toward a one-parameter family of new laws solving the nonlinear equation (1.1) if $\alpha = 1$.

Remark 1.3. The result from Theorem 1.1 and its proof hold true also for $\alpha = 2$ (observe that $(3-\alpha)/(\alpha-1) \to 1$ as $\alpha \to 2$). However, we pass over this case for simplicity of the exposition and because the large time asymptotics of solutions of the Burgers equation (1.7) is well known; see Lemma 2.4.

In the next section, we gather several preliminary properties of the operator Λ^α and of solutions of problem (1.1)–(1.2). Theorem 1.1 is shown in section 3. In section 4, we discuss possible generalizations of our main result.

Notation. For $1 \le p \le \infty$, the L^p-norm of a Lebesgue measurable, real-valued function v defined on $\mathbb{R}$ is denoted by $\|v\|_p$. For a finite signed measure m on $\mathbb{R}$, we put $\|m\| = |m|(\mathbb{R})$, where $|m|$ is the total variation of m. The Fourier transform of v is $\widehat{v}(\xi) \equiv (2\pi)^{-1/2} \int_{\mathbb{R}} e^{-ix\xi} v(x)\, dx$. Given a function $v = v(x)$, we are going to use the decomposition $v = v^+ - v^-$, where as usual $v^- = \max\{0, -v\}$ and $v^+ = \max\{0, v\}$. The constants (always independent of t) will be denoted by the same letter C, even if they may vary from line to line. Occasionally, we write, e.g., $C = C(\alpha, \ell)$ when we want to emphasize the dependence of C on parameters α and ℓ.

2. Preliminary results. We begin by recalling that the basic questions on the existence and the uniqueness of solutions of problem (1.1)–(1.2) have been answered in [8, 9].

THEOREM 2.1 (see [8, Thm. 1.1], [9, Thm. 7]). *Let $\alpha \in (1, 2)$ and $u_0 \in L^\infty(\mathbb{R})$. There exists the unique solution $u = u(x,t)$ of problem* (1.1)–(1.2) *in the following sense: for all $T > 0$,*

$$u \in C_b((0, T) \times \mathbb{R}) \textit{ and, for all } a \in (0, T),\ u \in C_b^\infty((a, T) \times \mathbb{R}),$$

$$u \textit{ satisfies } (1.1) \textit{ on } (0, T) \times \mathbb{R},$$

$$u(t, \cdot) \to u_0 \textit{ in } L^\infty(\mathbb{R}) \textit{ weak-}*, \textit{ as } t \to 0.$$

Moreover, the following inequality holds true:

$$\|u(t)\|_\infty \le \|u_0\|_\infty \quad \textit{for all} \quad t > 0. \tag{2.1}$$

The main goal of this section is to complete this result by additional properties of u_x if the initial conditions are of the form (1.4).

THEOREM 2.2. *Let $\alpha \in (1, 2)$. Assume that the initial datum u_0 can be written in the form* (1.4) *for a constant $c \in \mathbb{R}$ and a signed finite measure m on $\mathbb{R}$. Then the solution $u = u(x,t)$ of problem* (1.1)–(1.2) *satisfies $u_x \in C((0,T]; L^p(\mathbb{R}))$ for each $1 \le p \le \infty$ and every $T > 0$.*

Consider u and $\widetilde{u}$ to be two such solutions with initial conditions u_0 and $\widetilde{u}_0$, respectively. Suppose that $\widetilde{u}_x(x,t)$ is nonnegative a.e. and $u_0 - \widetilde{u}_0 \in L^1(\mathbb{R})$. Then

$$\|u(t) - \widetilde{u}(t)\|_1 \le \|u_0 - \widetilde{u}_0\|_1 \tag{2.2}$$

for all $t > 0$.

THEOREM 2.3. *Under the assumption of Theorem* 2.2, *if the measure m in the initial datum* (1.4) *is nonnegative, we have*

(i) *$u_x(x,t) \ge 0$ for all $x \in \mathbb{R}$ and $t > 0$;*

(ii) *for every $p \in [1, \infty]$ there exists $C = C(p) > 0$ such that*

$$\|u_x(t)\|_p \le t^{-1+1/p} \|m\|^{1/p}. \tag{2.3}$$

In the proofs of Theorems 2.2 and 2.3 as well as in our study of the large time asymptotics, we shall require several properties of the operator Λ^α and of the semigroup of linear operators generated by it. First of all, note that the operator defined by (1.3) has the integral representation for every $\alpha \in (1,2)$ (cf., e.g., [9, Thm. 1])

$$\Lambda^\alpha w(x) = -C(\alpha) \int_{\mathbb{R}} \frac{w(x+z) - w(x) - w_x(x) z}{|z|^{1+\alpha}}\, dz. \tag{2.4}$$

This formula allows us to apply Λ^α to functions which are bounded and sufficiently smooth but not, however, necessarily decaying at infinity.

LEMMA 2.1. *Let $1 < \alpha < 2$. For every $p \in [1, \infty]$ there exists $C = C(p, \alpha) > 0$ such that*

$$\|\Lambda^\alpha w\|_p \le C \|w_x\|_p^{2-\alpha} \|w_{xx}\|_p^{\alpha-1} \tag{2.5}$$

for all functions w satisfying $w_x, w_{xx} \in L^p(\mathbb{R})$.

Proof. We can easily deduce the interpolation inequality (2.5) from (2.4). Indeed, it follows from the Taylor formula that for any fixed $R > 0$ we have

$$\begin{aligned}\|\Lambda^\alpha w\|_p &\le C\|w_{xx}\|_p \int_{|z|\le R} |z|^{1-\alpha}\, dz + C\|w_x\|_p \int_{|z|>R} |z|^{-\alpha}\, dz \\ &\le C\left(R^{2-\alpha}\|w_{xx}\|_p + R^{1-\alpha}\|w_x\|_p\right).\end{aligned}$$

Choosing $R = \|w_x\|_p/\|w_{xx}\|_p$ we complete the proof of inequality (2.5). ∎

Now, we prove the Nash inequality for the operator Λ^α.

LEMMA 2.2. *Let $0 < \alpha$. There exists a constant $C_N > 0$ such that*

$$\|w\|_2^{2(1+\alpha)} \le C_N \|\Lambda^{\alpha/2} w\|_2^2 \|w\|_1^{2\alpha} \tag{2.6}$$

for all functions w satisfying $w \in L^1(\mathbb{R})$ and $\Lambda^{\alpha/2} w \in L^2(\mathbb{R})$.

Proof. For every $R > 0$, we decompose the L^2-norm of the Fourier transform of w as follows:

$$\begin{aligned}\|w\|_2^2 &= C\int_{\mathbb{R}} |\widehat{w}(\xi)|^2\, d\xi \\ &\le C\|\widehat{w}\|_\infty^2 \int_{|\xi|\le R} d\xi + CR^{-\alpha}\int_{|\xi|>R} |\xi|^\alpha |\widehat{w}(\xi)|^2\, d\xi \\ &\le CR\|w\|_1^2 + CR^{-\alpha}\|\Lambda^{\alpha/2} w\|_2^2.\end{aligned}$$

For $R = \left(\|\Lambda^{\alpha/2} w\|_2^2/\|w\|_1^2\right)^{1/(1+\alpha)}$ we obtain (2.6). ∎

LEMMA 2.3. *Let $0 \le \alpha \le 2$. For every $p > 1$, we have*

$$\int_{\mathbb{R}} (\Lambda^\alpha w)|w|^{p-2} w\, dx \ge \frac{4(p-1)}{p^2} \int_{\mathbb{R}} \left(\Lambda^{\frac{\alpha}{2}} |w|^{\frac{p}{2}}\right)^2\, dx \tag{2.7}$$

for all $w \in L^p(\mathbb{R})$ such that $\Lambda^\alpha w \in L^p(\mathbb{R})$. If $\Lambda^\alpha w \in L^1(\mathbb{R})$, we obtain

$$\int_{\mathbb{R}} (\Lambda^\alpha w) \operatorname{sgn} w\, dx \ge 0, \tag{2.8}$$

and if $w, \Lambda^\alpha w \in L^2(\mathbb{R})$, it follows that

$$\int_{\mathbb{R}} (\Lambda^\alpha w) w^+\, dx \ge 0 \quad \text{and} \quad \int_{\mathbb{R}} (\Lambda^\alpha w) w^-\, dx \ge 0, \tag{2.9}$$

where $w^+ = \max\{0, w\}$ and $w^- = \max\{0, -w\}$.

Inequality (2.7) is well known in the theory of sub-Markovian operators and its statement and proof are given, e.g., in [17, Thm. 2.1 combined with the Beurling–Deny condition (1.7)]; see also [7, 15]. Observe that if $\alpha = 2$, integrating by parts we obtain (2.7) with the equality. Inequality (2.8) (called the Kato inequality) is used in [9] to construct entropy solutions of (1.1) and it can be easily deduced from [9, Lem. 1] by an approximation argument (see also [3, inequality (3.5)]). The proof of (2.9) can be found, for example, in [17, Prop. 1.6].

We also recall that, by Duhamel's principle, the solution of problem (1.1)–(1.2) can be written in the equivalent integral form

$$u(t) = S_\alpha(t)u_0 - \int_0^t S_\alpha(t-\tau)u(\tau)u_x(\tau)\,d\tau, \tag{2.10}$$

where

$$S_\alpha(t)u_0 = p_\alpha(t) * u_0(x). \tag{2.11}$$

Here, the fundamental solution $p_\alpha(x,t)$ of the linear equation $\partial_t v + \Lambda^\alpha v = 0$ can be computed via the Fourier transform $\widehat{p}_\alpha(\xi,t) = e^{-t|\xi|^\alpha}$. Hence,

$$p_\alpha(x,t) = t^{-1/\alpha}P_\alpha(xt^{-1/\alpha}),$$

where P_α is the inverse Fourier transform of $e^{-|\xi|^\alpha}$. It is well known that for every $\alpha \in (0,2]$ the function P_α has the property $\int_{\mathbb{R}} P_\alpha(x)\ dx = 1$ and it is smooth, nonnegative, and satisfies

$$0 \le P_\alpha(x) \le C(1+|x|)^{-(\alpha+1)} \qquad \text{and} \qquad |\partial_x P_\alpha(x)| \le C(1+|x|)^{-(\alpha+2)} \tag{2.12}$$

for a constant C and all $x \in \mathbb{R}$. Using these properties of the convolution operator $S_\alpha(t)$ defined by (2.11) we obtain the estimates

$$\|S_\alpha(t)v\|_p \le Ct^{-(1-1/p)/\alpha}\|v\|_1, \tag{2.13}$$

$$\|(S_\alpha(t)v)_x\|_p \le Ct^{-(1-1/p)/\alpha-1/\alpha}\|v\|_1 \tag{2.14}$$

for every $p \in [1,\infty]$ and all $t > 0$. Moreover, we can replace v in (2.13) and (2.14) by any signed measure m. In that case, $\|v\|_1$ should be replaced by $\|m\|$.

Proof of Theorem 2.2. It follows from the integral equation (2.10) that u_x is the solution of

$$u_x(t) = S_\alpha(t)m - \int_0^t \partial_x S_\alpha(t-\tau)V(\tau)u_x(\tau)\,d\tau, \tag{2.15}$$

where $V(x,t) = u(x,t)$ is treated as given and is smooth and bounded. Now the standard argument involving the Banach fixed point theorem allows us to show that the "linear" equation (2.15) has a unique solution in $C((0,T];L^p(\mathbb{R}))$ for each $p \in [1,\infty]$ and every $T > 0$. Here, we should use the following estimate of the operator $\mathcal{T}(u)$ defined by the right-hand side of (2.15):

$$\begin{aligned}\|\mathcal{T}(u)(t)\|_p &\le \|S_\alpha(t)m\|_p + \int_0^t \|\partial_x S_\alpha(t-\tau)V(\tau)u_x(\tau)\|_p\,d\tau \\ &\le Ct^{-(1-1/p)/\alpha}\|m\| + C\sup_{\tau\in[0,T]}\|V(\tau)\|_\infty \int_0^t (t-\tau)^{-1/\alpha}\|u_x(\tau)\|_p\,d\tau,\end{aligned}$$

which is the immediate consequence of (2.13) and (2.14). Let us skip the other details of this well-known argument (cf. [22]).

Now, we prove inequality (2.2). A direct calculation shows that the function $v(x,t) = u(x,t) - \widetilde{u}(x,t)$ satisfies

$$v_t + \Lambda^\alpha v + \frac{1}{2}(v^2 + 2v\widetilde{u}))_x = 0. \tag{2.16}$$

First, we multiply (2.16) by $\operatorname{sgn} v = v|v|^{-1}$:

$$\frac{d}{dt}\int_{\mathbb{R}} |v|\,dx + \int_{\mathbb{R}} (\Lambda^\alpha v)\operatorname{sgn} v\,dx + \frac{1}{2}\int_{\mathbb{R}} [v^2 + 2v\widetilde{u})]_x \operatorname{sgn} v\,dx = 0.$$

The second term is nonnegative by (2.8). To show the same property for the third term, we replace the sgn function by smooth and nondecreasing $\varphi = \varphi(x)$. In this case, we obtain

$$\int_{\mathbb{R}} [v^2 + 2v\widetilde{u}]_x \varphi(v) dx = -\int_{\mathbb{R}} (v^2 + 2v\widetilde{u})\varphi'(v) v_x\,dx = -\int_{\mathbb{R}} \Psi(v)_x\,dx + \int_{\mathbb{R}} \widetilde{u}_x \Phi(v)\,dx,$$

where $\Psi(s) = \int_0^s z^2\varphi'(z)\,dz$ and $\Phi(s) = \int_0^s 2z\varphi'(z)\,dz$. Obviously, the first term on the right-hand side is equal to zero and the second one is nonnegative because $\widetilde{u}_x \ge 0$ and $\Phi(s) \ge 0$ for all $s \in \mathbb{R}$. Now, the standard approximation argument gives $\int_{\mathbb{R}} [v^2 + 2v\widetilde{u}]_x \operatorname{sgn} v\,dx \ge 0$. Hence, $\|v(t)\|_1 = \|u(t) - \widetilde{u}(t)\|_1 \le \|u_0 - \widetilde{u}_0\|_1 = \|v_0\|_1$ for all $t > 0$. $\square$

Proof of Theorem 2.3. To show part (i) of Theorem 2.3, we deal first with the smooth initial datum u_0 satisfying $u_{0,x}(x) \ge 0$ and $u_{0,x} \in L^p(\mathbb{R})$ for every $p \in [1,\infty]$. In this case, differentiating (1.1) with respect to x we have

$$(u_x)_t + \Lambda^\alpha u_x + (uu_x)_x = 0. \tag{2.17}$$

Note the well-known property

$$\int_{\mathbb{R}} v_t v^-\,dx = \int_{v\le 0} v_t^- v^-\,dx = \frac{1}{2}\frac{d}{dt}\int_{v\le 0} (v^-)^2\,dx.$$

Hence, multiplying (2.17) by u_x^-, integrating the resulting equation over $\mathbb{R}$, and integrating by parts on the right-hand side, we obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_{u_x\le 0} (u_x^-)^2\,dx + \int_{\mathbb{R}} (\Lambda^\alpha u_x)u_x^-\,dx &= -\int_{u_x\le 0} (uu_x^-)_x u_x^-\,dx \\ &= -\frac{1}{2}\int_{u_x\le 0} (u_x^-)^3\,dx.\end{aligned}$$

Since $\int_{\mathbb{R}} (\Lambda^\alpha u_x)u_x^-\,dx \ge 0$ by (2.9) and $\int_{u_x\le 0} (u_x^-(x,0))^2\,dx = 0$ by the assumption imposed on u_0, the Gronwall inequality implies $\int_{u_x\le 0} (u_x^-(x,t))^2\,dx = 0$ for all $t \ge 0$. Consequently, $u_x^-(x,t) \equiv 0$ and the proof of (i) for regular initial conditions is finished.

Now, the proof of (i) for the solution $u = u(x,t)$ corresponding to the initial datum u_0 of the form (1.4) with the nonnegative finite measure m can be completed by the following approximation argument. We consider the sequence of regular initial conditions u_0^n as in the first part of this proof. Moreover, we assume that $u_{0,x}^n$ converges weakly to m and $\|u_0^n - u_0\|_1 \to 0$ as $n \to \infty$. Inequality (2.2) allows us to prove that the corresponding solutions $u^n(\cdot,t)$ satisfy $\|u^n(\cdot,t) - u(\cdot,t)\|_1 \to 0$ as $n \to \infty$ for any $t > 0$. Hence, there is a subsequence $n_k \to \infty$ such that $u^{n_k}(x,t) \to u(x,t)$ a.e. Since each $u^n(x,t)$ is nondecreasing as a function of x, the same conclusion holds true for $u(x,t)$.

In order to show inequality (2.3), we first observe that integrating (2.15) over $\mathbb{R}$ and using the equalities

$$\int_{\mathbb{R}} S_\alpha(t)m\,dx = \int_{\mathbb{R}} m(dx) \quad \text{and} \quad \int_{\mathbb{R}} \partial_x S_\alpha(t-\tau)(u(\tau)u_x(\tau))\,dx = 0,$$

we obtain the identity $\int_{\mathbb{R}} u_x(x,t)\,dx = \int_{\mathbb{R}} m(dx)$, which for nonnegative u_x means

$$\|u_x(t)\|_1 = \|m\| \quad \text{for all} \quad t > 0. \tag{2.18}$$

Now, for fixed $p \in (1,\infty)$ and $u_x \geq 0$, we multiply (2.17) by u_x^{p-1} and integrate the resulting equation over $\mathbb{R}$. After some manipulations involving integrations by parts on the right-hand side, we arrive at

$$\begin{aligned} \frac{1}{p}\frac{d}{dt}\|u_x(t)\|_p^p + \int_{\mathbb{R}} u_x^{p-1}\Lambda^\alpha u_x\,dx &= -\int_{\mathbb{R}} (uu_x)_x u_x^{p-1}\,dx \\ &= -\frac{p-1}{p}\int_{\mathbb{R}} u_x^{p+1}\,dx. \end{aligned} \tag{2.19}$$

Recall now that $\int_{\mathbb{R}} u_x^{p-1}\Lambda^\alpha u_x\,dx \geq 0$ by inequality (2.7). Moreover, it follows from the Hölder inequality combined with (2.18) that

$$\|u_x(t)\|_p^{p^2/(p-1)} \leq \|u_x(t)\|_{p+1}^{p+1}\|m\|^{1/(p-1)}.$$

Applying those two inequalities to (2.19) (note that $u_x \geq 0$) we obtain the following differential inequality for $\|u_x(t)\|_p^p$:

$$\frac{d}{dt}\|u_x(t)\|_p^p \leq -(p-1)\|m\|^{-1/(p-1)}\left(\|u_x(t)\|_p^p\right)^{p/(p-1)}.$$

Integrating it we complete the proof of (2.3) for any $p \in (1,\infty)$.

The case of $p = \infty$ is obtained immediately by passing to the limit $p \to \infty$ in inequality (2.3). $\square$

We conclude this section by recalling some results on smooth approximations of rarefaction waves, namely, the solutions of the following Cauchy problem:

$$w_t - w_{xx} + ww_x = 0, \tag{2.20}$$

$$w(x,0) = w_0(x) = \begin{cases} u_-, & x < 0, \\ u_+, & x > 0. \end{cases} \tag{2.21}$$

LEMMA 2.4. *Let $u_- < u_+$. Problem* (2.20)–(2.21) *has the unique, smooth, global-in-time solution $w(x,t)$ satisfying that*

(i) *$u_- < w(t,x) < u_+$ and $w_x(t,x) > 0$ for all $(x,t) \in \mathbb{R}\times(0,\infty)$;*

(ii) *for every $p \in [1, \infty]$, there exists a constant $C = C(p,u_-,u_+) > 0$ such that*

$$\|w_x(t)\|_p \leq Ct^{-1+1/p}, \quad \|w_{xx}(t)\|_p \leq Ct^{-3/2+1/(2p)}$$

and

$$\|w(t) - w^R(t)\|_p \leq Ct^{-(1-1/p)/2}$$

for all $t > 0$, where $w^R(x,t)$ is the rarefaction wave (1.8).

All results stated in Lemma 2.4 are deduced from the explicit formula for solutions of (2.20)–(2.21) and detailed calculations can be found in [10], with some additional improvements contained in [16, sect. 3].

3. Convergence toward rarefaction waves. For simplicity of the exposition, we split the proof of Theorem 1.1 into a sequence of lemmas.

LEMMA 3.1. *Let $\alpha \in (1,2)$. Assume that u and $\widetilde{u}$ are two solutions of problem* (1.1)–(1.2) *with initial conditions u_0 and $\widetilde{u}_0$, both of the form* (1.4) *with finite signed measures m and $\widetilde{m}$, respectively. Suppose, moreover, that the measure $\widetilde{m}$ of $\widetilde{u}_0$ is nonnegative and $u_0 - \widetilde{u}_0 \in L^1(\mathbb{R})$. Then, for every $p \in [1,\infty]$ there exists a constant $C = C(p) > 0$ such that*

$$\|u(t) - \widetilde{u}(t)\|_p \le Ct^{-(1-1/p)/\alpha}\|u_0 - \widetilde{u}_0\|_1 \tag{3.1}$$

for all $t > 0$.

Proof. In our reasoning, we denote $v(x,t) = u(x,t) - \tilde{u}(x,t)$ which satisfies (2.16). It follows from Theorem 2.2, inequality (2.2), that $\|v(t)\|_1 \le \|v_0\|_1$.

Now, we multiply (2.16) by $|v|^{p-2}v$ with $p > 1$:

$$\frac{1}{p}\frac{d}{dt}\int_{\mathbb{R}} |v|^p\,dx + \int_{\mathbb{R}} (\Lambda^\alpha v)(|v|^{p-2}v)\,dx + \frac{1}{2}\int_{\mathbb{R}} [v^2 + 2v\widetilde{u}]_x |v|^{p-2}v\,dx = 0. \tag{3.2}$$

The third term on the left-hand side of (3.2) is nonnegative by the following calculations:

$$\begin{aligned} \int_{\mathbb{R}} [v^2 + 2v\widetilde{u}]_x |v|^{p-2}v\,dx &= \int_{\mathbb{R}} 2v_x|v|^p\,dx + \int_{\mathbb{R}} 2\widetilde{u}v_x|v|^{p-2}v\,dx + \int_{\mathbb{R}} 2\widetilde{u}_x|v|^p\,dx \\ &= 2\left(1 - \frac{1}{p}\right)\int_{\mathbb{R}} \widetilde{u}_x|v|^p\,dx \ge 0, \end{aligned} \tag{3.3}$$

because $\int_{\mathbb{R}} v_x|v|^p\,dx = 0$ and $\widetilde{u}_x \ge 0$. Hence, using inequality (2.7), we obtain from (3.2)

$$\frac{d}{dt}\int_{\mathbb{R}} |v|^p\,dx + 4\left(1 - \frac{1}{p}\right)\int_{\mathbb{R}} (\Lambda^{\alpha/2}|v|^{p/2})^2\,dx \le 0. \tag{3.4}$$

From now on, we proceed by induction. Applying the Nash inequality (2.6) combined with (2.2), we deduce from (3.4) with $p = 2$ the following differential inequality:

$$\frac{d}{dt}\|v(t)\|_2^2 + 2C_N^{-1}\|v_0\|_1^{-2\alpha}\|v(t)\|_2^{2(1+\alpha)} \le 0,$$

which, after integration, leads to

$$\|v(t)\|_2 \le C_1\|v_0\|_1 t^{-1/(2\alpha)} \quad \text{with} \quad C_1 = (C_N/2\alpha)^{1/(2\alpha)}. \tag{3.5}$$

This is estimate (3.1) with $p = 2$.

Suppose now that

$$\|v(t)\|_{2^n} \le C_n t^{-(1-2^{-n})/\alpha}\|v_0\|_1 \quad \text{for all} \quad t > 0. \tag{3.6}$$

We consider (3.4) with $p = 2^{n+1}$, where the second term is estimated, first, by the Nash inequality (2.6) with $w = |v|^{2^n}$, and next, by the inductive hypothesis (3.6). This two-step estimate leads to the differential inequality

$$\frac{d}{dt}\|v(t)\|_{2^{n+1}}^{2^{n+1}} + 4(1-2^{-n-1})C_N^{-1}(C_n\|v_0\|_1)^{-\alpha 2^{n+1}} t^{2^{n+1}-2}\left(\|v(t)\|_{2^{n+1}}^{2^{n+1}}\right)^{1+\alpha} \le 0.$$

Integrating it we obtain

$$\|v(t)\|_{2^{n+1}} \le C_{n+1} t^{-(1-2^{-n-1})/\alpha}\|v_0\|_1 \quad \text{for all} \quad t>0 \tag{3.7}$$

with

$$C_{n+1} = C_n\left((C_N/(2\alpha))^{1/\alpha}\right)^{2^{-n-1}}\left(2^{n2^{-n-1}}\right)^{1/\alpha}.$$

This is inequality (3.1) for any $p = 2^{n+1}$ with $n \in \mathbb{N}$.

We leave to the reader the proof that $\limsup_{n\to\infty} C_n < \infty$. Hence, passing to the limit $n \to \infty$ in (3.7) we obtain inequality (3.1) for $p = \infty$.

The Hölder inequality

$$\|v\|_p \le \|v\|_{2^n}^{2^{n+1}/p-1}\|v\|_{2^{n+1}}^{2-2^{n+1}/p}$$

completes the proof for every $p \in (2^n, 2^{n+1})$. $\square$

LEMMA 3.2. *Let $\alpha \in (1,2)$. Assume that $w = w(x,t)$ is the smooth approximation of the rarefaction wave, namely, the solution of problem* (2.20)–(2.21). *Then for each $t_0 > 0$ we have*

$$\int_{t_0}^{\infty}\|w_{xx}(t)\|_p\,dt < \infty \quad \textit{for every} \quad p \in (1,\infty]$$

and

$$\int_{t_0}^{t}\|\Lambda^\alpha w(t)\|_p\,dt \le C\log(2+t) \quad \textit{for} \quad p = \frac{3-\alpha}{\alpha-1}$$

for all $t \ge t_0$ and $C > 0$ independent of t.

Proof. It follows from the decay estimates recalled in Lemma 2.4 that

$$\int_{t_0}^{\infty}\|w_{xx}(t)\|_p\,dt \le C\int_{t_0}^{\infty} t^{-3/2+1/(2p)}\,dt < \infty \quad \text{for every} \quad p \in (1,\infty].$$

By the interpolation inequality (2.5) and Lemma 2.4, we obtain

$$\begin{aligned}\|\Lambda^\alpha w(t)\|_p &\le C(1+t)^{(-1+1/p)(2-\alpha)}(1+t)^{(-3/2+1/(2p))(\alpha-1)}\\ &= C(1+t)^{-(1+\alpha)/2+(3-\alpha)/(2p)}.\end{aligned}$$

Hence, the rate of decay on the right-hand side equals -1 for $p = (3-\alpha)/(\alpha-1)$. $\square$

LEMMA 3.3. *Let $\alpha \in (1,2)$. Assume that $u = u(x,t)$ is the solution of* (1.1)–(1.2) *and $w = w(x,t)$ is the solution of* (2.20)–(2.21). *Suppose that $u_0 - w_0 \in L^p(\mathbb{R})$ for $p = (3-\alpha)/(\alpha-1)$. Then*

$$\|u(t) - w(t)\|_p \le C\log(2+t).$$

Proof. Denoting $v = u - w$, we see that this new function satisfies

$$v_t + \Lambda^\alpha v + \frac{1}{2}[v^2 + 2vw]_x = -\Lambda^\alpha w + w_{xx}.$$

We multiply this equation by $|v|^{p-2}v$ and we integrate over $\mathbb{R}$ to obtain

$$\begin{aligned} &\frac{1}{p}\frac{d}{dt}\int |v|^p\, dx + \int (\Lambda^\alpha v)(|v|^{p-2}v)\, dx + \frac{1}{2}\int [v^2 + 2vw]_x |v|^{p-2}v\, dx \\ &= \int (-\Lambda^\alpha w + w_{xx})(|v|^{p-2}v)\, dx. \end{aligned} \tag{3.8}$$

It follows from Lemma 2.3 that $\int_{\mathbb{R}}(\Lambda^\alpha v)(|v|^{p-2}v)\ dx \geq 0$. The third term on the left-hand side of (3.8) is nonnegative by Lemma 2.4 and the same argument as the one used in the proof of Lemma 3.1; cf. identity (3.3). Moreover, using the Hölder inequality, we have

$$\left| \int_{\mathbb{R}} (-\Lambda^\alpha w + w_{xx})(|v|^{p-2}v)\, dx \right| \leq (\|\Lambda^\alpha w\|_p + \|w_{xx}\|_p)\|v\|_p^{p-1}.$$

Consequently, (3.8) implies the following differential inequality:

$$\frac{d}{dt}\|v(t)\|_p^p \leq p\left(\|\Lambda^\alpha w(t)\|_p + \|w_{xx}(t)\|_p\right)\|v(t)\|_p^{p-1},$$

which, after integration, leads to

$$\|v(t)\|_p \leq \|v(t_0)\|_p + \int_{t_0}^t \|\Lambda^\alpha w(\tau)\|_p + \|w_{xx}(\tau)\|_p\, d\tau.$$

The proof is completed by the result stated in Lemma 3.2. □

Now, we are in a position to prove the main result of this paper.

Proof of Theorem 1.1. First, we consider the auxiliary solution $\widetilde{u} = \widetilde{u}(x,t)$ of the fractal Burgers equation (1.1) with the step-like initial condition (1.10). In this case, the measure $\widetilde{m} = (u_+ - u_-)\delta_0$ is nonnegative; hence by Theorem 2.3, $\widetilde{u}_x \geq 0$, and by Lemma 3.1,

$$\|u(t) - \widetilde{u}(t)\|_p \leq Ct^{-(1-1/p)/\alpha}\|u_0 - \widetilde{u}_0\|_1$$

for every $p \in [1,\infty]$ and all $t > 0$.

Next, we compare $\widetilde{u}$ with the smooth approximation of the rarefaction wave that is with the solution $w = w(x,t)$ of (2.20)–(2.21) (note that $\widetilde{u}_0 = w_0$). By Theorem 2.2 and Lemma 2.4, we obtain

$$\|\widetilde{u}_x(t)\|_\infty + \|w_x(t)\|_\infty \leq Ct^{-1}.$$

Moreover, using the Gagliardo–Nirenberg inequality

$$\|v\|_p \leq C\|v_x\|_\infty^a \|v\|_{p_0}^{1-a}, \tag{3.9}$$

valid for any $1 < p_0 < p \leq \infty$ and $a = (1/p_0 - 1/p)/(1 + 1/p_0)$, we have

$$\begin{aligned} \|\widetilde{u}(t) - w(t)\|_p &\leq C(\|\widetilde{u}_x(t)\|_\infty + \|w_x(t)\|_\infty)^a \|\widetilde{u}(t) - w(t)\|_{p_0}^{1-a} \\ &\leq Ct^{-a}\|\widetilde{u}(t) - w(t)\|_{p_0}^{1-a}. \end{aligned}$$

Choosing $p_0 = (3-\alpha)/(\alpha-1)$ (hence $a = [\alpha - 1 - (3-\alpha)/p]/2$), by Lemma 3.3, we conclude that $\|\widetilde{u}(t) - w(t)\|_p \le Ct^{-a}\log(2+t)$ for every $p \in (p_0, \infty]$. Here, we are allowed to use Lemma 3.3 because $\widetilde{u}_0 - w_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R}) \subset L^p(\mathbb{R})$ for every $p \in [1,\infty]$.

Finally, it follows from Lemma 2.4 that the large time asymptotics of $w(t)$ is described in $L^p(\mathbb{R})$ by the rarefaction wave $w^R(x,t)$.

The proof is complete because for $1 < \alpha < 2$ we have $(1-1/p)/\alpha > (1-1/p)/2$. Moreover, since $1 < p_0 < p$, we have $(1-1/p)/2 > (1/p_0 - 1/p)/(1+1/p_0)$. $\square$

4. Additional comments and possible generalizations. Our main result is stated and shown in the simplest case of (1.1); however, several generalizations are possible.

First of all, the operator Λ^α can be replaced by the Lévy operator $\mathcal{L}$ which is a pseudodifferential operator defined by the symbol $a = a(\xi) \ge 0$, $\widehat{\mathcal{L}v}(\xi) = a(\xi)\widehat{v}(\xi)$. Here, the function $e^{-ta(\xi)}$ should be positive definite, so the symbol $a(\xi)$ can be represented by the Lévy–Khintchine formula in the Fourier variables

$$a(\xi) = ib\xi + Q(\xi) + \int_{\mathbb{R}} \left(1 - e^{-i\eta\xi} - i\eta\xi\, \mathbb{1}_{\{|\eta|<1\}}(\eta)\right) \Pi(d\eta). \tag{4.1}$$

Here, $b \in \mathbb{R}$ is fixed, $Q(\xi) = q\xi^2$ with some $q \ge 0$, and Π is a Borel measure such that $\Pi(\{0\}) = 0$ and $\int_{\mathbb{R}} \min(1, |\eta|^2)\,\Pi(d\eta) < \infty$.

Detailed analysis of conservation laws with the anomalous diffusion operator $\mathcal{L}$ is contained in [4, 5, 6]. Here, we would like to emphasize that the fundamental nature of the operator $\mathcal{L}$ is clear from the perspective of the probability theory. It represents the most general form of the generator of a stochastically continuous Markov process with independent and stationary increments. This fact was our basic motivation for the development of the theory presented above.

In order to show the convergence toward rarefaction waves of solutions of conservation laws with the Lévy operator, we need the counterparts of estimates (2.13)–(2.14) of the semigroup of linear operators $e^{-t\mathcal{L}}$ generated by $-\mathcal{L}$. They are valid, e.g., under the assumption that the symbol a of $\mathcal{L}$ has the form

$$a(\xi) = \ell|\xi|^\alpha + k(\xi), \tag{4.2}$$

where $\ell > 0$, $1 < \alpha \le 2$, and k is a symbol of another Lévy operator $\mathcal{K}$ such that

$$\lim_{\xi \to 0} \frac{k(\xi)}{|\xi|^\alpha} = 0. \tag{4.3}$$

The assumptions (4.2) and (4.3) are fulfilled, for example, by *multifractal diffusion operators*

$$\mathcal{L} = -a_0\partial_x^2 + \sum_{j=1}^{k} a_j(-\partial_x^2)^{\alpha_j/2}$$

with $a_0 \ge 0$, $a_j > 0$, $1 < \alpha_j < 2$, and $\alpha = \min_{1\le j\le k} \alpha_j$. We refer the reader to [5, 6] for the reasoning leading to the decay estimates of solutions of nonlinear problems with an operator $\mathcal{L}$ satisfying (4.2)–(4.3). That argument can be directly adapted to obtain counterparts of Theorem 2.2 and Lemma 3.1 with Λ^α replaced by $\mathcal{L}$. Note here that the L^p-L^q estimates of the semigroup $e^{-t\mathcal{L}}$ are equivalent to a certain Nash inequality; see [17, 5] and the references therein.

Our result also holds true if we replace the nonlinear term uu_x in (1.1) by $f(u)_x$ with a strictly convex C^2-function f (as in the paper of Il'in and Oleinik [11]) satisfying $f''(u) \geq \kappa$ for some fixed $\kappa > 0$ and all $u \in \mathbb{R}$. Under this assumption, we immediately generalize Theorem 2.2 and we obtain the decay estimate (2.3).

In order to show the counterpart of Lemma 3.1, we should use the assumption $f''(u) \geq \kappa$ and replace equalities (3.3) by the following (recall that $v = u - \widetilde{u}$):

$$\int_{\mathbb{R}} [f(u) - f(\widetilde{u})]_x |v|^{p-2} v \, dx \geq \kappa \left(1 - \frac{1}{p}\right) \int_{\mathbb{R}} \widetilde{u}_x |v|^p \, dx \geq 0.$$

This argument, however, is known and used systematically, e.g., in [24, inequality (3.5)] (see also [11, 20, 21, 23]); hence we skip the other details.

Acknowledgment. This paper was partially written while the first author was enjoying the hospitality and support of the Institute of Applied Physics and Computational Mathematics in Beijing.

REFERENCES

[1] N. Alibaud, *Entropy formulation for fractal conservation laws*, J. Evol. Equ., 7 (2007), pp. 145–175.

[2] N. Alibaud, J. Droniou, and J. Vovelle, *Occurrence and non-appearance of shocks in fractal Burgers equations*, J. Hyperbolic Differ. Equ., 4 (2007), pp. 479–499.

[3] P. Biler, T. Funaki, and W. Woyczyński, *Fractal Burgers equations*, J. Differential Equations, 148 (1998), pp. 9–46.

[4] P. Biler, G. Karch, and W. Woyczyński, *Asymptotics for multifractal conservation laws*, Studia Math., 135 (1999), pp. 231–252.

[5] P. Biler, G. Karch, and W. A. Woyczyński, *Asymptotics for conservation laws involving Lévy diffusion generators*, Studia Math., 148 (2001), pp. 171–192.

[6] P. Biler, G. Karch, and W. A. Woyczyński, *Critical nonlinearity exponent and self-similar asymptotics for Lévy conservation laws*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 18 (2001), pp. 613–637.

[7] A. Córdoba and D. Córdoba, *A maximum principle applied to quasi-geostrophic equations*, Comm. Math. Phys., 249 (2004), pp. 511–528.

[8] J. Droniou, T. Gallouët, and J. Vovelle, *Global solution and smoothing effect for a non-local regularization of a hyperbolic equation*, J. Evol. Equ., 3 (2002), pp. 499–521.

[9] J. Droniou and C. Imbert, *Fractal first order partial differential equations*, Arch. Ration. Mech. Anal., 182 (2006), pp. 299–331.

[10] Y. Hattori and K. Nishihara, *A note on the stability of the rarefaction wave of the Burgers equation*, Japan J. Indust. Appl. Math., 8 (1991), pp. 85–96.

[11] A. M. Il'in and O. A. Oleinik, *Asymptotic behavior of solutions of the Cauchy problem for some quasi-linear equations for large values of the time*, Mat. Sb., 51 (1960), pp. 191–216.

[12] C. Imbert and G. Karch, work in progress.

[13] B. Jourdain, S. Méléard, and W. Woyczyński, *A probabilistic approach for nonlinear equations involving the fractional Laplacian and singular operator*, Potential Anal., 23 (2005), pp. 55–81.

[14] B. Jourdain, S. Méléard, and W. Woyczyński, *Probabilistic approximation and inviscid limits for one-dimensional fractional conservation laws*, Bernoulli, 11 (2005), pp. 689–714.

[15] N. Ju, *The maximum principle and the global attractor for the dissipative 2D quasi-geostrophic equations*, Comm. Math. Phys., 255 (2005), pp. 161–181.

[16] S. Kawashima and Y. Tanaka, *Stability of rarefaction waves for a model system of a radiating gas*, Kyushu J. Math., 58 (2004), pp. 211–250.

[17] V. A. Liskevich and Yu. A. Semenov, *Some problems on Markov semigroups*, in Schrödinger Operators, Markov Semigroups, Wavelet Analysis, Operator Algebras, Math. Top. 11, Akademie Verlag, Berlin, 1996, pp. 163–217.

[18] T.-P. Liu, A. Matsumura, and K. Nishihara, *Behaviors of solutions for the Burgers equation with boundary corresponding to rarefaction waves*, SIAM J. Math. Anal., 29 (1998), pp. 293–308.

[19] J. A. Mann and W. A. Woyczyński, *Growing fractal interfaces in the presence of self-similar hopping surface diffusion*, Phys. A, 291 (2001), pp. 159–183.

[20] A. Matsumura and K. Nishihara, *Asymptotics toward the rarefaction waves of the solutions of a one-dimensional model system for compressible viscous gas*, Japan J. Appl. Math., 3 (1986), pp. 1–13.

[21] A. Matsumura and K. Nishihara, *Global stability of the rarefaction wave of a one-dimensional model system for compressible viscous gas*, Comm. Math. Phys., 144 (1992), pp. 325–335.

[22] C. Miao, B. Yuan, and B. Zhang, *Well-posedness of the Cauchy problem for fractional power dissipative equations*, Nonlinear Anal., 68 (2008), pp. 461–484.

[23] T. Nakamura, *Asymptotic decay toward the rarefaction waves of solutions for viscous conservation laws in a one-dimensional half space*, SIAM J. Math. Anal., 34 (2003), pp. 1308–1317.

[24] M. Nishikawa and K. Nishihara, *Asymptotics toward the planar rarefaction wave for viscous conservation law in two space dimensions*, Trans. Amer. Math. Soc., 352 (1999), pp. 1203–1215.

SIAM J. MATH. ANAL.
Vol. 39, No. 5, pp. 1550–1576

Γ-CONVERGENCE OF POWER-LAW FUNCTIONALS, VARIATIONAL PRINCIPLES IN L^∞, AND APPLICATIONS*

MARIAN BOCEA† AND VINCENZO NESI‡

Abstract. Two Γ-convergence results for a general class of power-law functionals are obtained in the setting of $\mathcal{A}$-quasiconvexity. New variational principles in L^∞ are introduced, allowing for the description of the yield set in the context of a simplified model of polycrystal plasticity. A number of highly degenerate nonlinear partial differential equations arise as Aronsson equations associated with these variational principles.

Key words. $\mathcal{A}$-quasiconvexity, Aronsson equations, Γ-convergence, Euler–Lagrange equations, lower semicontinuity, dielectric breakdown, polycrystal plasticity, strength set, yield set, yield surface

AMS subject classifications. 35F99, 35J70, 49K20, 49S05, 74C05

DOI. 10.1137/060672388

1. Introduction.

One of the main motivations of this work is the development of a general theory that allows for the characterization of the yield set in the context of *polycrystal plasticity* by means of suitable variational principles in L^∞. From the mathematical point of view the results of the present paper cover two related but simpler problems pertaining to highly nonlinear conducting media. These can be seen as simplified versions of polycrystal plasticity in the context of antiplane shear (see [37]). We will now describe the physical motivations in the more familiar framework of polycrystal plasticity.

A *polycrystal* is a collection of single crystals (*grains*) bonded together in different orientations; the microstructure of the polycrystal, which consists of the shapes and orientations of the grains, is called the *texture* of the polycrystal. In addition to the material properties of the grains, the behavior of polycrystalline solids is highly influenced by their texture. Understanding how the behavior of a polycrystal depends on its texture is crucial from the technological point of view in that it provides important guidelines for specific material processing.

The *yield* of a single crystalline material is described by a closed convex subset K of the space of stresses $\mathbb{M}^{3\times3}_{\mathrm{sym}}$ (symmetric 3×3 real matrices); any stress field $\sigma : \Omega \to \mathbb{M}^{3\times3}_{\mathrm{sym}}$ in the crystal occupying the region $\Omega \subset \mathbb{R}^3$ is subject to the pointwise constraint

$$\sigma(x) \in K,\ x \in \Omega.$$

In the context of rigid perfectly plastic crystals the set K is called the *yield set*. Rigid perfectly plastic behavior is characterized by the fact that the crystal can only withstand stresses in the interior of the set K; i.e., the material does not exhibit plastic deformation when subject to a stress σ such that $\sigma(x) \in \mathrm{int}(K)$, $x \in \Omega$, but,

*Received by the editors October 15, 2006; accepted for publication (in revised form) July 26, 2007; published electronically January 22, 2008.
http://www.siam.org/journals/sima/39-5/67238.html
†Department of Mathematics, North Dakota State University, 300 Minard Hall, Fargo, ND 58105-5075 (marian.bocea@ndsu.edu).
‡Dipartimento di Matematica, Università di Roma, “La Sapienza,” Piazzale A. Moro 2, 00185 Rome, Italy (nesi@mat.uniroma1.it). This author was partially supported by MIUR, PRIN 2006: “Omogeneizzazione e metodi variazionali in matematica applicata.”

on the other hand, the material will deform at a certain strain rate when $\sigma(x) \in \partial K$ for some $x \in \Omega$. The boundary ∂K of the yield set is called the *yield surface*.

The texture of a polycrystal is described by a piecewise constant rotation-valued function $R : \Omega \to \mathrm{SO}(3)$, where $R(x)$ is constant in each grain and indicates the orientation of the grain which contains the point $x \in \Omega$. If the yield set of the reference single crystal is K, the stress in the polycrystal occupying the region $\Omega \subset \mathbb{R}^3$ must satisfy the constraint

$$\sigma(x) \in R(x)KR^T(x), \; x \in \Omega. \tag{1.1}$$

The effective behavior of the polycrystal is then described by the constraints induced on the average stress $\overline{\sigma} := \fint_\Omega \sigma(x)dx$; the set of all average stresses $\overline{\sigma}$ when σ satisfies the pointwise constraint (1.1) and the equilibrium equation,

$$\operatorname{div} \sigma = 0 \text{ in } \Omega, \tag{1.2}$$

is called the *yield set of the polycrystal*. Precisely,

$$K_{\text{eff}} := \left\{ \overline{\sigma} := \fint_\Omega \sigma(x)dx \; : \; (1.1) \text{ and } (1.2) \text{ hold} \right\}.$$

Yield in a crystalline solid is usually associated with a finite number of *slip systems* which depend on the atomic lattice, each determined by a pair (n_k, m_k) of orthogonal vectors, where n_k is the normal to the *slip plane* and m_k is the *direction of slip*. In this case, we have

$$K = \left\{ A \in \mathbb{M}^{3\times3}_{\text{sym}} \; : \; \langle A, \mu_k \rangle \le \tau_k^{\text{critical}}, k = 1, \ldots, s \right\},$$

where s stands for the number of slip systems, τ_k^{critical} is the *critical shear stress* for the kth slip system, and

$$\mu_k := \frac{1}{2}\left(m_k \otimes n_k + n_k \otimes m_k\right)$$

is the kth *slip tensor*. The goal of polycrystal plasticity is to describe the yield set K_{eff}, given K and some information on the texture of the polycrystal. The complexity of the problem has led to the introduction of alternative schemes designed to estimate the macroscopic response of polycrystals without solving the equilibrium equation (1.2) directly. The estimate introduced by Sachs [42] assumes that the stress is constant throughout the polycrystal, leading to

$$K_{\text{Sachs}} := \left\{ \sigma \in \mathbb{M}^{3\times3}_{\text{sym}} \; : \; (1.1) \text{ holds} \right\}.$$

On the other hand, the estimate introduced by Taylor [49] assumes that the strain rate is constant throughout the polycrystal, giving

$$K_{\text{Taylor}} := \left\{ \overline{\sigma} := \fint_\Omega \sigma(x)dx \; : \; (1.1) \text{ holds} \right\}.$$

Since

$$K_{\text{Sachs}} \subseteq K_{\text{eff}} \subseteq K_{\text{Taylor}},$$

it follows that the Sachs and Taylor estimates lead to bounds on the yield set K_{eff}. This was first recognized by Bishop and Hill [11]. In much more recent developments, Kohn and Little [37] studied in detail a simpler model working in particular in the context of antiplane shear. They showed that in two dimensions the Sachs bound is optimal and gave a better estimate for the Taylor bound. Goldsztein [30] improved the results in [37] by proposing a better estimate for K_{Taylor} having the same scaling law. In [31] he showed that this scaling law is essentially optimal. The corresponding problems in three dimensions for model problems involving gradients and divergence-free fields were addressed by Garroni and Kohn [27]. In both settings they found that the Sachs bound is optimal; in the case of gradients they provided an improved estimate for the Taylor bound which scales differently for certain reference crystals, while in the divergence-free case they proved that the Taylor bound has the optimal scaling law.

From the mathematical standpoint, polycrystal plasticity has been justified for many years by means of the dual of a $W^{1,1}$ problem. This formulation has been studied in a classical series of papers by Bouchittè and Suquet. We refer the reader to [12] for the original work and to [37] for a minimal overview. Much more recently, the study of these problems has been justified in a different way. The first paper in this direction is due to Garroni, Nesi, and Ponsiglione [28], where the authors considered first failure dielectric breakdown. Mathematically, they treated a case in which the variable is a gradient field. The main contribution is to show that the so-called *strength* or *yield* set can be characterized via Γ-convergence in an alternative way. Here we address a similar but mathematically distinct problem, namely, the characterization of effective yield sets defined in terms of divergence-free fields. We achieve this goal in the context of electrical resistivity in section 6 of the paper. We will comment later on the differences between this alternative point of view and the classical one; we point out that both the new and the classical derivations give rise to the same effective yield set, but the new derivation preserves some memory of the approximation that has been used to derive it. We refer the reader to sections 5 and 6 of the paper. For example, in the "gradient" case, treated in [28], one is naturally led to an interesting connection with the study of some generalizations of the infinity-Laplace equation

$$\Delta_\infty u := \sum_{i,j=1}^{N} \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_j} \frac{\partial^2 u}{\partial x_i \partial x_j} = 0. \tag{1.3}$$

This equation has been introduced by Aronsson in his influential work in connection with the problem of minimal extensions of Lipschitz maps [2], [3], [4], [5], [6], and its study has received a great deal of attention in the last decade. An excellent account on this topic is given in the survey paper by Aronsson, Crandall, and Juutinen [7]. We also refer the reader to the recent work of Evans and Yu [24] for a review of some basic properties of the solution to (1.3), as well as for some new interesting results.

Let us go back to the main contribution of the present paper. It has long been recognized that in the mathematical study of materials the key question that one must understand is the interaction between *linear* partial differential equations (the balance laws) and *nonlinear* pointwise constraints (the constitutive laws). This is the underlying motivation for the general theory of compensated compactness as developed by Murat and Tartar (see, e.g., [39], [47], [48]). In this paper we generalize the Γ-convergence results of [28] to more general linear PDE constraints on the underlying fields, in the framework of $\mathcal{A}$-quasiconvexity (see [19], [25]). This allows us to consider situations which have not been studied so far, most notably the case of solenoidal fields (divergence-free fields), which is relevant to treating extreme resistivity. Actual plasticity would require an extra constraint, namely, one should consider divergence-free matrix fields which are symmetric. This further generalization does not seem to be a serious obstacle to the analysis, but it would require a rather long digression so we will address this issue elsewhere.

The paper is organized as follows. In section 2 we give the definition of $\mathcal{A}$-quasiconvexity, state a lower semicontinuity result of Fonseca and Müller [25], and recall the definition of De Giorgi's Γ-convergence. In section 3 we prove two Γ-convergence results for power-law functionals in the general setting of $\mathcal{A}$-quasiconvexity. Section 4 is devoted to specializing the abstract results of the previous section to two different special cases. The first one deals with curl-free fields, where we generalize results in [28] by making milder assumptions on the functionals under consideration. We focus on power-law-type materials and discuss the meaning of our results in the context of effective conductivity. In the second part of section 4 we specialize the Γ-convergence results of section 3 to the case where the underlying fields are divergence free. This leads to a new derivation, whose meaning is discussed in the context of effective resistivity. We conclude this section by briefly recalling the duality behind effective conductivity and effective resistivity, and we explain the connection with the literature treating bounds for effective properties of nonlinear composites in the context of constitutive laws of power-law type (see [28], [40], [41], [44], [45], [46]). Section 5 treats dielectric breakdown as a particular instance of our general results. We define the effective strength domain for electrically conducting media and explain the relation with the new variational principle and the associated Aronsson equation, leading to the study of the infinity-Laplace equation. In section 6 we define the effective strength set K_{eff} (which one may call yield set, with a slight abuse of language) in the context of electrical resistivity, we propose a new variational principle by means of a minimization problem in L^∞, and we provide the first characterization of K_{eff} in terms of this new object. We end the first part of section 6 by explaining some of the issues that remain to be addressed via homogenization. The new variational principle in L^∞ leads naturally to interesting new systems of partial differential equations which arise as the Aronsson equations associated with the supremal functionals under consideration. The formal derivation of these systems, as well as several remarks, are left for the second part of section 6.

2. Preliminaries. Let N, d, l be positive integers, $1 \leq p \leq +\infty$, and consider a family $A^{(1)}, A^{(2)}, \dots, A^{(N)} \in \operatorname{Lin}(\mathbb{R}^d; \mathbb{R}^l)$. We define $\mathcal{A} : L^p(\Omega; \mathbb{R}^d) \to W^{-1,p}(\Omega; \mathbb{R}^l)$ by

$$\mathcal{A}v := \sum_{i=1}^{N} A^{(i)} \frac{\partial v}{\partial x_i}, \tag{2.1}$$

where Ω is an open, bounded domain in $\mathbb{R}^N$. Note that each $A^{(i)}$ is represented by an $l \times d$ matrix, and $\frac{\partial v}{\partial x_i}(x)$ is represented by a d-dimensional column vector, so that each term $A^{(i)} \frac{\partial v}{\partial x_i}(x)$ is an l-dimensional column vector. Consider the operator $\mathbb{A} : \mathbb{R}^N \to \operatorname{Lin}(\mathbb{R}^d; \mathbb{R}^l)$, defined as a linear combination of the given family $A^{(1)}, A^{(2)}, \dots, A^{(N)}$ with real coefficients. Precisely, we set

$$\mathbb{A}(w) := \sum_{i=1}^{N} w_i A^{(i)},$$

where $w = (w_1, \ldots, w_N) \in \mathbb{R}^N$. In what follows we will assume that $\mathcal{A}$ satisfies the *constant rank property* (see [39]):

$$\text{there exists } r \in \mathbb{N} \text{ such that rank } \mathbb{A}(w) = r \text{ for all } w \in S^{N-1}. \tag{2.2}$$

The constant rank property plays a central role in the theory of compensated compactness developed by Murat [39] and Tartar [47], [48]. Situations where (2.2) fails are much less understood (see, e.g., [38]). In what follows, and throughout the paper, $Q := (0,1)^N$ is the unit cube in $\mathbb{R}^N$.

DEFINITION 2.1. *A function $g : \mathbb{R}^d \to \mathbb{R}$ is said to be $\mathcal{A}$-quasiconvex if*

$$g(A) \leq \int_Q g(A + w(x))dx$$

for all $A \in \mathbb{R}^d$ and all Q-periodic $w \in C^\infty(Q; \mathbb{R}^d)$ such that $\mathcal{A}(w) = 0$ and $\int_Q w(x)dx = 0$.

The notion of $\mathcal{A}$-quasiconvexity (without the assumption of periodicity of the test functions) has been first investigated by Dacorogna [19]. Under the assumption of constant rank property (2.2) of the operator $\mathcal{A}$, Fonseca and Müller have shown in [25] that if $\Omega \subset \mathbb{R}^N$ is an open, bounded set, $(u, v) : \Omega \to \mathbb{R}^m \times \mathbb{R}^d$ is measurable, and $g : \Omega \times \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}$ is a normal integrand, then the $\mathcal{A}$-quasiconvexity of $g(x, u, \cdot)$ is a necessary and sufficient condition for the sequential lower semicontinuity of integrals of the form

$$(u, v) \mapsto \int_\Omega g(x, u(x), v(x))dx$$

along sequences that satisfy $u_n \to u$ in measure, $v_n \to v$ in L^p, and $\mathcal{A}v_n \to 0$ in $W^{-1,p}$. We will explicitly need only the sufficiency part of their result for $1 \leq p < +\infty$. Precisely, we have the following proposition.

PROPOSITION 2.2 (see [25, Theorem 3.7]). *Let $1 \leq p < +\infty$, and suppose that $g : \Omega \times \mathbb{R}^m \times \mathbb{R}^d \to [0, +\infty)$ is a normal integrand such that $z \mapsto g(x, u, z)$ is $\mathcal{A}$-quasiconvex and continuous for $\mathcal{L}^N$-a.e. $x \in \Omega$ and all $u \in \mathbb{R}^m$. Assume further that there exists a locally bounded function $a : \Omega \times \mathbb{R}^m \to [0, +\infty)$ such that*

$$0 \leq g(x, u, v) \leq a(x, u)(1 + |v|^p)$$

for $\mathcal{L}^N$-a.e. $x \in \Omega$ and all $(u, v) \in \mathbb{R}^m \times \mathbb{R}^d$. If

$$u_n \to u \textit{ in measure},$$

$$v_n \rightharpoonup v \textit{ in } L^p(\Omega; \mathbb{R}^d), \tag{2.3}$$

and

$$\mathcal{A}v_n \to 0 \textit{ in } W^{-1,p}(\Omega; \mathbb{R}^l), \tag{2.4}$$

then

$$\int_\Omega g(x, u(x), v(x))dx \leq \liminf_{n\to\infty} \int_\Omega g(x, u_n(x), v_n(x))dx. \tag{2.5}$$

We remark that if $p = +\infty$, then (2.5) still holds provided that in (2.3) the weak convergence of v_n to v in $L^p(\Omega; \mathbb{R}^d)$ is replaced by the weak* convergence in

$L^\infty(\Omega;\mathbb{R}^d)$, and in (2.4) $\mathcal{A}v_n \to 0$ in $W^{-1,p}(\Omega;\mathbb{R}^l)$ is replaced by $\mathcal{A}v_n \equiv 0$. In this case the growth condition on g is not needed (see [25]). We end this section by recalling the definition of De Giorgi's Γ-convergence (see [21], [22]) in metric spaces. For a comprehensive introduction to the subject we refer the reader to the monograph by Dal Maso [20]. See also [13].

DEFINITION 2.3. *Let X be a metric space. A sequence $\{I_p\}$ of functionals $I_p : X \to \overline{\mathbb{R}} := \mathbb{R}\cup\{+\infty\}$ is said to $\Gamma(X)$-converge to $I : X \to \overline{\mathbb{R}}$ (we write $\Gamma(X) - \lim_{p\to\infty} I_p = I$) if the following hold:*

(i) *for every $u \in X$ and $\{u_p\} \subset X$ such that $u_p \to u$ in X, we have*

$$I(u) \leq \liminf_{p\to\infty} I_p(u_p);$$

(ii) *for every $u \in X$ there exists a sequence $\{u_p\} \subset X$ such that $u_p \to u$ in X and*

$$I(u) = \lim_{p\to\infty} I_p(u_p).$$

The sequence $\{u_p\}$ in (ii) *is called a* recovery sequence *for the Γ-limit.*

3. Γ-convergence results in a general setting. In this section we prove two Γ-convergence results. The first one should be seen as a justification of a simplified plasticity model by a nonclassical route. Let Ω be an open, bounded domain in $\mathbb{R}^N (N \geq 1)$, and let $f : \Omega \times \mathbb{R}^d \to [0,+\infty)$ be a Carathéodory integrand satisfying the following hypotheses:

$$f(x,\cdot) \text{ is } \mathcal{A}\text{-quasiconvex for } \mathcal{L}^N\text{-a.e. } x \in \Omega; \tag{3.1}$$

there exist constants $c, C > 0$ such that

$$c|v| \leq f(x,v) \leq C\,(1+|v|) \text{ for } \mathcal{L}^N\text{-a.e. } x \in \Omega \text{ and all } v \in \mathbb{R}^d. \tag{3.2}$$

Let $I_p : L^1(\Omega;\mathbb{R}^d) \to [0,+\infty]$ be defined by

$$I_p(w) := \begin{cases} \dfrac{1}{p}\displaystyle\int_\Omega f(x,w(x))^p dx & \text{if } w \in L^p(\Omega;\mathbb{R}^d)\cap \ker\mathcal{A}, \\ +\infty & \text{otherwise.} \end{cases}$$

THEOREM 3.1. *Define $I_\infty : L^1(\Omega;\mathbb{R}^d) \to [0,+\infty]$ by*

$$I_\infty(w) := \begin{cases} 0 & \text{if } f(x,w(x)) \leq 1 \text{ for } \mathcal{L}^N\text{-a.e. } x\in\Omega,\ \mathcal{A}w = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Then

(i) *for every $w \in L^1(\Omega;\mathbb{R}^d)$ and $\{w_p\} \subset L^1(\Omega;\mathbb{R}^d)$ such that $w_p \rightharpoonup w$ weakly in $L^1(\Omega;\mathbb{R}^d)$, we have*

$$I_\infty(w) \leq \liminf_{p\to\infty} I_p(w_p); \tag{3.3}$$

(ii) *for every $w \in L^1(\Omega;\mathbb{R}^d)$, there exists a sequence $\{w_p\} \subset L^1(\Omega;\mathbb{R}^d)$ such that $w_p \to w$ in $L^1(\Omega;\mathbb{R}^d)$, and*

$$\limsup_{p\to\infty} I_p(w_p) \leq I_\infty(w). \tag{3.4}$$

In particular,

$$\Gamma(L^1(\Omega;\mathbb{R}^d)) - \lim_{p\to\infty} I_p = I_\infty.$$

Proof. Let $\{w_p\} \subset L^1(\Omega;\mathbb{R}^d)$ be such that $w_p \rightharpoonup w$ weakly in $L^1(\Omega;\mathbb{R}^d)$. We will show that (3.3) holds. In view of the coercivity condition in (3.2), we may assume, without loss of generality, that

$$w_p \in L^p(\Omega;\mathbb{R}^d),\ \ \mathcal{A}w_p = 0, \tag{3.5}$$

and

$$\liminf_{p\to\infty} I_p(w_p) = \lim_{p\to\infty} I_p(w_p) < +\infty. \tag{3.6}$$

Let $x \in \Omega$ be a Lebesgue point for $f(\cdot, w(\cdot)) \in L^1(\Omega)$. For any ball $B(x,r) \subset \Omega$ and for p sufficiently large, we have

$$\begin{aligned}\int_{B(x,r)} f(y, w_p(y))dy &\le \left(\int_\Omega (f(y,w_p(y)))^p\, dy\right)^{1/p} \left(\mathcal{L}^N(B(x,r))\right)^{(p-1)/p}\\ &= (I_p(w_p))^{1/p}\, p^{1/p} \left(\mathcal{L}^N(B(x,r))\right)^{(p-1)/p},\end{aligned}$$

where we have used Hölder's inequality. Letting $p \to \infty$, we obtain that

$$\limsup_{p\to\infty} \int_{B(x,r)} f(y, w_p(y))dy \le \mathcal{L}^N(B(x,r)). \tag{3.7}$$

Taking (3.1), (3.2), and (3.5) into account, we may apply Proposition 2.2 and we deduce that

$$\int_{B(x,r)} f(y, w(y))dy \le \liminf_{p\to\infty} \int_{B(x,r)} f(y, w_p(y))dy.$$

Thus, in view of (3.7),

$$\frac{1}{\mathcal{L}^N(B(x,r))} \int_{B(x,r)} f(y, w(y))dy \le 1.$$

Since $\mathcal{L}^N$-a.e. $x \in \Omega$ is a Lebesgue point for $f(\cdot, w(\cdot))$, passing to the limit $r \to 0^+$ in the last equation yields

$$f(x, w(x)) \le 1 \quad \text{for } \mathcal{L}^N\text{-a.e. } x \in \Omega.$$

It follows that $I_\infty(w) = 0$, and this implies that (3.3) holds. To complete the proof, we need to show that for any $w \in L^1(\Omega;\mathbb{R}^d)$, there exists a recovery sequence for the Γ-limit, that is, a sequence $\{w_p\} \subset L^1(\Omega;\mathbb{R}^d)$ with $w_p \to w$ in $L^1(\Omega;\mathbb{R}^d)$ and such that (3.4) holds.

To this aim, assume, without loss of generality, that $I_\infty(w) = 0$. Thus, $f(x, w(x)) \le 1$ for $\mathcal{L}^N$-a.e. $x \in \Omega$ and $\mathcal{A}w = 0$. By (3.2), $w \in L^\infty(\Omega;\mathbb{R}^d)$, and thus we may choose $\{w_p\} = \{w\}$ as a recovery sequence. Indeed,

$$I_p(w_p) = \frac{1}{p}\int_\Omega f(x, w(x))^p dx \le \frac{1}{p}\mathcal{L}^N(\Omega),$$

which implies that $\lim_{p\to\infty} I_p(w_p) = 0 = I_\infty(w)$. We conclude that (3.4) holds. ☐

We will now describe an alternative derivation of the model. The goal is to obtain a limiting functional which is less degenerate than I_∞. We follow [28], where a similar approach was used in the gradient case.

Define $J_p : L^1(\Omega;\mathbb{R}^d) \to [0,+\infty]$ by

$$J_p(w) := \begin{cases} \left(\displaystyle\int_\Omega f(x,w(x))^p dx\right)^{1/p} & \text{if } w \in L^p(\Omega;\mathbb{R}^d) \cap \ker\mathcal{A}, \\ +\infty & \text{otherwise.} \end{cases}$$

THEOREM 3.2. *Let $J_\infty : L^1(\Omega;\mathbb{R}^d) \to [0,+\infty]$ be defined by*

$$J_\infty(w) := \begin{cases} \operatorname{ess\,sup}_{x\in\Omega} f(x,w(x)) & \textit{if } w \in L^\infty(\Omega;\mathbb{R}^d) \cap \ker\mathcal{A}, \\ +\infty & \textit{otherwise.} \end{cases}$$

Then

(i) *for every $w \in L^1(\Omega;\mathbb{R}^d)$ and $\{w_p\} \subset L^1(\Omega;\mathbb{R}^d)$ such that $w_p \rightharpoonup w$ weakly in $L^1(\Omega;\mathbb{R}^d)$, we have*

$$J_\infty(w) \le \liminf_{p\to\infty} J_p(w_p); \tag{3.8}$$

(ii) *for every $w \in L^1(\Omega;\mathbb{R}^d)$, there exists a sequence $\{w_p\} \subset L^1(\Omega;\mathbb{R}^d)$ such that $w_p \to w$ in $L^1(\Omega;\mathbb{R}^d)$ and*

$$\limsup_{p\to\infty} J_p(w_p) \le J_\infty(w). \tag{3.9}$$

In particular,

$$\Gamma(L^1(\Omega;\mathbb{R}^d)) - \lim_{p\to\infty} J_p = J_\infty.$$

Proof. Let $\{w_p\} \subset L^1(\Omega;\mathbb{R}^d)$ be such that $w_p \rightharpoonup w$ weakly in $L^1(\Omega;\mathbb{R}^d)$. We need to show that (3.8) holds. Extract a subsequence (not relabeled) such that

$$w_p \in L^p(\Omega;\mathbb{R}^d),\ \mathcal{A}w_p = 0, \tag{3.10}$$

and

$$\liminf_{p\to\infty} J_p(w_p) = \lim_{p\to\infty} J_p(w_p) < +\infty. \tag{3.11}$$

In view of (3.1) and Jensen's inequality, for any $q \ge 1$, $f(x,\cdot)^q$ is $\mathcal{A}$-quasiconvex for $\mathcal{L}^N$-a.e. $x \in \Omega$. In addition, by (3.2),

$$c^q|v|^q \le f(x,v)^q \le 2^{q-1}C^q\,(1+|v|^q) \tag{3.12}$$

for $\mathcal{L}^N$-a.e. $x \in \Omega$ and all $v \in \mathbb{R}^d$. For any $p > q > 1$, we have

$$\|w_p\|_{L^q(\Omega;\mathbb{R}^d)} \le C J_p(w_p),$$

where $C > 0$ is a constant which depends only on $\mathcal{L}^N(\Omega)$. Thus, by (3.11), $\{w_p\}$ is bounded in $L^q(\Omega;\mathbb{R}^d)$. Up to a subsequence (not relabeled) $w_p \rightharpoonup w$ weakly in

$L^q(\Omega;\mathbb{R}^d)$, as $p \to \infty$. Taking (3.10) and (3.12) into account and in view of the $\mathcal{A}$-quasiconvexity and continuity of $f(x,\cdot)^q$ for $\mathcal{L}^N$-a.e. $x \in \Omega$, we are again in position to apply Proposition 2.2. We obtain

$$\int_\Omega f(x,w(x))^q dx \le \liminf_{p\to\infty} \int_\Omega f(x,w_p(x))^q dx. \tag{3.13}$$

On the other hand, by means of elementary arguments, we have that

$$\left(\liminf_{p\to\infty} \int_\Omega f(x,w_p(x))^q dx\right)^{1/q} \le \limsup_{p\to\infty} \left(\int_\Omega f(x,w_p(x))^q dx\right)^{1/q}. \tag{3.14}$$

In view of (3.13) and (3.14), we deduce that

$$J_q(w) \le \limsup_{p\to\infty} J_q(w_p) \le \mathcal{L}^N(\Omega)^{\frac{1}{q}} \lim_{p\to\infty} J_p(w_p) \tag{3.15}$$

for all $q \ge 1$, where the last inequality follows from the fact that

$$J_q(w_p) \le \mathcal{L}^N(\Omega)^{\frac{1}{q}-\frac{1}{p}} J_p(w_p).$$

By a localization argument similar to the one used in the proof of Theorem 3.1 it can be shown that $f(\cdot,w(\cdot)) \in L^\infty(\Omega)$. It follows that $J_q(w) \to \|f(\cdot,w(\cdot))\|_{L^\infty(\Omega)}$ as $q \to \infty$. Thus, sending $q \to \infty$ in (3.15) we obtain that (3.8) holds.

It remains to prove (3.9). Let $w \in L^1(\Omega;\mathbb{R}^d)$, and consider the sequence $\{w_p\} \subset L^1(\Omega;\mathbb{R}^d)$, where $w_p := w$ for all $p \in \mathbb{N}$. To verify that (3.9) holds we assume, without loss of generality, that $w \in L^\infty(\Omega;\mathbb{R}^d)$ and $\mathcal{A}w = 0$, so that $J_\infty(w) < +\infty$. We have

$$\limsup_{p\to\infty} J_p(w_p) = \limsup_{p\to\infty} \|f(\cdot,w(\cdot))\|_{L^p(\Omega)} = J_\infty(w),$$

which concludes the proof. ∎

4. Effective conductivity and resistivity. In this section we specialize the previous results to cover effective conductivity and effective resistivity for the important class of power-law materials. In this context, the use of convex duality has been exploited by many authors both from the mathematical point of view (see, e.g., [23] and [26]) and from the point of view of applications to materials science, more specifically to the issue of bounding effective behavior in this nonlinear setting (see [44], [45], [46]).

4.1. Conductivity: The curl-free case. In this section we consider the situation where the differential constraint $\mathcal{A}w = 0$ reduces to requiring $w : \Omega \subset \mathbb{R}^N \to \mathbb{R}^N$ to satisfy curl $w = 0$, i.e.,

$$\frac{\partial w_k}{\partial x_i} - \frac{\partial w_i}{\partial x_k} = 0,\ 1 \le i,k \le N.$$

In this case $d = N$, and we may rewrite these partial differential equations as $\mathcal{A}w = 0$, where

$$\mathcal{A}w := \sum_{r=1}^N A^{(r)} \frac{\partial w}{\partial x_r}$$

with $A^{(r)} \in \mathrm{Lin}(\mathbb{R}^N, \mathbb{R}^{N^2})$ given by

$$A^{(r)}_{(ik),p} := \delta_{ri}\delta_{pk} - \delta_{rk}\delta_{pi},\ 1 \le i,k,p,r \le N.$$

It can be shown that $\mathcal{A}$ satisfies the constant rank property (see (2.2) and [25]). In fact, for every $w \in S^{N-1}$, $\ker(\mathbb{A}(w)) = \mathrm{span}\{w\}$, and thus we have that $\dim(\ker(\mathbb{A}(w))) = 1$.

Let $f : \Omega \times \mathbb{R}^N \to [0, +\infty)$ be defined by $f(x, v) := \langle \Sigma(x)v, v\rangle^{\frac{1}{2}}$, where $\Sigma \in L^\infty\left(\Omega; \mathbb{M}^{N\times N}_{\mathrm{sym}}\right)$ satisfies the ellipticity bounds

$$\frac{1}{K}|\xi|^2 \le \langle \Sigma(x)\xi, \xi\rangle \le K|\xi|^2 \tag{4.1}$$

for $\mathcal{L}^N$-a.e. $x \in \Omega$ and all $\xi \in \mathbb{R}^N$, where $K \ge 1$ is a real constant. Note that f satisfies our hypotheses (3.1) and (3.2). With these particular choices one can recover the results of Garroni, Nesi, and Ponsiglione [28] as corollaries of Theorems 3.1 and 3.2.

Corollary 4.1 (see [28, Proposition 2.1]). *Let $G_p : L^1(\Omega; \mathbb{R}^d) \to [0, +\infty]$ be defined by*

$$G_p(u) := \begin{cases} \dfrac{1}{p}\displaystyle\int_\Omega \langle \Sigma(x)\nabla u(x), \nabla u(x)\rangle^{\frac{p}{2}}\,dx & \text{if } u \in W^{1,p}(\Omega), \\ +\infty & \text{otherwise.} \end{cases}$$

Then

$$\Gamma(L^1(\Omega; \mathbb{R}^d)) - \lim_{p\to\infty} G_p = G_\infty,$$

where $G_\infty : L^1(\Omega; \mathbb{R}^d) \to [0, +\infty]$ is given by

$$G_\infty(u) := \begin{cases} 0 & \text{if } \left|\Sigma^{\frac{1}{2}}(x)\nabla u(x)\right| \le 1 \text{ for } \mathcal{L}^N\text{-a.e. } x \in \Omega, \\ +\infty & \text{otherwise.} \end{cases}$$

Corollary 4.2 (see [28, Proposition 2.6]). *Define $F_p : L^1(\Omega; \mathbb{R}^d) \to [0, +\infty]$ by*

$$F_p(u) := \begin{cases} \left(\displaystyle\int_\Omega \langle \Sigma(x)\nabla u(x), \nabla u(x)\rangle^{\frac{p}{2}}\,dx\right)^{1/p} & \text{if } u \in W^{1,p}(\Omega), \\ +\infty & \text{otherwise.} \end{cases}$$

Then

$$\Gamma(L^1(\Omega; \mathbb{R}^d)) - \lim_{p\to\infty} F_p = F_\infty,$$

where $F_\infty : L^1(\Omega; \mathbb{R}^d) \to [0, +\infty]$ is given by

$$F_\infty(u) := \begin{cases} \left\|\Sigma^{\frac{1}{2}}\nabla u\right\|_{L^\infty(\Omega)} & \text{if } u \in W^{1,\infty}(\Omega), \\ +\infty & \text{otherwise.} \end{cases}$$

In [28] the authors are interested in the case with isotropic phases and therefore they set $\Sigma(x) = \lambda^2(x)\mathrm{Id},\ x \in \Omega$, for some $\lambda \in L^\infty(\Omega)$.

Let us now focus on the periodic case for simplicity. In what follows, for $1 \le p \le +\infty$, we denote by $L^p_\sharp(Q)$ and $W^{1,p}_\sharp(Q)$ the spaces of Q-periodic functions belonging to $L^p_{\rm loc}(\mathbb{R}^N)$ and $W^{1,p}_{\rm loc}(\mathbb{R}^N)$, respectively, where $Q := (0,1)^N$ is the unit cube in $\mathbb{R}^N$. We assume that Σ is defined on Q, and we define the effective conductivity as follows:

$$g_p(\xi) := \inf \left\{ \mathcal{G}_p(u) \,:\, u \in W^{1,p}_{\sharp}(Q) + \xi \cdot x,\ \int_Q u(y)\,dy = 0 \right\}, \tag{4.2}$$

where

$$\mathcal{G}_p(u) := \int_Q \langle \Sigma(x)\nabla u(x), \nabla u(x)\rangle^{\frac{p}{2}}\,dx. \tag{4.3}$$

Next, we make the (very strong) assumption that g_p depends on ξ only through its norm. This gives

$$g_p(\xi) = (h_{p,\rm eff})^{\frac{p}{2}}|\xi|^p \tag{4.4}$$

for some number $h_{p,\rm eff}$ which will be interpreted as the effective conductivity. Let us remark for later purposes that the minimizer of (4.2) (see also (4.3)) is a weak solution of the anisotropic p-Laplace-type equation

$$\begin{cases} \operatorname{div}\left(\langle \Sigma(x)\nabla u(x), \nabla u(x)\rangle^{\frac{p-2}{2}} \Sigma(x)\nabla u(x)\right) = 0 & \text{if } x \in Q, \\ u \in W^{1,p}_{\sharp}(Q) + \xi \cdot x. \end{cases} \tag{4.5}$$

4.2. Resistivity: The divergence-free case. Now we consider the situation where the differential constraint $\mathcal{A}w = 0$ becomes div $w = 0$. In this case $d = N$, and we take

$$A^{(i)}_j := \delta_{ij},\ 1 \le i,j \le N.$$

Note that the constant rank condition is again satisfied, since for every $w \in S^{N-1}$ we have

$$\ker(\mathbb{A}(w)) = \left\{ v \in \mathbb{R}^N \ :\ \sum_{i=1}^N A^{(i)} w_i(v) = 0 \right\} = \{ v \in \mathbb{R}^N \ :\ w \cdot v = 0\},$$

which gives rank $(\mathbb{A}(w)) = \dim(\ker \mathbb{A}(w)) = N - 1$.

Let $f : \Omega \times \mathbb{R}^N \to [0, +\infty)$ be defined by $f(x,v) := \langle A(x)v, v\rangle^{\frac{1}{2}}$, where $A \in L^\infty\left(\Omega; \mathbb{M}^{N\times N}_{\rm sym}\right)$ satisfies the ellipticity bounds

$$\frac{1}{K}|\xi|^2 \le \langle A(x)\xi, \xi\rangle \le K|\xi|^2 \tag{4.6}$$

for $\mathcal{L}^N$-a.e. $x \in \Omega$ and all $\xi \in \mathbb{R}^N$, where $K \ge 1$ is a real constant. Note that f satisfies (3.1) and (3.2).

With these particular choices, we obtain the following corollaries of Theorems 3.1 and 3.2.

COROLLARY 4.3. *Let $T_q, T_\infty : L^1(\Omega;\mathbb{R}^N) \to [0, +\infty]$ be defined by*

$$T_q(b) := \begin{cases} \dfrac{1}{q}\displaystyle\int_\Omega \langle A(x)b(x), b(x)\rangle^{\frac{q}{2}}\,dx & \textit{if } b \in L^q(\Omega;\mathbb{R}^N),\ \operatorname{div} b = 0, \\ +\infty & \textit{otherwise} \end{cases}$$

and

$$T_\infty(b) := \begin{cases} 0 & \textit{if } \langle A(x)b(x), b(x)\rangle \le 1 \textit{ for } \mathcal{L}^N\textit{-a.e. } x \in \Omega, \\ +\infty & \textit{otherwise.} \end{cases}$$

Then

$$\Gamma(L^1(\Omega;\mathbb{R}^N)) - \lim_{q\to\infty} T_q = T_\infty.$$

Corollary 4.4. *Define* $S_q, S_\infty : L^1(\Omega;\mathbb{R}^N) \to [0,+\infty]$ *by*

$$S_q(b) := \begin{cases} \left(\int_\Omega \langle A(x)b(x), b(x)\rangle^{\frac{q}{2}}\, dx\right)^{1/q} & \textit{if } b \in L^q(\Omega;\mathbb{R}^N),\ \mathrm{div}\ b = 0, \\ +\infty & \textit{otherwise} \end{cases}$$

and

$$S_\infty(b) := \begin{cases} \|\langle Ab, b\rangle^{\frac{1}{2}}\|_{L^\infty(\Omega)} & \textit{if } b \in L^\infty(\Omega;\mathbb{R}^N),\ \mathrm{div}\ b = 0, \\ +\infty & \textit{otherwise.} \end{cases}$$

Then

$$\Gamma(L^1(\Omega;\mathbb{R}^N)) - \lim_{q\to\infty} S_q = S_\infty.$$

Let us now focus again on the periodic case, for simplicity. We assume that A is defined on the unit cube $Q = (0,1)^N$ and define the effective resistivity as follows:

$$t_q(\eta) := \inf\left\{\mathcal{T}_q(b) \,:\, b \in L^q_\sharp(Q;\mathbb{R}^N) + \eta,\ \mathrm{div}\ b = 0\right\}, \tag{4.7}$$

where

$$\mathcal{T}_q(b) := \int_Q \langle \Sigma^{-1}(x)b(x), b(x)\rangle^{\frac{q}{2}}\, dx. \tag{4.8}$$

Note that we have chosen

$$A \equiv \Sigma^{-1}, \tag{4.9}$$

where $\Sigma \in L^\infty\left(Q;\mathbb{M}^{N\times N}_{\mathrm{sym}}\right)$ is such that (4.1) holds. As in the previous case, we assume the dependence of $t_q(\eta)$ upon η is only through its norm. In this way, one has that

$$t_q(\eta) = (\rho_{q,\mathrm{eff}})^{\frac{q}{2}}|\eta|^q \tag{4.10}$$

for some number $\rho_{q,\mathrm{eff}}$ which we interpret as effective resistivity. The crucial fact is that, with the definitions (4.4) and (4.10), one has

$$h_{p,\mathrm{eff}} = \frac{1}{\rho_{q,\mathrm{eff}}}, \tag{4.11}$$

where $q = p'$ is the Hölder conjugate exponent of p, i.e., $\frac{1}{p} + \frac{1}{q} = 1$. Let us briefly explain where condition (4.11) comes from. We need the following well-known observations. Given $p \in (1,+\infty)$, if u_p solves (4.5), then the vector field

$$b_p := \langle \Sigma\nabla u_p, \nabla u_p\rangle^{\frac{p-2}{2}}\Sigma\nabla u_p \tag{4.12}$$

belongs to $L^q(Q;\mathbb{R}^N)$ and it is divergence free. Moreover, $b_p \in L^\infty_\sharp(Q;\mathbb{R}^N) + \eta$ for some vector $\eta \in \mathbb{R}^N$ which can be explicitly computed in terms of ξ and $h_{p,\mathrm{eff}}$. The Euler–Lagrange equations associated with (4.7) (see also (4.8)) are given by

$$\begin{cases} \operatorname{curl}\left(\left\langle \Sigma^{-1}b, b\right\rangle^{\frac{q-2}{2}} \Sigma^{-1}b\right) = 0 & \text{in } Q, \\ \operatorname{div} b = 0, & \text{in } Q, \\ b \in L^\infty_\sharp(Q;\mathbb{R}^N) + \eta. \end{cases} \tag{4.13}$$

We will derive such systems in the next section, where we need to address the issue of the limiting case $q \to +\infty$. It is easy to check that b_p does indeed solve the system (4.13) if one defines the vector η by

$$\eta := h_{p,\mathrm{eff}}^{\frac{p}{2}} |\xi|^{p-2}\xi. \tag{4.14}$$

In addition, one has the pointwise relation

$$\langle \Sigma\nabla u_p, \nabla u_p\rangle^{\frac{p}{2}} = \langle \Sigma^{-1}b_p, b_p\rangle^{\frac{q}{2}}, \tag{4.15}$$

which implies that

$$\int_Q \langle \Sigma\nabla u_p, \nabla u_p\rangle^{\frac{p}{2}}\, dx = \int_Q \langle \Sigma^{-1}b_p, b_p\rangle^{\frac{q}{2}}\, dx.$$

The above equation is equivalent to

$$h_{p,\mathrm{eff}}^{\frac{p}{2}} |\xi|^p = \rho_{q,\mathrm{eff}}^{\frac{q}{2}} |\eta|^q,$$

which, in view of (4.14), gives (4.11). The above calculations explain why in the literature about bounding nonlinear energies one looks simultaneously at both a variational principle in terms of gradient fields and one in terms of divergence-free fields. A lower bound on the effective conductivity relative to the exponent p is equivalent to an upper bound on the effective resistivity relative to the exponent q, and conversely.

5. Power-law and dielectric breakdown. In this section we formulate the strategy for our new definition of the effective yield set. This issue will be given a clear physical interpretation. However, as we shall see, a number of mathematical questions arise. We explain our results by specializing them to electrically conducting materials, since in this context our analysis is complete. We follow in part the exposition in [28], focusing on the simplest possible nontrivial situation.

5.1. Effective strength set for conducting materials. Suppose that we are given a material that occupies the domain $\Omega := \Omega_1 \cup \Omega_2$ which, for simplicity, will be taken to be the unit cube $Q = (0,1)^N$ $(N \geq 2)$. Given $0 < \alpha_1 < \alpha_2$, let $\alpha \in L^\infty(\Omega;[\alpha_1,\alpha_2])$ and, for $p > 1$ and $\xi \in \mathbb{R}^N$, consider

$$F_p(u) := \left(\frac{1}{p}\int_\Omega \alpha(x)^{\frac{p}{2}} |\nabla u(x)|^p dx\right)^{\frac{1}{p}} \tag{5.1}$$

and set

$$f_p(\xi) := \inf\left\{F_p(u) \,:\, u \in W^{1,p}_\sharp(\Omega) + \xi\cdot x\right\}. \tag{5.2}$$

We regard $f_p(\xi)$ as the overall (or effective) conductivity of the body in the direction $\xi \in \mathbb{R}^N$. With the notation of section 4, if we assume isotropy, we have

$$f_p(\xi) = \left(\frac{h_{p,\mathrm{eff}}^{\frac{1}{2}}}{p^{\frac{1}{p}}} \right) |\xi|. \tag{5.3}$$

The terminology is as follows. The material occupying the domain Ω is a two-phase composite whose conductivity is locally isotropic, given by the measurable function $\alpha(x) = \chi_{\Omega_1}(x)\alpha_1 + \chi_{\Omega_2}(x)\alpha_2$. The unknown is the "microgeometry," i.e., the set Ω_1, which is completely determined by its characteristic function χ_{Ω_1} with $\chi_{\Omega_2} = 1 - \chi_{\Omega_1}$. It is assumed that the relative volume fractions θ_1 and $\theta_2 = 1 - \theta_1$ are given with

$$\theta_1 = \frac{\mathcal{L}^N(\Omega_1)}{\mathcal{L}^N(\Omega)}. \tag{5.4}$$

The goal is to characterize (for every vector ξ) the closure of the range of $f_p(\xi)$ as χ_1 varies within the class of characteristic functions of measurable subsets Ω_1 of Ω satisfying (5.4). This kind of material is often of power-law type and has been studied extensively. The typical approach, in a large part of the engineering literature (see, for instance, the influential work of Hutchinson [32]), is to consider p as a "fitting parameter." In other words, one takes the viewpoint that due to too many nonlinear effects, it is hard to choose a priori the right "constitutive law." For given p, one would have the electric field of the form $e(x) = \nabla u(x)$ and the nonlinear response $b(x) = \alpha(x)^{\frac{p}{2}} |\nabla u(x)|^{p-2} \nabla u(x)$ with the constitutive law div $b = 0$. In other words, $u \in W^{1,p}_{\mathrm{loc}}(\Omega)$ and it solves (4.5) exactly as the minimizers of (4.3).

For given "microgeometry" (i.e., for given sets Ω_1 and Ω_2) one obtains the real number $f_p(\xi)$ given by (5.2) and (5.3). However, while the microgeometry is kept fixed, $f_p(\xi)$ varies with p (monotonically, by Jensen's inequality). Therefore, if one can measure the effective response of the material, one can choose p in such a way that will make the theoretical answer as accurate as possible. With this in mind, one can immediately understand why long ago people have been trying to take limits as $p \to \infty$ for this type of problem. Actually, as in, for instance, [32], calculations are often performed for "large" p. In [32, p. 109], "large" p means $p = 8$, and $p = +\infty$ is "extrapolated."

The mathematical approach has been quite different. We refer the reader to the work of Bouchitté and Suquet (see, e.g., [12]). A more direct approach has been proposed in [28], and this is the ancestor of what we do in the present paper. The idea is to pass to the limit as $p \to \infty$ in the sense of Γ-convergence. In particular, this new idea makes it possible to rigorously justify statements which were previously taken for granted. In fact, one achieves more naturally a very direct link between the minimizers of (5.2) and (at least one) minimizer of the new variational problem associated with F_∞ (the Γ-limit). Let us briefly review the classical approach versus the new one.

Within the classical approach, one considers $\alpha \in L^\infty(\Omega; [\alpha_1, \alpha_2])$ as above, $\lambda := \alpha^{\frac{1}{2}}$, and defines, for $u \in W^{1,\infty}(\Omega)$,

$$G_\infty(u) := \begin{cases} 0 & \text{if } |\lambda(x)\nabla u(x)| \le 1 \text{ for } \mathcal{L}^N\text{-a.e. } x \in \Omega, \\ +\infty & \text{otherwise} \end{cases}$$

and

$$g_\infty(\xi) := \inf \left\{ G_\infty(u) \,:\, u \in W^{1,\infty}_\sharp(\Omega) + \xi \cdot x \right\}. \tag{5.5}$$

Then the "strength" set (which is called yield set in plasticity) K_{eff} can be defined by

$$K_{\text{eff}} := \{\xi \in \mathbb{R}^N \,:\, g_\infty(\xi) = 0\}. \tag{5.6}$$

In this approach, the link between the minimizers of F_p and those of G_∞ is unclear. In particular it is not clear, a priori, whether $\lim_{p\to\infty} f_p(\xi)$ (which exists, by monotonicity) is indeed related to the value $g_\infty(\xi)$. The new approach is more transparent and direct: define

$$F_\infty(u) := ||\lambda \nabla u||_{L^\infty(\Omega)} \tag{5.7}$$

and

$$f_\infty(\xi) := \inf\left\{F_\infty(u) \,:\, u \in W^{1,\infty}_\sharp(\Omega) + \xi \cdot x\right\}. \tag{5.8}$$

As a direct consequence of Corollary 4.2 (see also [28]) one has

$$K_{\text{eff}} = \{\xi \in \mathbb{R}^N \,:\, f_\infty(\xi) \le 1\}. \tag{5.9}$$

Moreover, up to a subsequence, the minimizers of F_p converge to *some* minimizer of F_∞. In addition, $p \mapsto f_p(\xi)$ is continuous.

Another important advantage of this approach is that it strongly suggests an interesting connection with the infinity-Laplace equation. This is highlighted in the next subsection.

5.2. Infinity-harmonic functions. Let us explain this in the context developed in [28]. Assume that the material under consideration fills a regular body Ω, and that it is not only isotropic but also homogeneous. In this case $\alpha(x) \equiv \lambda(x) \equiv 1$, and the functional F_∞ in (5.7) becomes

$$F_\infty(u) = ||\nabla u||_{L^\infty(\Omega)}. \tag{5.10}$$

The problem of minimizing (in a properly interpreted sense) this functional represents the prototype for a series of related problems belonging to the emerging area of *calculus of variations in* L^∞. In this context, the functional F_∞ is strongly connected to the infinity-Laplace equation (1.3)—its associated Euler–Lagrange (Aronsson) equation, whose solutions must be understood in the viscosity sense. We will give more details shortly. However, before doing this, let us remark that these solutions lose many of the properties enjoyed by the solutions of the Euler–Lagrange equations associated with the functionals F_p. From the physical perspective, these phenomena are much less surprising. We refer the reader to sections 3 and 4 of [28] and to [27] for some examples.

The minimization of the functional F_p, subject to given boundary conditions, leads in the usual way to the associated Euler–Lagrange equation. It was first observed by Aronsson [2], [3], [4], [5], [6] in the one-dimensional case that minimization problems (properly interpreted) for supremal functionals such as (5.10) also lead to partial differential equations. We outline below, for the convenience of the reader, the formal derivation of the Aronsson equation associated with F_∞, as a limiting ($p \to \infty$) case of the Euler–Lagrange equations associated with F_p. For $\lambda \equiv 1$, this has been made rigorous in [10], where it was shown that weak solutions of the Euler–Lagrange equations associated with the p-Dirichlet integrals converge (up to a subsequence) to *viscosity* solutions of the limiting equation.

For $p \geq 1$ and when λ is smooth, minimizers of $F_p(u) := \frac{1}{p}\int_\Omega |\lambda(x)\nabla u(x)|^p dx$ in $W^{1,p}(\Omega)$, subject to given boundary conditions, solve the associated Euler–Lagrange equation

$$-\mathrm{div}\left(\lambda^p|\nabla u|^{p-2}\nabla u\right) = 0, \tag{5.11}$$

which may be rewritten as

$$\frac{p-2}{2}\lambda^{p-4}|\nabla u|^{p-4}\langle\nabla(\lambda^2|\nabla u|^2), \lambda^2\nabla u\rangle + \lambda^{p-2}|\nabla u|^{p-2}\mathrm{div}(\lambda^2\nabla u) = 0. \tag{5.12}$$

For $p > 2$, after division by $\frac{p-2}{2}\lambda^{p-4}|\nabla u|^{p-4}$, (5.12) becomes

$$\langle\nabla(\lambda^2|\nabla u|^2), \lambda^2\nabla u\rangle + \frac{2}{p-2}\lambda^2|\nabla u|^2\mathrm{div}(\lambda^2\nabla u) = 0. \tag{5.13}$$

Note that at least the necessary condition that must hold in order to divide by $\frac{p-2}{2}\lambda^{p-4}|\nabla u|^{p-4}$ in (5.12) is satisfied because the left-hand side of (5.13) also vanishes when either $\lambda = 0$ or $\nabla u = 0$. Formally letting $p \to +\infty$, we obtain the equation

$$\langle\nabla(\lambda^2|\nabla u|^2), \lambda^2\nabla u\rangle = 0. \tag{5.14}$$

In the general case, where F_p is given by the formula in Corollary 4.2 with Σ not necessarily isotropic, (5.14) should be replaced by

$$\langle\nabla\langle\Sigma\nabla u, \nabla u\rangle, \Sigma\nabla u\rangle = 0. \tag{5.15}$$

When $\lambda \equiv 1$ in (5.14), (5.11) reduces to the p-Laplace equation

$$\Delta_p u := \mathrm{div}(|\nabla u|^{p-2}\nabla u) = 0, \tag{5.16}$$

while (5.14) becomes

$$\Delta_\infty u = 0, \tag{5.17}$$

where Δ_∞ denotes the infinity-Laplace operator, which on smooth, real-valued functions u is defined by the formula

$$\Delta_\infty u := \sum_{i,j=1}^{N} \frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_j}\frac{\partial^2 u}{\partial x_i \partial x_j}. \tag{5.18}$$

The infinity-Laplace equation (5.17) has been introduced in the fundamental work of Aronsson [4], [5] as the candidate for the Euler–Lagrange equation associated with a properly interpreted minimization problem for supremal functionals (5.10). Precisely, we say that $u \in W^{1,\infty}(\Omega) \cap C(\overline{\Omega})$ is an *absolutely minimizing Lipschitz extension* (abbreviated AMLE) if

$$\|\nabla u\|_{L^\infty(D)} \leq \|\nabla v\|_{L^\infty(D)}$$

whenever $D \subset \Omega$ is open, and $v \in W^{1,\infty}(D) \cap C(\overline{D})$ is such that $u_{|\partial D} = v_{|\partial D}$. Aronsson showed that a necessary and sufficient condition for a smooth u to be an AMLE is that u solves (5.17) in the classical sense. On the other hand, he also noticed that (5.17) does not always have a classical solution. Motivated by the work in [4], [5], Jensen showed in [33] that every AMLE u is indeed a solution of the infinity-Laplace equation in the viscosity sense (see [18], [17]). We recall that u is called a viscosity solution of the infinity-Laplace equation if it is simultaneously a viscosity subsolution and a viscosity supersolution of (5.17):

(i) (u is a viscosity subsolution of (5.17)) for every local maximum point $x \in \Omega$ of $u - \varphi$, where φ is C^2 in some neighborhood of x, we have $-\Delta_\infty \varphi(x) \leq 0$;

(ii) (u is a viscosity supersolution of (5.17)) for every local minimum point $x \in \Omega$ of $u - \varphi$, where φ is C^2 in some neighborhood of x, we have $-\Delta_\infty \varphi(x) \geq 0$.

In addition, Jensen proved the maximum principle for solutions of (5.17) and, consequently, settled the crucial question regarding uniqueness of solutions. Generalizations of these ideas to minimization problems for generic supremal functionals of the form $\|f(\cdot, u(\cdot), \nabla u(\cdot))\|_{L^\infty(\Omega)}$ have been considered by Juutinen [34] and Barron, Jensen, and Wang [9], among others. We need to emphasize here that in order to be truly relevant for the study of composite materials, a theory would need to be valid with no smoothness assumptions on the coefficients. To our knowledge, such a theory is not yet available. A recent work by Juutinen [35] describes a promising approach.

6. The effective strength set for electrical resistivity. In this section we explore some applications of the Γ-convergence results obtained in section 3. Our results are motivated by the study of the yield set in polycrystal plasticity. Here, we address the mathematically very similar problem of electrical resistivity. The mathematical gap consists of the fact that in the former problem one should consider divergence-free matrix fields which are constrained to be symmetric, a constraint which we are allowed to neglect in the latter. Handling the extra constraint necessary to treat polycrystal plasticity does not seem to be a serious obstacle to the analysis, but it would require a rather long digression so we will address this issue elsewhere. We begin by providing a characterization of the effective strength set in the context of electrical resistivity in terms of the new L^∞ functionals acting on divergence-free fields. A formal derivation of the Aronsson equations associated with our new variational principle follows.

6.1. Effective strength for resistive materials. In this subsection the goal is to characterize the effective strength set K_{eff}, defined by (6.7), in the context of electrical resistivity. This is achieved in Proposition 6.2. Consider

$$j_q^{\text{eff}}(\eta) := \inf \left\{ \left(\int_Q f(x, b(x) + \eta)^q dx \right)^{1/q} : b \in L^q(Q; \mathbb{R}^N), \int_Q b \, dx = 0, \operatorname{div} b = 0 \right\}. \tag{6.1}$$

We note that the infimum in (6.1) is attained. This is an easy consequence of the lower semicontinuity result stated in Proposition 2.2.

PROPOSITION 6.1. *Let $f : Q \times \mathbb{R}^N \to \mathbb{R}$ be a Carathéodory integrand satisfying* (3.1) *and* (3.2). *For any $\eta \in \mathbb{R}^N$, $j_q^{\text{eff}}(\eta)$ converges to $j_\infty^{\text{eff}}(\eta)$ given by*

$$j_\infty^{\text{eff}}(\eta) := \inf \left\{ \operatorname*{ess\,sup}_{x \in Q} f(x, b(x) + \eta) : \ b \in L^\infty(Q; \mathbb{R}^N), \int_Q b \, dx = 0, \operatorname{div} b = 0 \right\}. \tag{6.2}$$

Moreover, if $b_q^{(\eta)}$ is a minimizer of the problem (6.1), *then, up to a subsequence, the sequence $\{b_q^{(\eta)}\}$ converges weakly in $L^1(Q; \mathbb{R}^N)$ to a minimizer $b^{(\eta)}$ of* (6.2).

Proposition 6.1 can be seen as a consequence of standard arguments regarding the convergence of minima and minimizers of Γ-converging functionals (see, e.g., [14] and [20]). For the convenience of the reader, we provide below a more elementary self-contained proof.

Proof. Let $\eta \in \mathbb{R}^N$, and let $b_q^{(\eta)} \in L^q(Q;\mathbb{R}^N)$ be such that $\int_Q b_q^{(\eta)}\, dx = 0$, $\operatorname{div} b_q^{(\eta)} = 0$, and

$$\left(\int_Q f(x, b_q^{(\eta)}(x) + \eta)^q dx\right)^{1/q} = j_q^{\text{eff}}(\eta). \tag{6.3}$$

We distinguish two cases.

Case 1.

$$\limsup_{q\to\infty} \left(\int_Q f(x, b_q^{(\eta)}(x) + \eta)^q dx\right)^{1/q} = +\infty.$$

We claim that in this case $j_\infty^{\text{eff}}(\eta) = +\infty$ as well. Indeed, if this were not the case, there would exist $b \in L^\infty(Q;\mathbb{R}^N)$ such that $\int_Q b\, dx = 0$, $\operatorname{div} b = 0$, and

$$\operatorname*{ess\,sup}_{x\in Q} f(x, b(x) + \eta) < j_\infty^{\text{eff}}(\eta) + 1 < +\infty.$$

Since b is admissible for $j_q^{\text{eff}}(\eta)$, we have

$$\left(\int_Q f(x, b_q^{(\eta)}(x) + \eta)^q dx\right)^{1/q} \leq \left(\int_Q f(x, b(x) + \eta)^q dx\right)^{1/q}$$

$$\leq \left(\int_Q \left(\operatorname*{ess\,sup}_{x\in Q} f(x, b(x) + \eta)\right)^q dx\right)^{1/q} \leq j_\infty^{\text{eff}}(\eta) + 1.$$

It follows that

$$\limsup_{q\to\infty} \left(\int_Q f(x, b_q^{(\eta)}(x) + \eta)^q dx\right)^{1/q} \leq j_\infty^{\text{eff}}(\eta) + 1,$$

a contradiction.

Case 2.

$$\limsup_{q\to\infty} \left(\int_Q f(x, b_q^{(\eta)}(x) + \eta)^q dx\right)^{1/q} < +\infty.$$

In view of the coercivity condition in (3.2), the sequence $\{b_q^{(\eta)} + \eta\}$ is bounded in $L^q(Q;\mathbb{R}^N)$. Extract a subsequence (not relabeled) such that $b_q^{(\eta)} \rightharpoonup b^{(\eta)}$ weakly in $L^1(Q;\mathbb{R}^N)$ with $\operatorname{div} b^{(\eta)} = 0$ and $\int_Q b^{(\eta)} dx = 0$. By Theorem 3.2, we deduce that

$$J_\infty(b^{(\eta)} + \eta) \leq \liminf_{q\to\infty} \left(\int_Q f(x, b_q^{(\eta)}(x) + \eta)^q dx\right)^{1/q} < +\infty. \tag{6.4}$$

In particular, $b^{(\eta)} \in L^\infty(Q;\mathbb{R}^N)$. We have

$$\begin{aligned}\limsup_{q\to\infty} j_q^{\rm eff}(\eta) &= \limsup_{q\to\infty}\left(\int_Q f(x, b_q^{(\eta)}(x)+\eta)^q dx\right)^{1/q}\\ &\le \limsup_{q\to\infty}\left(\int_Q f(x, b^{(\eta)}(x)+\eta)^q dx\right)^{1/q} = \operatorname*{ess\,sup}_{x\in Q} f(x, b^{(\eta)}(x)+\eta)\\ &\le \liminf_{q\to\infty} j_q^{\rm eff}(\eta),\end{aligned}$$

where we have used (6.3) and (6.4). Thus, $\lim_{q\to\infty} j_q^{\rm eff}(\eta)$ exists, and

$$\operatorname*{ess\,sup}_{x\in Q} f(x, b^{(\eta)}(x)+\eta) = \lim_{q\to\infty} j_q^{\rm eff}(\eta).$$

It remains to show that

$$\operatorname*{ess\,sup}_{x\in Q} f(x, b^{(\eta)}(x)+\eta) = j_\infty^{\rm eff}(\eta). \tag{6.5}$$

Clearly, $j_\infty^{\rm eff}(\eta) \le \operatorname{ess\,sup}_{x\in Q} f(x, b^{(\eta)}(x)+\eta)$. To show the converse inequality, consider $b \in L^\infty(Q;\mathbb{R}^N)$ such that div $b = 0$ and $\int_Q b\; dx = 0$. We have

$$\begin{aligned}\operatorname*{ess\,sup}_{x\in Q} f(x, b(x)+\eta) &= \lim_{q\to\infty}\left(\int_Q f(x, b(x)+\eta)^q dx\right)^{1/q}\\ &\ge \lim_{q\to\infty} j_q^{\rm eff}(\eta) = \operatorname*{ess\,sup}_{x\in Q} f(x, b^{(\eta)}(x)+\eta).\end{aligned}$$

Passing to the infimum over all fields b satisfying the admissibility conditions, we obtain that $j_\infty^{\rm eff}(\eta) \ge \operatorname{ess\,sup}_{x\in Q} f(x, b^{(\eta)}(x)+\eta)$. It follows that (6.5) holds. ∎

Consider now the situation where the pointwise constraint on the stress (current) may be written in the form

$$\sigma(x) \in \left\{\eta \in \mathbb{R}^N \;:\; f(x,\eta) \le 1\right\}, \tag{6.6}$$

where $f : Q \times \mathbb{R}^N \to \mathbb{R}$ satisfies our hypotheses (3.1) and (3.2). In the context of (first failure) models of dielectric breakdown for composites made of two isotropic phases, considered by Garroni, Nesi, and Ponsiglione in [28], the constraint (6.6) is replaced by the requirement that the electric field ∇u satisfy $\nabla u(x) \in K(x)$, where

$$K(x) = \left\{\eta \in \mathbb{R}^2 \;:\; \lambda(x)|\eta| \le 1\right\}$$

with $\lambda(x)$ being a combination of the characteristic functions of the phases. Thus, in that case, $K(x)$ is a disc whose radius varies from point to point. Such a model is concerned with electrical conductivity, and therefore the relevant fields are curl free. In dimension $N \ge 3$, if one wants to model electrical resistivity, then the right differential constraint is that the relevant fields are divergence free. In view of the pointwise constraint (6.6), the yield set of the polycrystal becomes

$$K_{\text{eff}} = \Bigg\{\eta \in \mathbb{R}^N \ : \ \text{there exists } \sigma \in L^\infty(Q;\mathbb{R}^N) \text{ such that } \eta = \int_Q \sigma(x)dx,$$
$$\text{div } \sigma = 0, f(x,\sigma(x)) \leq 1, \mathcal{L}^N\text{-a.e. } x \in Q\Bigg\}$$
$$= \Bigg\{\eta \in \mathbb{R}^N \ : \ \text{there exists } \sigma \in L^\infty(Q;\mathbb{R}^N) \text{ such that } \int_Q \sigma(x)dx = 0,$$
$$\text{div } \sigma = 0, f(x,\sigma(x)+\eta) \leq 1, \mathcal{L}^N\text{-a.e. } x \in Q\Bigg\}. \tag{6.7}$$

We are now in position to describe K_{eff} by means of the L^∞ variational principle (6.2).

PROPOSITION 6.2. *Let $N \in \{2,3\}$, and let $f : Q \times \mathbb{R}^N \to \mathbb{R}$ be a Carathéodory function satisfying* (3.1) *and* (3.2). *The yield set of a conducting polycrystal whose current (stress) σ satisfies the pointwise constraint* (6.6) *at almost every point and the equilibrium equation* div $\sigma = 0$ *is described by*

$$K_{\text{eff}} = \left\{\eta \in \mathbb{R}^N \ : \ j_\infty^{\text{eff}}(\eta) \leq 1\right\}. \tag{6.8}$$

Proof. Let $\eta \in K_{\text{eff}}$. By (6.7), there exists $\sigma \in L^\infty(Q;\mathbb{R}^N)$ such that $\int_Q \sigma(x)dx = 0$, div $\sigma = 0$, and with $f(x,\sigma(x)+\eta) \leq 1$ for $\mathcal{L}^N$-a.e. $x \in Q$. We obtain that $j_\infty^{\text{eff}}(\eta) \leq \operatorname{ess\,sup}_{x\in Q} f(x,\sigma(x)+\eta) \leq 1$. Conversely, let $\eta \in \mathbb{R}^N$ be such that

$$j_\infty^{\text{eff}}(\eta) \leq 1. \tag{6.9}$$

Let $\{\sigma_n\} \subset L^\infty(Q;\mathbb{R}^N)$ be a sequence such that div $\sigma_n = 0$, $\int_Q \sigma_n(x)dx = 0$ for any $n \in \mathbb{N}$, and

$$\lim_{n\to\infty} \operatorname*{ess\,sup}_{x\in Q} f(x,\sigma_n(x)+\eta) = j_\infty^{\text{eff}}(\eta). \tag{6.10}$$

Taking (3.2) into account, we may extract a subsequence of $\{\sigma_n\}$ (not relabeled) such that $\sigma_n \rightharpoonup \sigma$ weakly* in $L^\infty(Q;\mathbb{R}^N)$, with div $\sigma = 0$, and $\int_Q \sigma(x)dx = 0$. Let $x \in Q$ be a Lebesgue point for $f(\cdot,\sigma(\cdot)+\eta)$. By (3.1), and in view of Proposition 2.2, we have that

$$\int_{B(x,r)} f(y,\sigma(y)+\eta)dy \leq \liminf_{n\to\infty} \int_{B(x,r)} f(y,\sigma_n(y)+\eta)dy$$

for sufficiently small $r > 0$. Thus, (6.10) yields

$$\frac{1}{\mathcal{L}^N(B(x,r))} \int_{B(x,r)} f(y,\sigma(y)+\eta)dy \leq j_\infty^{\text{eff}}(\eta).$$

Letting $r \to 0^+$, and since a.e. point $x \in Q$ is a Lebesgue point for $f(\cdot,\sigma(\cdot)+\eta)$, we deduce that $f(x,\sigma(x)+\eta) \leq j_\infty^{\text{eff}}(\eta)$ for $\mathcal{L}^N$-a.e. $x \in Q$. Taking (6.9) into account, we deduce that $\eta \in K_{\text{eff}}$. Thus, (6.8) holds. ∎

Let us remark that, although our new definition of K_{eff} provided by Proposition 6.2 is novel, the study of bounds on the effective strength set in this context is not. Kohn and Little [37] considered the case of antiplane shear in the two-dimensional case, and there the corresponding sets $K(x)$ are rectangles whose orientation varies from point to point. In this case their analysis is equivalently performed on both curl-free fields and symmetrized gradients. A similar study is considered in [27], where some interesting new bounds are proved in three dimensions for both gradients and divergence-free fields.

An interesting open question is whether j_∞^{eff}, which enters the expression of K_{eff} in (6.8), can be recovered in the general case as a relevant energy density via homogenization. Let us assume that $f : \mathbb{R}^N \times \mathbb{R}^d \to \mathbb{R}$ is a Carathéodory function, $(0,1)^N$-periodic in its first variable. For $\varepsilon > 0$, define $J_{p,\varepsilon} : L^1(\Omega;\mathbb{R}^d) \to [0,+\infty]$ by

$$J_{p,\varepsilon}(v) := \begin{cases} \left(\int_\Omega f\left(\frac{x}{\varepsilon}, v(x)\right)^p dx\right)^{1/p} & \text{if } v \in L^p(\Omega;\mathbb{R}^d) \cap \ker\mathcal{A}, \\ +\infty & \text{otherwise,} \end{cases}$$

where $\mathcal{A}$ is a constant rank operator (see (2.1)). If $f(x,\cdot)$ is $\mathcal{A}$-quasiconvex and if (3.2) holds we have, in view of Theorem 3.2, that $J_{p,\varepsilon}$ $\Gamma(L^1(\Omega;\mathbb{R}^d))$-converges as $p \to \infty$ to $J_{\infty,\varepsilon} : L^1(\Omega;\mathbb{R}^d) \to [0,+\infty]$ defined by

$$J_\varepsilon(u;\Omega) := \begin{cases} \operatorname*{ess\,sup}_{x\in\Omega} f\left(\frac{x}{\varepsilon}, v(x)\right) & \text{if } v \in L^\infty(\Omega;\mathbb{R}^d) \cap \ker\mathcal{A}, \\ +\infty & \text{otherwise.} \end{cases}$$

A first step toward answering the above question is to decide whether the Γ-limit (as $\varepsilon \to 0^+$) of F_ε with respect to the weak* topology of $L^\infty(\Omega;\mathbb{R}^d)$ may itself be written in supremal form. If this is the case, under which assumptions on f is j_∞^{eff} the supremand of the Γ-limit? General homogenization and relaxation results for integral functionals in the context of $\mathcal{A}$-quasiconvexity have been studied by Braides, Fonseca, and Leoni in [15]. We also refer the reader to the recent paper by Ansini and Garroni [1] where, in the divergence-free case $\mathcal{A}v = \operatorname{div} v$, some of the general arguments of [15] are simplified and more explicit constructions are given. Except for some recent progress in the unconstrained case and in the curl-free case (see, e.g., [29]), the corresponding questions for supremal functionals are yet to be addressed. We remark that the condition on f which should be relevant in this context is the necessary and sufficient condition for the weak* lower semicontinuity in $L^\infty(\Omega;\mathbb{R}^d)$ of supremal functionals of the type

$$v \;\mapsto\; \operatorname*{ess\,sup}_{x\in\Omega} f(x, v(x)).$$

For the general case of constant rank PDE constraints on the fields v such a condition has not yet been identified. In the unconstrained scalar case, which corresponds to $\mathcal{A} \equiv 0$, it is known (see Barron, Jensen, and Wang [8]) that the necessary and sufficient condition for lower semicontinuity is the *level set convexity* of $f(x,\cdot)$, that is, for $\mathcal{L}^N$-a.e. $x \in \Omega$ and every $u \in \mathbb{R}^m$, the sets $\{v \in \mathbb{R}^d : f(x,v) \le \gamma\}$ are convex for any choice of $\gamma \in \mathbb{R}$. For the case $v = \nabla u$ in a vectorial setting, corresponding to the particular choice $\mathcal{A} = 0 \Leftrightarrow \operatorname{curl} u = 0$, the necessary and sufficient condition for the lower semicontinuity of the functional $W^{1,\infty}(\Omega;\mathbb{M}^{m\times N}) \ni u \mapsto \operatorname{ess\,sup}_{x\in\Omega} f(x,\nabla u(x))$ with respect to the weak* topology in $W^{1,\infty}(\Omega;\mathbb{M}^{m\times N})$ has been identified in [8] as being the *strong Morrey quasiconvexity* of f in the last variable. Let $Q = (0,1)^N$. A Borel measurable map $g : \mathbb{M}^{m\times N} \to \mathbb{R}$ is said to be strong Morrey quasiconvex if for any $\varepsilon > 0$, each $A \in \mathbb{M}^{m\times N}$, and any $K > 0$, there exists $\delta = \delta(\varepsilon, K, A) > 0$ such that if $\varphi \in W^{1,\infty}(Q;\mathbb{R}^m)$ satisfies

$$\|\nabla\varphi\|_{L^\infty(Q;\mathbb{M}^{m\times N})} \le K, \quad \max_{x\in\partial Q} |\varphi(x)| \le \delta,$$

then

$$g(A) \le \operatorname*{ess\,sup}_{x\in Q} g(A + \nabla\varphi(x)) + \varepsilon.$$

In the curl-free case the question of whether the corresponding j_∞^{eff} arises directly as a relevant energy density via homogenization has been answered affirmatively by Briani, Garroni, and Prinari in [16] under the assumption of level set convexity on $f(x,\cdot)$.

6.2. The new Aronsson equations suggested by power-law resistivity. In what follows we will focus on Corollary 4.4, and motivated by the results of the previous subsection we derive the Euler–Lagrange equation associated with the minimization of S_q. Next we identify, via a formal limiting procedure as $q \to \infty$, the candidate for the Euler–Lagrange equation (the Aronsson equation) corresponding to the minimization problem of the Γ-limit S_∞.

For $q \ge 1$ and A smooth, any minimizer b of S_q subject to appropriate boundary conditions must satisfy

$$S_q(b) \le S_q(b + t\varphi)$$

for any $t \in \mathbb{R}$ and any $\varphi \in C_c^\infty(\Omega;\mathbb{R}^N)$ such that div $\varphi = 0$ in Ω. Thus, for any such φ, we must have

$$\frac{\mathrm{d}}{\mathrm{d}t} S_q(b + t\varphi)_{|t=0} = 0. \tag{6.11}$$

Since

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} S_q(b + t\varphi)\\
&\quad = \frac{1}{q}\left(\int_\Omega \langle Ab + tA\varphi, b + t\varphi\rangle^{\frac{q}{2}}\, dx\right)^{\frac{1}{q}-1} \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \langle Ab + tA\varphi, b + t\varphi\rangle^{\frac{q}{2}}\, dx\\
&\quad = \frac{1}{2}\left(\int_\Omega \langle Ab + tA\varphi, b + t\varphi\rangle^{\frac{q}{2}}\, dx\right)^{\frac{1}{q}-1}\\
&\qquad \times \int_\Omega \langle Ab + tA\varphi, b + t\varphi\rangle^{\frac{q-2}{2}} \Big(\langle A\varphi, b\rangle + \langle Ab, \varphi\rangle + 2t\langle A\varphi, \varphi\rangle\Big) dx,
\end{aligned}$$

(6.11) becomes

$$\int_\Omega \Big\langle \langle A(x)b(x), b(x)\rangle^{\frac{q-2}{2}} \left(\left(A^T(x) + A(x)\right) b(x)\right), \varphi(x)\Big\rangle\, dx = 0 \tag{6.12}$$

for all $\varphi \in C_c^\infty(\Omega;\mathbb{R}^N)$, where A^T stands for the transpose of A.

Testing in (6.12) with $\varphi = \text{curl}\,\xi$, for $\xi \in C_c^\infty(\Omega)$, we find that b must satisfy the system

$$\begin{cases} \text{curl}\left[\langle Ab,b\rangle^{\frac{q-2}{2}}\left(A^T+A\right)b\right]=0 & \text{in } \Omega,\\ \text{div}\; b=0 & \text{in } \Omega \end{cases} \tag{6.13}$$

in the sense of distributions.

We now focus on the physically more relevant case $N=3$. Note that we may rewrite the left-hand side of the first equation in (6.13) as

$$\begin{aligned} &\text{curl}\left[\langle Ab,b\rangle^{\frac{q-2}{2}}\left(A^T+A\right)b\right]\\ &\quad=\langle Ab,b\rangle^{\frac{q-2}{2}}\text{curl}\left[\left(A^T+A\right)b\right]+\left(\frac{q-2}{2}\langle Ab,b\rangle^{\frac{q-4}{2}}\nabla\langle Ab,b\rangle\right)\wedge\left(A^T+A\right)b\\ &\quad=\langle Ab,b\rangle^{\frac{q-2}{2}}\text{curl}\left[\left(A^T+A\right)b\right]\\ &\qquad+\frac{q-2}{2}\langle Ab,b\rangle^{\frac{q-4}{2}}\left(\left([D(Ab)]^T b+[Db]^T Ab\right)\wedge\left(A^T+A\right)b\right), \end{aligned}$$

where $u\wedge v := (u_2v_3-u_3v_2, u_3v_1-u_1v_3, u_1v_2-u_2v_1)$ is the exterior (cross) product of $u=(u_1,u_2,u_3)$ and $v=(v_1,v_2,v_3)$ in $\mathbb{R}^3$.

Thus, for $q>4$, we may rewrite the system (6.13) as

$$\begin{cases} \left([D(Ab)]^T b+[Db]^T Ab\right)\wedge\left(A^T+A\right)b & \\ \qquad +\frac{2}{q-2}\langle Ab,b\rangle\text{curl}\left[\left(A^T+A\right)b\right]=0 & \text{in } \Omega,\\ \text{div}\; b=0 & \text{in } \Omega \end{cases} \tag{6.14}$$

after formally dividing the first equation in (6.13) by $\frac{q-2}{2}\langle Ab,b\rangle^{\frac{q-4}{2}}$. Note that at least the necessary condition which must hold in order to perform the division is satisfied, since on the set where $\langle Ab,b\rangle=0$ the resulting expression also vanishes.

The limiting system as $q\to\infty$ reads

$$\begin{cases} \left([D(Ab)]^T b+[Db]^T Ab\right)\wedge\left(A^T+A\right)b=0 & \text{in } \Omega,\\ \text{div}\; b=0 & \text{in } \Omega. \end{cases} \tag{6.15}$$

In the isotropic and homogeneous case, where A is taken to be the identity matrix in $\mathbb{M}^{3\times 3}$, the systems (6.14) and (6.15) become

$$\begin{cases} \left([Db]^T b\right)\wedge b+\frac{1}{q-2}|b|^2\text{curl}\; b=0 & \text{in } \Omega,\\ \text{div}\; b=0 & \text{in } \Omega \end{cases} \tag{6.16}$$

and

$$\begin{cases} \left([Db]^T b\right)\wedge b=0 & \text{in } \Omega,\\ \text{div}\; b=0 & \text{in } \Omega, \end{cases} \tag{6.17}$$

respectively. Thus, smooth solutions of (6.17) are solenoidal vector fields b whose values at $x \in \Omega$ are eigenvectors of $[Db(x)]^T$. Equivalently, (6.17) can be written as

$$\begin{cases} b \wedge \nabla |b|^2 = 0 & \text{in } \Omega, \\ \operatorname{div} b = 0 & \text{in } \Omega. \end{cases} \tag{6.18}$$

Comparing with the curl-free situation, the system (6.13) (in the isotropic, homogeneous case $A = \mathrm{Id} \in \mathbb{M}^{N \times N}$) can be viewed as being dual to the p-Laplace equation (5.16) which reduces to the usual Laplace equation $\Delta u = 0$ (u is *harmonic*) when $p = 2$. Similarly, if $q = 2$, (6.13) becomes (again, take $A = \mathrm{Id} \in \mathbb{M}^{N \times N}$)

$$\begin{cases} \operatorname{curl} b = 0 & \text{in } \Omega, \\ \operatorname{div} b = 0 & \text{in } \Omega. \end{cases} \tag{6.19}$$

Equivalently, with $b(x_1, \ldots, x_N) = (b_1(x_1, \ldots, x_N), \ldots, b_N(x_1, \ldots, x_N))$, this is a Riesz system of partial differential equations

$$\begin{cases} \frac{\partial b_i}{\partial x_j} = \frac{\partial b_j}{\partial x_i} \ (i, j = 1, \ldots, N) & \text{in } \Omega, \\ \sum_{i=1}^N \frac{\partial b_i}{\partial x_i} = 0 & \text{in } \Omega, \end{cases}$$

a generalized Cauchy–Riemann system (see Stein and Weiss [43]). Indeed, in the two-dimensional case (6.19) is, simply, the system of Cauchy–Riemann equations

$$\begin{cases} \frac{\partial b_2}{\partial x_1} = \frac{\partial b_1}{\partial x_2} & \text{in } \Omega, \\ \frac{\partial b_2}{\partial x_2} = -\frac{\partial b_1}{\partial x_1} & \text{in } \Omega. \end{cases} \tag{6.20}$$

In this case $b_1 - ib_2$ is an analytic function of $z = x_1 + ix_2$, and $b = (b_1, b_2)$ is the gradient of a harmonic function in the region Ω. Vector-valued functions $b = (b_1, \ldots, b_N) : \Omega \subset \mathbb{R}^N \to \mathbb{R}^N$ which are gradients of harmonic functions in Ω are called *generalized analytic* [43].

Viscosity solutions of the infinity-Laplace equation (5.17) arising in the curl-free case are called *infinity-harmonic functions*. In view of the duality described above, we will loosely refer to the solutions of (6.17) as *infinity analytic* (although the notion of solution to (6.17) will later need to be understood in an appropriate sense). We conjecture that the system (6.17) (or, equivalently, (6.18)) is the correct "Aronsson equation" associated with the minimization of the supremal functional

$$b \ \mapsto \ \operatorname*{ess\,sup}_{x \in \Omega} |b(x)| \tag{6.21}$$

considered on divergence-free fields b. However, many basic questions about the system (6.17) will need to be answered in order to justify this properly. For example, is it true that, in the spirit of the results available in the curl-free case, absolute (local) minimizers of (6.21) are infinity analytic? In particular, in which sense do absolute minimizers solve the system (6.17)? Are they the unique solution of this system? Unlike in the curl-free case, where one shows that absolute minimizers of $u \mapsto \operatorname{ess\,sup}_{x \in \Omega} |\nabla u(x)|$ are solutions to the infinity-Laplace equation in the viscosity sense (note that here u is a scalar function), the method of viscosity solutions may not be the right tool for the study of (6.17) (whose solutions must be vector fields). To overcome this difficulty, we need to pursue alternative methods, capable of handling the vectorial case. Perhaps the most outstanding problem along these lines is to understand whether an efficient weak formulation for the system (6.15) can be identified in the case of nonsmooth coefficients A.

Finally, we would like to comment very briefly on the limiting cases of the usual (p, q)-Hölder duality (here, $q = p'$, with $\frac{1}{p} + \frac{1}{p'} = 1$) when $p \to \infty$ (which corresponds to $q \to 1^+$) or $q \to \infty$ (equivalently, $p \to 1^+$). In the context of power-law materials at "level" p, convex duality is fully understood. Guided by the duality between the p-Laplace equation (5.16) and the system (6.16) (as we have already seen, these equations correspond to dual variational principles), it is not hard to see, formally, that when $p \to 1^+$ in (5.16) one has that u satisfies the 1-Laplace equation (see, e.g., [36])

$$\operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right) = 0, \tag{6.22}$$

while when $q \to 1^+$ in (6.16), b satisfies the system

$$\operatorname{curl}\left(\frac{b}{|b|}\right) = 0. \tag{6.23}$$

Again, heuristically, if u_1 is a solution to (6.22), then the field b_1 defined by

$$b_1 := \frac{\nabla u_1}{|\nabla u_1|} \tag{6.24}$$

is divergence free, and we have $|b_1|^2 = 1$ a.e. Therefore, b_1 is a solution of (6.18) (or, equivalently, (6.17)), which is the limiting ($q \to \infty$) system associated with (6.16). On the other hand, if one starts from a field b_1 which satisfies (6.23), one is given a function u_1 such that

$$\nabla u_1 := \frac{b_1}{|b_1|}. \tag{6.25}$$

It is not hard to check that u_1 formally satisfies $\Delta_\infty u_1 = 0$.

Whether a satisfactory mathematical theory can be developed to justify these statements in some appropriate weak sense is an intriguing open question which we believe deserves further investigation.

Acknowledgment. The authors would like to thank the anonymous referees for several constructive suggestions which have helped improve the exposition.

REFERENCES

[1] N. Ansini and A. Garroni, *Γ-convergence of functionals on divergence free fields*, ESAIM Control Optim. Calc. Var., 13 (2007), pp. 809–828.

[2] G. Aronsson, *Minimization problems for the functional* $\sup_x F(x, f(x), f'(x))$, Ark. Mat., 6 (1965), pp. 33–53.

[3] G. Aronsson, *Minimization problems for the functional* $\sup_x F(x, f(x), f'(x))$. II, Ark. Mat., 6 (1966), pp. 409–431.

[4] G. Aronsson, *Extensions of functions satisfying Lipschitz conditions*, Ark. Mat., 6 (1967), pp. 551–561.

[5] G. Aronsson, *On the partial differential equation* $u_x^2 u_{xx} + 2u_x u_y u_{xy} + u_y^2 u_{yy} = 0$, Ark. Mat., 7 (1968), pp. 395–425.

[6] G. Aronsson, *Minimization problems for the functional* $\sup_x F(x, f(x), f'(x))$. III, Ark. Mat., 7 (1969), pp. 509–512.

[7] G. Aronsson, M. G. Crandall, and P. Juutinen, *A tour of the theory of absolutely minimizing functions*, Bull. Amer. Math. Soc. (N.S.), 41 (2004), pp. 439–505.

[8] E. N. Barron, R. R. Jensen, and C. Y. Wang, *Lower semicontinuity of L^∞ functionals*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 18 (2001), pp. 495–517.

[9] E. N. Barron, R. R. Jensen, and C. Y. Wang, *The Euler equation and absolute minimizers of L^∞ functionals*, Arch. Ration. Mech. Anal., 157 (2001), pp. 255–283.

[10] T. Bhattacharya, E. DiBenedetto, and J. Manfredi, *Limits as $p \to \infty$ of $\Delta_p\, u_p = f$ and related extremal problems*, Rend. Sem. Mat. Univ. Politec. Torino, special issue (1991), pp. 15–68.

[11] J. Bishop and R. Hill, *A theory for the plastic distortion of a polycrystalline aggregate under combined stresses*, Phil. Mag. A, 42 (1951), pp. 414–427.

[12] G. Bouchittè and P. Suquet, *Homogenization, plasticity and yield design*, in Composite Media and Homogenization Theory (Trieste, 1990), Progr. Nonlinear Differential Equations Appl. 5, Birkhäuser, Boston, 1991, pp. 107–133.

[13] A. Braides, *Γ-Convergence for Beginners*, Oxford University Press, Oxford, UK, 2002.

[14] A. Braides and A. Defranceschi, *Homogenization of Multiple Integrals*, Oxford Lecture Ser. Math. Appl. 12, Oxford University Press, Oxford, UK, 1998.

[15] A. Braides, I. Fonseca, and G. Leoni, *$\mathcal{A}$-quasiconvexity: Relaxation and homogenization*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 539–577.

[16] A. Briani, A. Garroni, and F. Prinari, *Homogenization of L^∞ functionals*, Math. Models Methods Appl. Sci., 14 (2004), pp. 1761–1784.

[17] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[18] M. G. Crandall and P.-L. Lions, *Viscosity solutions and Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[19] B. Dacorogna, *Weak Continuity and Weak Lower Semicontinuity for Nonlinear Functionals*, Lecture Notes in Math. 922, Springer-Verlag, New York, 1982.

[20] G. Dal Maso, *An Introduction to Γ-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkäuser, Boston, 1993.

[21] E. De Giorgi, *Sulla convergenza di alcune succesioni di integrali del tipo dell'area*, Rend. Mat., 8 (1975), pp. 277–294.

[22] E. De Giorgi and T. Franzoni, *Su un tipo di convergenza variazionale*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur., 58 (1975), pp. 842–850.

[23] I. Ekeland and R. Témam, *Convex Analysis and Variational Problems*, Classics in Appl. Math. 28, SIAM, Philadelphia, PA, 1999.

[24] L. C. Evans and Y. Yu, *Various properties of solutions of the infinity-Laplacian equation*, Comm. Partial Differential Equations, 30 (2005), pp. 1401–1428.

[25] I. Fonseca and S. Müller, *$\mathcal{A}$-quasiconvexity, lower semicontinuity, and Young measures*, SIAM J. Math. Anal., 30 (1999), pp. 1355–1390.

[26] N. Fusco and G. Moscariello, *An application of duality to homogenization of integral functionals*, Atti Accad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. Sez. Ia (8), 17 (1984), pp. i–iii, 361–372.

[27] A. Garroni and R. V. Kohn, *Some three-dimensional problems related to dielectric breakdown and polycrystal plasticity*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 459 (2003), pp. 2613–2625.

[28] A. Garroni, V. Nesi, and M. Ponsiglione, *Dielectric breakdown: Optimal bounds*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 2317–2335.

[29] A. Garroni, M. Ponsiglione, and F. Prinari, *From 1-homogeneous supremal functionals to difference quotients: Relaxation and Γ-convergence*, Calc. Var. Partial Differential Equations, 27 (2006), pp. 397–420.

[30] G. H. Goldsztein, *Rigid perfectly plastic two-dimensional polycrystals*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 2789–2798.

[31] G. H. Goldsztein, *Two-dimensional rigid polycrystals whose grains have one ductile direction*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 459 (2003), pp. 1949–1968.

[32] J. W. Hutchinson, *Bounds and self-consistent estimates for creep of polycrystalline materials*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 348 (1976), pp. 101–127.

[33] R. Jensen, *Uniqueness of Lipschitz extensions: Minimizing the sup norm of the gradient*, Arch. Ration. Mech. Anal., 123 (1993), pp. 51–74.

[34] P. Juutinen, *Minimization problems for Lipschitz functions via viscosity solutions*, Ann. Acad. Sci. Fenn. Math. Diss., No. 115 (1998), pp. 1–53.

[35] P. Juutinen, *Quasiminima of the Lipschitz extension problem*, Ann. Mat. Pura Appl. (4), 186 (2007), pp. 303–316.

[36] B. Kawohl, *On a family of torsional creep problems*, J. Reine Angew. Math., 410 (1990), pp. 1–22.
[37] R. V. Kohn and T. D. Little, *Some model problems of polycrystal plasticity with deficient basic crystals*, SIAM J. Appl. Math., 59 (1998), pp. 172–197.
[38] S. Müller, *Rank-one convexity implies quasiconvexity on diagonal matrices*, Int. Math. Res. Not., 20 (1999), pp. 1087–1095.
[39] F. Murat, *Compacité par compensation: Condition necessaire et suffisante de continuité faible sous une hypothése de rang constant*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 8 (1981), pp. 68–102.
[40] P. Ponte Castañeda, *New variational principles in plasticity and their application to composite materials*, J. Mech. Phys. Solids, 40 (1992), pp. 1757–1788.
[41] P. Ponte Castañeda and P. Suquet, *Nonlinear composites*, Adv. Appl. Mech., (1998) 34, pp. 171–302.
[42] G. Sachs, *Zur Albeitung einer Fleissbedingun*, Z. Ver. Dtsch. Ing., 72 (1928), pp. 734–736.
[43] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.
[44] P. Suquet, *Overall potentials and extremal surfaces of power law or ideally plastic composites*, J. Mech. Phys. Solids, 41 (1993), pp. 981–1002.
[45] D. R. S. Talbot and J. R. Willis, *Variational principles for inhomogeneous nonlinear media*, IMA J. Appl. Math., 35 (1985), pp. 39–54.
[46] D. R. S. Talbot and J. R. Willis, *Bounds for the effective constitutive relation of a nonlinear composite*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 2705–2723.
[47] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Vol. IV, Pitman Res. Notes Math. 39, R. Knops, ed., Longman, Harlow, UK, 1979, pp. 136–212.
[48] L. Tartar, *The compensated compactness method applied to systems of conservation laws*, in Systems of Nonlinear Partial Differential Equations, J. M. Ball, ed., D. Reidel, Dordrecht, 1983, pp. 263–285.
[49] G. Taylor, *Plastic strains in metals*, J. Inst. Metals, 62 (1938), pp. 307–324.

SIAM J. MATH. ANAL.
Vol. 39, No. 5, pp. 1577–1594

NEWTONIAN LIMIT FOR WEAKLY VISCOELASTIC FLUID FLOWS OF OLDROYD TYPE*

LUC MOLINET† AND RAAFAT TALHOUK‡

Abstract. This paper is concerned with regular flows of incompressible weakly viscoelastic fluids which obey a differential constitutive law of Oldroyd type. We study the Newtonian limit for weakly viscoelastic fluid flows in $\mathbb{R}^N$ or $\mathbb{T}^N$ for $N = 2, 3$, when the Weissenberg number (relaxation time measuring the elasticity effect in the fluid) tends to zero. More precisely, we prove that the velocity field and the extrastress tensor converge in their existence spaces (we examine the Sobolev-H^s theory and the Besov-$B_2^{s,1}$ theory to reach the critical case $s = N/2$) to the corresponding Newtonian quantities. This convergence results are established in the case of "ill-prepared" data. We deduce, in the two-dimensional case, a new result concerning the global existence of weakly viscoelastic fluid flow. Our approach makes use of essentially two ingredients: the stability of the null solution of the viscoelastic fluid flow and the damping effect, on the difference between the extrastress tensor and the tensor of rate of deformation, induced by the constitutive law of the fluid.

Key words. viscoelastic fluids, global existence, Newtonian limit

AMS subject classifications. 76D03, 35B05

DOI. 10.1137/070681259

1. Introduction, main results, and notations.

In this paper we investigate the Newtonian limit of weakly viscoelastic fluid flows of Oldroyd type in $\Omega = \mathbb{R}^N$ or $\Omega = \mathbb{T}^N$.

The dynamics of homogeneous, isothermic, and incompressible fluid flows is described by the partial differential derivatives system given by

$$\left\{ \begin{array}{rcl} \rho\Big(u' + (u.\nabla)u\Big) & = & f + \operatorname{div}\sigma, \\ \operatorname{div} u & = & 0. \end{array} \right. \tag{1}$$

Here $\rho > 0$ is the (constant) density and f is the external density body forces. $u = u(t, x)$ is the velocity vector field and $\sigma = \sigma(t, x)$ is the symmetric stress tensor, which is split into two parts: $\sigma = -pId + \tau$, where $-pId$ is the spherical part ($p = p(t, x)$ the hydrodynamics pressure) and τ is the tangential part or the extrastress tensor. The fluid is called Newtonian if τ can be expressed linearly in terms of the rate of strain tensor $\mathbf{D}[u] = \frac{1}{2}(\nabla u + \nabla u^T)$, i.e.,

$$\tau = 2\eta\mathbf{D}[u], \tag{2}$$

where η is the viscosity coefficient of the fluid (in this case (1) is the Navier–Stokes system). A fluid for which (2) is not valid is called non-Newtonian or complex fluid.

Unfortunately no universal constitutive law exists for non-Newtonian fluids (see, for instance, [10]). In this paper we consider a class of fluids with memory. For this

*Received by the editors January 29, 2007; accepted for publication (in revised form) July 30, 2007; published electronically January 30, 2008. This work was partially supported by a contrat CEDRE, Université Paris-Nord, Université Libanaise.
http://www.siam.org/journals/sima/39-5/68125.html

†L.A.G.A., Institut Galilée, Université Paris-Nord, 93430 Villetaneuse, France (molinet@math.univ-paris13.fr).
‡Mathématiques, Faculté des Sciences I, Université Libanaise, Beyrouth, Lebanon (rtalhouk@ul.edu.lb). The author was partially supported by "le programme d'appui des projets de recherche de l'université Libanaise."

kind of fluid, the extrastress tensor at a time t depends on $\mathbf{D}[u]$ and its history. A model taking into account this property is the Oldroyd one. The constitutive law of Oldroyd's type [9] is given by

$$\tau + \lambda_1 \frac{\mathcal{D}_a \tau}{\mathcal{D}t} = 2\eta \left(\mathbf{D} + \lambda_2 \frac{\mathcal{D}_a \mathbf{D}}{\mathcal{D}t} \right), \tag{3}$$

where $0 \leq \lambda_2 < \lambda_1$, λ_1 is the relaxation time, and λ_2 is the retardation time. The symbol $\frac{\mathcal{D}_a}{\mathcal{D}t}$ denotes an objective (frame indifferent) tensor derivative (see [10]). More precisely,

$$\frac{\mathcal{D}_a \tau}{\mathcal{D}t} = \tau' + (u \cdot \nabla)\tau + \tau \mathbf{W} - \mathbf{W}\tau - a(\mathbf{D}\tau + \tau \mathbf{D})$$

with $\mathbf{W}[u] = \frac{1}{2}(\nabla u - \nabla u^T)$ the vorticity tensor and a a real number verifying $-1 \leq a \leq 1$. The limit case $\lambda_1 > 0$ and $\lambda_2 = 0$ corresponds to a purely elastic fluid (which is excluded in our analysis), while the limit case $\lambda_1 = \lambda_2 = 0$ corresponds to a viscous Newtonian fluid.

The constitutive law (3) is not under an evolution form. This equation can be transformed into a transport equation by splitting the extrastress tensor into two parts $\tau_s + \tau_p$, where τ_s corresponds to a Newtonian part (the solvant) and τ_p to the elastic part (the polymer). Setting $\tau_s = 2\eta(1-\omega)\mathbf{D}[u]$, with ω defined by $0 \leq \omega = 1 - \frac{\lambda_2}{\lambda_1} \leq 1$, it follows from (3) that τ_p satisfies the following transport equation:

$$\tau_p + \lambda_1 \frac{\mathcal{D}_a \tau_p}{\mathcal{D}t} = 2\eta\omega \mathbf{D}[u]. \tag{4}$$

From now on we shall denote $\tau_p = \tau$ and rewrite (1) and (4) by using dimensionless variables, and we obtain the following partial differential system:

$$\left\{ \begin{array}{rcl} \mathrm{Re}\,(u' + (u.\nabla)u) - (1-\omega)\Delta u + \nabla p &=& f + \mathrm{div}\,\tau \\ \mathrm{div}\, u &=& 0 \\ \varepsilon\Big(\tau' + (u.\nabla)\tau + \mathbf{g}(\nabla u, \tau)\Big) + \tau &=& 2\omega \mathbf{D}[u] \end{array} \right. \quad \text{in } \Omega, \tag{5}$$

where $\mathbf{g}$ is a bilinear tensor-valued mapping defined by

$$\mathbf{g}(\nabla u, \tau) = \tau \mathbf{W}[u] - \mathbf{W}[u]\tau - a(\mathbf{D}[u]\tau + \tau\, \mathbf{D}[u]),$$

$\mathrm{Re} = \rho \frac{UL}{\eta}$ and $\varepsilon = \lambda_1 \frac{U}{L}$ are, respectively, the well-known Reynolds number and the Weissenberg number (U and L represent a typical velocity and typical length of the flow). It is worth noticing that the Weissenberg number is usually denoted by We. Here, since we will make the Weissenberg number tend to zero, we prefer to denote it by ε. It is crucial to note that when $\varepsilon = 0$, (5) reduces to the incompressible Navier–Stokes system

$$\left\{ \begin{array}{rcl} \mathrm{Re}\,(v' + (v.\nabla)v) - \Delta v + \nabla p &=& f \\ \mathrm{div}\, v &=& 0 \end{array} \right. \quad \text{in } \Omega\,. \tag{6}$$

On the other hand, from the definition of the retardation parameter we observe that $\omega = 1 - \mu/\varepsilon$, where $0 \leq \mu < \varepsilon$ is given by $\mu = \lambda_2 \frac{U}{L}$. Therefore, the Newtonian limit of (5) is actually a limit with two parameters ε and μ. To simplify the study we could drop a parameter by assuming that the rate μ/ε (or equivalently ω) is constant as

ε tends to zero. In this work, instead of doing this, we will only impose a uniform upper bound on $\omega\,(= 1 - \mu/\varepsilon)$ with respect to ε.

System (5) is completed by the following initial conditions:

$$u_{|t=0} = u_0 \quad \text{and} \quad \tau_{|t=0} = \tau_0. \tag{7}$$

Our approach is quite general and uses the two following ingredients.

• The stability of the null solution of (5) for a fixed ε (see [1] on $\mathbb{R}^N$ or $\mathbb{T}^N$ and [6], [4], [8] for the case of a bounded domain).

• The damping of factor $1/\varepsilon$ on the quantity $\tau - 2\omega\mathbf{D}[u]$ induced by equation $(5)_3$.

Our results in the Sobolev spaces are valid for $\Omega = \mathbb{R}^N$ or $\mathbb{T}^N$, but to simplify the expository we will only consider $\Omega = \mathbb{R}^N$ and give the necessary modification to handle the periodic case.

The main idea is to cut u and τ in low and high frequencies at a level depending on $1/\varepsilon$. Roughly speaking, forgetting the nonlinear terms, the high frequency part of $u-v$ (v is the Newtonian solution; see (6) associated with the initial data u_0) will satisfy the homogeneous system linearized around the null solution plus a nonhomogeneous part containing a high frequency term of v. But by the Lebesgue monotone convergence theorem, this term will tend to zero in the appropriate norms. The stability of the null solution (cf. [1], [6]) will then force the high frequencies of $u - v$ and $\varepsilon^{1/2}\tau$ to remain small (recall that $(u - v)(0) = 0$). On the other hand, the remaining frequencies will tend to zero due to the damping effect on $\tau - 2\omega\mathbf{D}[u]$ which we will use in the same time as a smoothing effect. We will describe the main steps of the proof in section 1.3.

Note that our analysis is in the spirit of numerous works on the incompressible limit of compressible Navier–Stokes equations (see, for instance, [2] and the references therein). However, our analysis is in some aspects easier since there is a damping effect relating to the small parameter whereas in the incompressible limit it is a dispersive effect.

To our knowledge, no such result exists in the literature concerning our study, i.e., the Newtonian limit of non-Newtonian fluid flows. Moreover, our global existence result for regular weakly viscoelastic fluids flow in dimension two (see Corollary 1.1) is new and, in particular, not contained in the global existence results of [1].

1.1. Function spaces and notations. In what follows, C denotes a positive constant which may differ at each appearance. When writing $x \backsimeq y$ (for x and y two nonnegative real numbers), we mean that there exist C_1 and C_2 two positive constants (which do not depend on x and y) such that $C_1 x \le y \le C_2 x$. When writing $x \lesssim y$ (for x and y two nonnegative real numbers), we mean that there exists C_1 a positive constant (which does not depend on x and y) such that $x \le C_1 y$.

$\mathcal{P}$ will denote the Leray projector on solenoidal vector fields. For $1 \le p, q \le \infty$, we denote by $\|\cdot\|_{L^p}$ the usual Lebesgue norm on $\Omega = \mathbb{R}^N$,

$$\|v\|_{L^p} = \left(\int_{\mathbb{R}^N} |v|^p(x)\,dx\right)^{1/p}$$

and by $\|\cdot\|_{L^q_t L^p}$ the space-time Lebesgue norm on $]0,t[\times\Omega$,

$$\|v\|_{L^q_t L^p} = \left[\int_0^t \|v(\tau)\|^q_{L^p}\,d\tau\right]^{1/q}$$

with the obvious modification for $p, q = \infty$. For $s \in \mathbb{R}$, we denote by $\|\cdot\|_{H^s}$ the usual Sobolev norms on $\Omega = \mathbb{R}^N$,

$$\|v\|_{H^s} = \left(\int_{\mathbb{R}^N} (1+|\xi|^2)^s |\hat{v}(\xi)|^2 \, d\xi\right)^{1/2},$$

where $\hat{v}$ is the Fourier transform of v. The corresponding scalar product will be denoted by $((\cdot,\cdot))_{H^s}$. Finally, for any $\varepsilon > 0$ we introduce the following Fourier projectors:

$$\widehat{P_\varepsilon f}(\xi) = \chi_{[0,\varepsilon^\alpha]}(|\xi|)\hat{f}(\xi) \quad \text{and} \quad \widehat{Q_\varepsilon f}(\xi) = \chi_{]\varepsilon^\alpha,\infty[}(|\xi|)\hat{f}(\xi), \tag{8}$$

where $\alpha > 0$ will be specified later.

1.1.1. Homogeneous Besov spaces. Let ψ be in $\mathcal{S}(\mathbb{R})$ such that $\hat{\psi}$ is supported by the set $\{z \,/\, 2^{-1} \le |z| \le 2\}$ and such that

$$\sum_{j\in\mathbb{Z}} \hat{\psi}(2^{-j}z) = 1\,,\ z \neq 0. \tag{9}$$

Define φ by

$$\hat{\varphi} = 1 - \sum_{j\ge 1} \hat{\psi}(2^{-j}z), \tag{10}$$

and observe that $\varphi \in \mathcal{D}(\mathbb{R})$, $\hat{\varphi}$ is supported by the ball $\{z \,/\, |z| \le 2\}$, and $\hat{\varphi} = 1$ for $|z| \le 1$. We denote now by Δ_j and S_j the convolution operators on $\mathbb{R}^N$ whose symbols are, respectively, given by $\hat{\psi}(2^{-j}|\xi|)$ and $\hat{\varphi}(2^{-j}|\xi|)$ where $\xi \in \mathbb{R}^N$ and $|\xi| = \sqrt{\xi_1^2 + \cdots + \xi_N^2}$. Also we define the operator $\tilde{\Delta}_j$ by

$$\tilde{\Delta}_j = \Delta_{j-1} + \Delta_j + \Delta_{j+1}$$

which satisfies

$$\tilde{\Delta}_j \circ \Delta_j = \Delta_j.$$

For $s \in \mathbb{R}$, the homogeneous Besov space $B_2^{s,1}(\mathbb{R}^N)$ (to simplify the notation we will simply denote it by $B^s(\mathbb{R}^N)$) is the completion of $\mathcal{S}(\mathbb{R}^N)$ with respect to the seminorm

$$\|f\|_{B^s} = \big\|\{2^{js}\|\Delta_j(f)\|_{L^2}\}\big\|_{l^1(\mathbb{Z})}. \tag{11}$$

1.2. Main results.

THEOREM 1.1. *Let $N = 2, 3$ and let $(u_0, \tau_0) \in H^s(\mathbb{R}^N) \times H^s(\mathbb{R}^{N^2})$ and $f \in L^2_{\mathrm{loc}}(\mathbb{R}; H^{s-1})$ with $s > N/2$. Let v be the Newtonian solution satisfying* (6) *with initial data u_0 and let $0 < T_0 \le \infty$ such that $v \in C([0,T_0]; H^s)$. Then for any $\delta \in]0,1[$ there exists*

$$\varepsilon_0 = \varepsilon_0(N, \mathrm{Re}, \delta, \|v\|_{L^\infty_{T_0}H^s}, \|\nabla v\|_{L^2_{T_0}H^s}, \|\tau_0\|_{H^s}, \|\mathcal{P}f\|_{L^2_{T_0}H^{s-1}}) > 0$$

such that for any $0 < \varepsilon < \varepsilon_0$ the system (5)*, with*

$$0 < \omega \le 1 - \delta, \tag{12}$$

admits a unique solution

$$u_\varepsilon \in C([0,T_0];H^s), \quad \nabla u_\varepsilon \in L^2(0,T_0;H^s),\ \tau_\varepsilon \in C([0,T_0];H^s).$$

Moreover,

$$u_\varepsilon \underset{\varepsilon\to 0}{\longrightarrow} v \text{ in } C([0,T_0];H^s), \tag{13}$$

$$\tau_\varepsilon - 2\omega\mathbf{D}[u_\varepsilon] \underset{\varepsilon\to 0}{\longrightarrow} 0 \text{ in } L^2(0,T_0;H^s), \tag{14}$$

$$\varepsilon^{1/2}\tau_\varepsilon \underset{\varepsilon\to 0}{\longrightarrow} 0 \text{ in } C([0,T_0];H^s). \tag{15}$$

Recalling that in dimension two, the solution of the Newtonian problem exists for all positive times, we deduce the following result.

COROLLARY 1.1. *In dimension two there exists*

$$\varepsilon_0 = \varepsilon_0(\mathrm{Re},\delta,\|v\|_{L^\infty_\infty H^s},\|\nabla v\|_{L^2_\infty H^s},\|\tau_0\|_{H^s},\|\mathcal{P}f\|_{L^2_\infty H^{s-1}}) > 0$$

such that for any $0<\varepsilon<\varepsilon_0$ *the solution of* (5) *given by Theorem* 1.1 *exists for all positive times.*

Remark 1.1. Note that Theorem 1.1 is a convergence result for "ill-prepared" data. Indeed the quantity $\tau_0 - 2\omega\mathbf{D}[u_0]$ is not assumed to be small with ε. Moreover, this is a singular limit result since τ and $\mathbf{D}[u]$ do not belong to the same function space. In particular, $\mathbf{D}[u_0]$ is not as the same level of Sobolev regularity as τ_0.

Remark 1.2. According to the introduction, the Newtonian limit process is actually a limit process with two parameters ε and μ tending to zero with $0\le\mu<\varepsilon$. The assumption (12) of Theorem 1.1 means that we impose the following additional conditions on the rate μ/ε as (ε,μ) tends to zero $(0,0)$:

$$\delta \le \frac{\mu}{\varepsilon} = \frac{\lambda_2}{\lambda_1} < 1$$

for some fixed $1>\delta>0$.

As mentioned in the introduction, the use of Besov spaces enables us to reach the critical index $s=N/2$.

THEOREM 1.2. *Let* $N=2,3$ *and let* $(u_0,\tau_0)\in B^{N/2-1}(\mathbb{R}^N)\times B^{N/2}(\mathbb{R}^{N^2})$ *and* $f\in L^1_{\mathrm{loc}}(B^{N/2-1})$. *Let* v *be the Newtonian solution satisfying* (6) *with initial data* u_0 *and let* $0<T_0\le\infty$ *such that* $v\in C([0,T_0];B^{N/2-1})$. *There exist* $0<\omega_0<1$ *and* $\varepsilon_0=\varepsilon_0(N,\mathrm{Re},\omega_0,\|\tau_0\|_{B^{N/2}},\mathcal{P}f,u_0)>0$ *such that for any* $0<\varepsilon<\varepsilon_0$ *the system* (5), *with* $0<\omega\le\omega_0$, *admits a unique solution*

$$u_\varepsilon \in C([0,T_0];B^{N/2-1}), \quad u_\varepsilon \in L^1(0,T_0;B^{N/2+1}),\ \tau_\varepsilon \in C([0,T_0];B^{N/2}).$$

Moreover,

$$u_\varepsilon \underset{\varepsilon\to 0}{\longrightarrow} v \text{ in } C([0,T_0];B^{N/2-1}), \tag{16}$$

$$\tau_\varepsilon - 2\omega\mathbf{D}[u_\varepsilon] \underset{\varepsilon\to 0}{\longrightarrow} 0 \text{ in } L^1(0,T_0;B^{N/2}), \tag{17}$$

$$\varepsilon^{1/2}\tau_\varepsilon \underset{\varepsilon\to 0}{\longrightarrow} 0 \text{ in } C([0,T_0];B^{N/2})\,. \tag{18}$$

In dimension two, using the classical global existence result in B^0 for the Newtonian problem (see, for instance, [2]), we get the following corollary.

Corollary 1.2. *In dimension two there exists*

$$\varepsilon_0 = \varepsilon_0(\mathrm{Re}, \omega_0, \|v\|_{L^\infty_\infty B^0}, \|\nabla v\|_{L^1_\infty B^2}, \|\tau_0\|_{B^1}, \|\mathcal{P}f\|_{L^1_{T_0}B^0})$$

such that for any $0<\varepsilon<\varepsilon_0$ *the solution of* (5) *given by Theorem* 1.2 *exists for all positive times.*

In the surcritical case, $s>N/2$, we get similar results by considering nonhomogeneous Besov spaces. Note that ε_0 then depends explicitly on some norms of v.

Theorem 1.3. *For* $s>N/2$*, Theorem* 1.2 *and Corollary* 1.2 *also hold by replacing the function spaces* $B^{N/2-1}$ *by* $B^{s-1}\cap B^{N/2-1}$ *and* $B^{N/2}$ *by* $B^s\cap B^{N/2}$*. Moreover,* ε_0 *will depend now explicitly on some norms of* v *and* $\mathcal{P}f$*. More precisely, for* $s>N/2$*, we have*

$$\varepsilon_0 = \varepsilon_0(N, \mathrm{Re}, \omega_0, \|v\|_{L^\infty_{T_0}B^{N/2-1}}, \|\nabla v\|_{L^1_{T_0}B^s}, \|\tau_0\|_{B^{N/2}}, \|\mathcal{P}f\|_{L^1_{T_0}B^{N/2-1}}).$$

1.3. Sketch of the proof of Theorem 1.1. In this subsection we want to explain the main steps of the proof of Theorem 1.1. Note that Theorem 1.2 follows from the same arguments. To simplify we drop the nonlinear terms in (5). The first step consists in noticing that $W := u - v$ satisfies the following system:

$$\begin{cases} \mathrm{Re}\, W_t - (1-\omega)Q_\varepsilon\Delta W - P_\varepsilon\Delta W = P_\varepsilon\mathcal{P}(\mathrm{div}\,\tau - \omega\Delta u) \\ \qquad\qquad\qquad\qquad -\omega Q_\varepsilon\Delta v + Q_\varepsilon\mathcal{P}\mathrm{div}\,\tau, \\ \mathrm{div}\, W = 0, \\ \varepsilon\tau_t + Q_\varepsilon\tau = 2\omega Q_\varepsilon\mathbf{D}[W] + 2\omega Q_\varepsilon\mathbf{D}[v] - P_\varepsilon(\tau - 2\omega\mathbf{D}[u]), \end{cases} \tag{19}$$

where P_ε and Q_ε are the projectors on, respectively, the low and the high frequencies defined in (8).

Projecting on the high frequencies with Q_ε (see (8) for the definition), proceeding as in [1], it is easy to check that we get a differential inequality close to

$$\frac{d}{dt}\Big(\|Q_\varepsilon W\|^2_{H^s} + \varepsilon\|Q_\varepsilon\tau\|^2_{H^s}\Big) + \|Q_\varepsilon\nabla W\|^2_{H^s} + \|Q_\varepsilon\tau\|^2_{H^s} \lesssim \|Q_\varepsilon\nabla v\|^2_{H^s},$$

where we drop all the constants to clarify the presentation. Therefore, since $W(0)=0$, $\varepsilon\to 0$, and, by the Lebesgue monotone convergence theorem, $\|Q_\varepsilon\nabla v\|_{L^2_{T_0}H^s}\to 0$, we infer that $\|Q_\varepsilon W\|_{L^\infty_{T_0}H^s}$ goes to zero with ε. Now, to treat the low frequency part, we observe that, computing $P_\varepsilon(19)_3 - \frac{2\omega}{\mathrm{Re}}\mathbf{D}[(19)_1]$ and taking the H^{s-1}-scalar product of the resulting equation with $Z := \tau - 2\omega\mathbf{D}[u]$, we obtain something like

$$\frac{d}{dt}\|P_\varepsilon Z\|^2_{H^{s-1}} + \frac{1}{\varepsilon}\|P_\varepsilon Z\|^2_{H^{s-1}} \lesssim \|P_\varepsilon\tau\|^2_{H^s} + \|P_\varepsilon f\|^2_{H^{s-1}}. \tag{20}$$

On the other hand, $P_\varepsilon(19)_3$ can be rewritten as

$$\varepsilon P_\varepsilon\tau_t + \varepsilon^\beta P_\varepsilon\tau = 2\omega\varepsilon^\beta P_\varepsilon\mathbf{D}[W] + 2\omega\varepsilon^\beta P_\varepsilon\mathbf{D}[v] - (1-\varepsilon^\beta)P_\varepsilon Z,$$

where $0 < \beta < 1$ will be specified later. Therefore, taking the H^s-scalar product of this last equality with τ and adding with the scalar product of $(19)_1$ with W, we get a differential inequality close to

$$\frac{d}{dt}\Big(\|P_\varepsilon W\|_{H^s}^2+\varepsilon\|P_\varepsilon\tau\|_{H^s}^2\Big)+\|P_\varepsilon\nabla W\|_{H^s}^2+\varepsilon^\beta\|P_\varepsilon\tau\|_{H^s}^2\lesssim\varepsilon^{-\beta}\|P_\varepsilon Z\|_{H^s}^2+\varepsilon^\beta\|P_\varepsilon\nabla v\|_{H^s}^2.$$

Adding this last inequality and $\varepsilon^{2\beta}(20)$ we finally obtain

$$\begin{aligned}\frac{d}{dt}\Big(\|P_\varepsilon W\|_{H^s}^2 + \varepsilon\|P_\varepsilon\tau\|_{H^s}^2\Big) + \|P_\varepsilon\nabla W\|_{H^s}^2 + \varepsilon^\beta\|P_\varepsilon\tau\|_{H^s}^2 + \varepsilon^{2\beta-1}\|P_\varepsilon Z\|_{H^{s-1}}^2 \\ \lesssim \varepsilon^{2\beta}\|P_\varepsilon f\|_{H^{s-1}}^2 + \varepsilon^\beta\|P_\varepsilon\nabla v\|_{H^s}^2\end{aligned}$$

since $\varepsilon^{-\beta}\|P_\varepsilon Z\|_{H^s}^2 \le \varepsilon^{-\beta}\varepsilon^{-2\alpha}\|P_\varepsilon Z\|_{H^{s-1}}^2 \le \frac{\varepsilon^{2\beta-1}}{4}\|P_\varepsilon Z\|_{H^{s-1}}^2$ as soon as $1-3\beta-2\alpha > 0$. This last inequality enables us to conclude for the low frequency part. Note that we used the damping effect also as a smoothing effect.

2. Proof of Theorem 1.1. Let us recall the following existence theorem proven by Chemin and Masmoudi [1].

THEOREM 2.1. *Let $(u_0,\tau_0) \in H^s(\mathbb{R}^N) \times H^s(\mathbb{R}^{N^2})$ with $s > N/2$. Then there exists a unique positive maximal time T^* and a unique solution*

$$(u,\tau) \in C([0,T^*[;H^s) \cap L^2_{\mathrm{loc}}(0,T^*;H^{s+1}) \times C([0,T^*[;H^s).$$

Moreover, if $T^* < \infty$, then for all $N/2 < s' \le s$

$$\limsup_{t\nearrow T^*}\Big(\|u(t)\|_{H^{s'}} + \|\tau(t)\|_{H^{s'}}\Big) = +\infty. \tag{21}$$

Remark 2.1. Actually in [1] the following sharper blowup condition is derived:

$$T^* < \infty \Longrightarrow \int_0^{T^*} \|\nabla u(t)\|_{L^\infty} + \|\tau(t)\|_{L^\infty}^2\,dt = +\infty,$$

but for our purpose the classical blowup condition (21) will be sufficient.

Let us also recall a commutator estimate and classical Leibniz rules for fractional derivatives.

LEMMA 2.2. *Let Δ be the Laplace operator on $\mathbb{R}^N$, $N \ge 1$. Denote by J^s the operator $(1-\Delta)^{s/2}$.*

- *For every $s > N/2$,*

$$\|[J^s,f]g\|_{L^2(\mathbb{R}^N)} \lesssim \|\nabla f\|_{H^s(\mathbb{R}^N)}\|g\|_{H^{s-1}(\mathbb{R}^N)}. \tag{22}$$

- *For every $s > 0$, $1 < q,q' \le \infty$, and $1 < r,p,p' < \infty$ with $1/p+1/q = 1/p'+1/q' = 1/r$,*

$$\|J^s(fg)\|_{L^r(\mathbb{R}^N)} \lesssim \|J^s f\|_{L^p(\mathbb{R}^N)}\|g\|_{L^q(\mathbb{R}^N)} + \|f\|_{L^{q'}(\mathbb{R}^N)}\|J^s g\|_{L^{p'}(\mathbb{R}^N)}. \tag{23}$$

- *For every p,r,t such that $r,p \ge t$ and $r+p-t > N/2$,*

$$\|fg\|_{H^t(\mathbb{R}^N)} \lesssim \|f\|_{H^p(\mathbb{R}^N)}\|g\|_{H^r(\mathbb{R}^N)}. \tag{24}$$

Proof. Estimates (23) and (24) are classical and can be found in [7] and [5]. Estimate (22) is a variant of Kato–Ponce's commutator estimates. It is proven in [12] in dimension one but the proof works also in dimensions two and three. ∎

To treat some nonlinear terms in dimension two we will need, moreover, the following Gagliardo–Nirenberg-type inequality (see, for instance, [3]).

LEMMA 2.3. *Let $N \geq 2$, for $u \in H^1(\mathbb{R}^N)$ the following Sobolev-type inequality holds for any $2 \leq p < +\infty$ such that $\frac{1}{2} - \frac{1}{N} \leq \frac{1}{p}$:*

$$\|u\|_{L^p(\mathbb{R}^N)} \lesssim \|u\|_{L^2(\mathbb{R}^N)}^{(\frac{N}{p}-\frac{N}{2}+1)} \|\nabla u\|_{L^2(\mathbb{R}^N)}^{(\frac{N}{2}-\frac{N}{p})}. \tag{25}$$

2.1. Estimate on $W = u - v$ and $\varepsilon^{1/2}\tau$. We start by deriving a differential inequality for the H^s-norms of W and $\varepsilon^{1/2}\tau$. The high frequency part of this inequality is directly inspired by the stability proof of the null solution in [1]. This will enable us to control the very high frequency part $(Q_\varepsilon u, Q_\varepsilon \tau)$ of the solution. The other part $(P_\varepsilon u, P_\varepsilon \tau)$ will be treated by using the damping effect.

For $\varepsilon > 0$ fixed, Theorem 2.1 gives the existence and uniqueness of the solution $(u_\varepsilon, \tau_\varepsilon)$ of (5) in $C([0, T_\varepsilon^*[; H^s) \cap L^2_{\mathrm{loc}}(0, T_\varepsilon^*; H^{s+1}) \times C([0, T_\varepsilon^*[; H^s)$ for some $T_\varepsilon^* > 0$. To simplify the notations, we drop the index ε on u and τ in what follows. Setting

$$Z = \tau - 2\omega \mathbf{D}[u]$$

we have the following estimates.

LEMMA 2.4. *For $\varepsilon > 0$ small enough, the solution (u, τ) of (5) satisfies, for all $0 < t < T_\varepsilon^*$ and $0 < \beta < 1$,*

$$\begin{aligned}
&\frac{d}{dt}\Big(\frac{\mathrm{Re}}{2}\|W\|_{H^s}^2 + \frac{\varepsilon}{4\omega}\|\tau\|_{H^s}^2\Big) + \frac{1}{4}\|P_\varepsilon \nabla W\|_{H^s}^2 + \frac{(1-\omega)}{2}\|Q_\varepsilon \nabla W\|_{H^s}^2 \\
&\quad + \frac{1}{4\omega}\|Q_\varepsilon \tau\|_{H^s}^2 + \frac{\varepsilon^\beta}{4\omega}\|P_\varepsilon \tau\|_{H^s}^2 \\
&\quad \leq \left(1 + \frac{4\varepsilon^{-\beta}}{\omega}\right)\|P_\varepsilon Z\|_{H^s}^2 + 4\omega\|Q_\varepsilon \nabla v\|_{H^s}^2 + 8\omega\varepsilon^\beta\|P_\varepsilon \nabla v\|_{H^s}^2 \\
&\quad + \frac{C\,\mathrm{Re}}{(1-\omega)^2}(\|\nabla u\|_{H^s}^2 + \|\nabla v\|_{H^s}^2)\|W\|_{H^s}^2 + \frac{C}{\omega}\,\varepsilon^{2-\beta}\|\nabla u\|_{H^s}^2\|\tau\|_{H^s}^2
\end{aligned} \tag{26}$$

whenever $0 < \omega < 1$. Moreover, for $0 < \omega \leq 10^{-2}$, it holds that

$$\begin{aligned}
&\frac{d}{dt}\Big(\frac{\mathrm{Re}}{4}\|W\|_{H^s}^2 + \frac{\varepsilon}{2}\|\tau\|_{H^s}^2\Big) + \frac{1}{8}\|P_\varepsilon \nabla W)\|_{H^s}^2 + \frac{(1-\omega)}{4}\|Q_\varepsilon \nabla W\|_{H^s}^2 \\
&\quad + \frac{1}{4}\|Q_\varepsilon \tau\|_{H^s}^2 + \frac{\varepsilon^\beta}{4}\|P_\varepsilon \tau\|_{H^s}^2 \\
&\quad \leq (1 + 4\varepsilon^{-\beta})\|P_\varepsilon Z\|_{H^s}^2 + 8\omega^2\|Q_\varepsilon \nabla v\|_{H^s}^2 + 8\omega^2\varepsilon^\beta\|P_\varepsilon \nabla v\|_{H^s}^2 \\
&\quad + \frac{C\ \mathrm{Re}}{(1-\omega)^2}(\|\nabla u\|_{H^s}^2 + \|\nabla v\|_{H^s}^2)\|W\|_{H^s}^2 + C\,\varepsilon^{2-\beta}\|\nabla u\|_{H^s}^2\|\tau\|_{H^s}^2.
\end{aligned} \tag{27}$$

Proof. Notice that W verifies the equation

$$\begin{aligned}
\mathrm{Re}\Big(W_t + \mathcal{P}(u.\nabla)W\Big) - \Delta W &= \mathcal{P}\mathrm{div}\,\tau - \omega\Delta u - \ \mathrm{Re}\ \mathcal{P}(W.\nabla)v \\
&= P_\varepsilon \mathcal{P}\Big(\mathrm{div}\,\tau - \omega\Delta u\Big) - \omega Q_\varepsilon \Delta v \\
&\quad + Q_\varepsilon \mathcal{P}\mathrm{div}\,\tau - \omega Q_\varepsilon \Delta W - \ \mathrm{Re}\ \mathcal{P}(W.\nabla)v.
\end{aligned} \tag{28}$$

Therefore, multiplying scalarly (28) by W in $H^s(\mathbb{R}^N)$, using the Cauchy–Schwarz inequality, Lemma 2.2, and that u is divergence free, we obtain

$$\begin{aligned}\frac{1}{2}\mathrm{Re}\frac{d}{dt}\|W\|^2_{H^s} &+ \|P_\varepsilon \nabla W\|^2_{H^s} + (1-\omega)\|Q_\varepsilon \nabla W\|^2_{H^s}\\ &\leq ((Q_\varepsilon \mathrm{div}\,\tau, W))_{H^s} + \|P_\varepsilon Z\|_{H^s}\|\nabla W\|_{H^s} + \omega\|Q_\varepsilon\nabla v\|_{H^s}\|Q_\varepsilon \nabla W\|_{H^s}\\ &\quad + C\ \mathrm{Re}\left|((J^s(u.\nabla)W, J^sW))_{L^2}\right| +\ \mathrm{Re}\left|((J^s(W.\nabla)v, J^sW))_{L^2}\right|.\end{aligned}\tag{29}$$

To estimate the second to the last term of (29), we rewrite it with the help of a commutator and apply the Cauchy–Schwarz inequality to the term containing this commutator to get

$$\left|((J^s(u.\nabla)W, J^sW))_{L^2}\right| \leq \left|(((u.\nabla)J^sW, J^sW))\right| + \left\|[J^s,(u.\nabla)]W\right\|_{L^2}\|J^sW\|_{L^2}.$$

Since u is divergence free, the first term of the right-hand side of this last inequality cancels by integration by parts. Estimating the second term, thanks to Lemma 2.2, we then obtain

$$\left|((J^s(u.\nabla)W, J^sW))_{L^2}\right| \leq C\,\|\nabla u\|_{H^s}\|\nabla W\|_{H^s}\|W\|_{H^s}.$$

Now, to estimate the last term of the right-hand side of (29) we have to distinguish the cases $N = 2$ and $N = 3$.

• $N = 3$. Then by Lemma 2.2 and the Hölder, Sobolev, and Young inequalities, we get

$$\begin{aligned}\left|((J^s(W.\nabla)v, J^sW))_{L^2}\right| &\leq \|J^s(W.\nabla)v\|_{L^{6/5}}\|J^sW\|_{L^6}\\ &\lesssim \Big(\|J^sW\|_{L^2}\|\nabla v\|_{L^3} + \|W\|_{L^3}\|J^s\nabla v\|_{L^2}\Big)\|J^s\nabla W\|_{L^2}\\ &\lesssim \|W\|_{H^s}\|\nabla v\|_{H^s}\|\nabla W\|_{H^s}.\end{aligned}$$

• $N = 2$. In this case, using Hölder's inequality and Lemmas 2.2 and 2.3 we infer that

$$\begin{aligned}\left|((J^s(W.\nabla)v), J^sW))_{L^2}\right| &\leq \|J^s(W.\nabla)v\|_{L^{3/2}}\|J^sW\|_{L^3}\\ &\lesssim \Big(\|J^sW\|_{L^6}\|\nabla v\|_{L^2} + \|W\|_{L^6}\|J^s\nabla v\|_{L^2}\Big)\|J^sW\|_{L^3}\\ &\lesssim \|J^sW\|_{L^6}\|J^s\nabla v\|_{L^2}\|J^sW\|_{L^3}\\ &\lesssim \|W\|_{H^s}\|\nabla v\|_{H^s}\|\nabla W\|_{H^s}.\end{aligned}$$

By Young's inequalities it thus follows from (29) that

$$\begin{aligned}\frac{\mathrm{Re}}{2}\frac{d}{dt}\|W\|^2_{H^s} &+ \frac{3}{4}\|P_\varepsilon\nabla W\|^2_{H^s} + \frac{(1-\omega)}{2}\|Q_{>\varepsilon}\nabla W\|^2_{H^s}\\ &\leq ((Q_\varepsilon \mathrm{div}\,\tau, W))_{H^s} + \|P_\varepsilon Z\|^2_{H^s} + \frac{\omega^2}{4(1-\omega)}\|Q_\varepsilon\nabla v\|^2_{H^s}\\ &\quad + \frac{C\ \mathrm{Re}}{(1-\omega)^2}\Big(\|\nabla u\|^2_{H^s} + \|\nabla v\|^2_{H^s}\Big)\|W\|^2_{H^s}.\end{aligned}\tag{30}$$

On the other hand, for $0 < \beta < 1$, observing that

$$\begin{aligned}\tau - 2\omega\mathbf{D}[u] &= Q_\varepsilon\tau - 2\omega Q_\varepsilon(\mathbf{D}[W] + \mathbf{D}[v])\\ &\quad + (1-\varepsilon^\beta)P_\varepsilon Z + \varepsilon^\beta\Big(P_\varepsilon\tau - 2\omega P_\varepsilon(\mathbf{D}[W] + \mathbf{D}[v])\Big),\end{aligned}$$

we deduce from (5) that τ satisfies the equation

$$\varepsilon\Big(\tau_t + (u.\nabla)\tau + g(\nabla u,\tau)\Big) + Q_\varepsilon\tau + \varepsilon^\beta P_\varepsilon\tau = 2\omega Q_\varepsilon\mathbf{D}[W] + 2\omega Q_\varepsilon\mathbf{D}[v] \\ +2\omega\varepsilon^\beta P_\varepsilon\mathbf{D}[W] + 2\omega\varepsilon^\beta P_\varepsilon\mathbf{D}[v] - (1-\varepsilon^\beta)P_\varepsilon Z.$$

Taking the H^s-scalar product of this equation with τ and using Lemma 2.2 and the Cauchy–Schwarz and Young inequalities we get

$$\frac{\varepsilon}{2}\frac{d}{dt}\|\tau\|^2_{H^s} + \frac{1}{2}\|Q_\varepsilon\tau\|^2_{H^s} + \frac{\varepsilon^\beta}{2}\|P_\varepsilon\tau\|^2_{H^s} \le 2\omega((Q_\varepsilon\mathbf{D}[W],\tau))_{H^s} + 8\varepsilon^{-\beta}\|P_\varepsilon Z\|^2_{H^s} \\ +8\omega^2\|Q_\varepsilon\nabla v\|^2_{H^s} + 8\omega^2\varepsilon^\beta(\|P_\varepsilon\nabla W\|^2_{H^s} + \|P_\varepsilon\nabla v\|^2_{H^s}) \\ +C\,\varepsilon^{2-\beta}\|\nabla u\|^2_{H^s}\|\tau\|^2_{H^s}. \tag{31}$$

We now separate the two cases.

• $\omega \neq 0$. Then, adding (30) and (31)/2ω we notice that the first term in the right-hand side of (30) and (31) cancel each other and (26) follows. This gives (26) for ε small enough since $\beta > 0$.

• $0 < \omega \le 10^{-2}$. Then adding (30)/2+(31), estimating the two remaining H^s-scalar products by integration by parts and using the Cauchy–Schwarz and Young inequalities, one obtains (27). □

2.2. Estimate on $Z = \tau - 2\omega\mathbf{D}[u]$. We will now take advantage of the damping effect on $Z = \tau - 2\omega\mathbf{D}[u]$.

LEMMA 2.5. *The solution (u,τ) of (5) satisfies, for all ε small enough and $0 < t < T^*_\varepsilon$,*

$$\frac{1}{2}\frac{d}{dt}\|Z\|^2_{H^{s-1}} + \frac{1}{2\varepsilon}\|Z\|^2_{H^{s-1}} \le \frac{4\omega}{\mathrm{Re}\,(1-\omega)}\|\mathcal{P}f\|^2_{H^{s-1}} + \frac{(1+\omega)^2}{\mathrm{Re}\,(1-\omega)}\|\tau\|^2_{H^s} \\ +\frac{4}{1-\omega}(\mathrm{Re}\|\nabla u\|^2_{H^s} + \|\tau\|^2_{H^s})\|u\|^2_{H^s}. \tag{32}$$

Proof. We apply $\frac{2\omega}{\mathrm{Re}}\mathbf{D}[\cdot]$ to $(5)_1$ and substract the resulting equation from $(5)_2$ to obtain

$$Z_t - \frac{(1-\omega)}{\mathrm{Re}}\Delta Z + \frac{1}{\varepsilon}Z = -f_1 - f_2, \tag{33}$$

where

$$f_1 = \frac{2\omega}{\mathrm{Re}}\mathbf{D}[\mathcal{P}\mathrm{div}\,\tau] - \frac{(1-\omega)}{\mathrm{Re}}\Delta\tau + \frac{2\omega}{\mathrm{Re}}\mathbf{D}[\mathcal{P}f] - 2\omega\mathbf{D}[\mathcal{P}(u.\nabla)u]$$

and

$$f_2 = \mathcal{P}(u.\nabla)\tau + g(\nabla u,\tau).$$

Taking the H^{s-1}-scalar product of (33) with Z we get

$$\frac{1}{2}\frac{d}{dt}\|Z\|^2_{H^{s-1}} + \frac{(1-\omega)}{4\ \mathrm{Re}}\|\nabla Z\|^2_{H^{s-1}} + \frac{1}{\varepsilon}\|Z\|^2_{H^{s-1}} \\ \le C\,\frac{(1+\omega)^2}{\mathrm{Re}\,(1-\omega)}\|\tau\|^2_{H^s} + \frac{4\omega}{\mathrm{Re}\,(1-\omega)}\|\mathcal{P}f\|^2_{H^{s-1}} \\ +\frac{4\omega\ \mathrm{Re}}{1-\omega}\|(u.\nabla)u\|^2_{H^{s-1}} + 4\|(u.\nabla)\tau\|^2_{H^{s-1}} + \|\mathbf{g}(\nabla u,\tau)\|^2_{H^{s-1}}, \tag{34}$$

where we used that

$$\begin{aligned}\frac{2\omega}{\mathrm{Re}}\Big|((\mathbf{D}[\mathcal{P}\mathrm{div}\,\tau], Z))_{H^{s-1}}\Big| &\le C\frac{2\omega}{\mathrm{Re}}\|\mathrm{div}\,\tau\|_{H^{s-1}}\|\nabla Z\|_{H^{s-1}}\\ &\le \frac{1-\omega}{8\ \mathrm{Re}}\|\nabla Z\|^2_{H^{s-1}} + C\,\frac{\omega^2}{\mathrm{Re}\,(1-\omega)}\|\tau\|^2_{H^s}.\end{aligned}$$

Finally to control the nonlinear terms we notice that, thanks to (24),

$$\|a.\nabla b\|_{H^{s-1}} \lesssim \|a\|_{H^s}\|\nabla b\|_{H^{s-1}}$$

which concludes the proof of (32). $\square$

2.3. Convergence to the Newtonian flow. We give here the proof in the case $10^{-2} \le \omega \le 1-\delta$. The case $0 < \omega \le 10^{-2}$ is simpler and can be handled in the same way by using (27) instead of (26).

Adding (26) and $\varepsilon^{2\beta}$(32), we obtain for ε small enough

$$\begin{aligned}&\frac{d}{dt}\left(\frac{\mathrm{Re}}{2}\|W\|^2_{H^s} + \frac{\varepsilon}{4\omega}\|\tau\|^2_{H^s} + \frac{\varepsilon^{2\beta}}{2}\|Z\|^2_{H^{s-1}}\right)\\ &\quad + \frac{(1-\omega)}{4}\|\nabla W\|^2_{H^s} + \frac{1}{8\omega}(\|Q_\varepsilon\tau\|^2_{H^s} + \varepsilon^\beta\|P_\varepsilon\tau\|^2_{H^s}) + \frac{\varepsilon^{2\beta-1}}{4}\|Z\|^2_{H^{s-1}}\\ &\quad \le 8\omega^2\|Q_\varepsilon\nabla v\|^2_{H^s} + 8\omega^2\varepsilon^\beta\|P_\varepsilon\nabla v\|^2_{H^s} + C\,\varepsilon^{1-\beta}\|\nabla u\|^2_{H^s}\,\frac{\varepsilon}{4\omega}\|\tau\|^2_{H^s}\\ &\quad\quad + \frac{C\ \mathrm{Re}}{(1-\omega)^2}(\|\nabla v\|^2_{H^s} + \|\nabla u\|^2_{H^s})\|W\|^2_{H^s}\\ &\quad\quad + C\frac{\varepsilon^{2\beta}}{(1-\omega)}\left[\frac{1}{\mathrm{Re}}\|\mathcal{P}f\|^2_{H^{s-1}} + (\mathrm{Re}\|\nabla u\|^2_{H^s} + \|\tau\|^2_{H^s})\|u\|^2_{H^s}\right].\end{aligned}\tag{35}$$

Here, we used that for ε small enough,

$$\frac{(1+\omega)^2}{\mathrm{Re}\,(1-\omega)}\varepsilon^{2\beta}\|\tau\|^2_{H^s} \le \frac{\varepsilon^\beta}{8\omega}\|\tau\|^2_{H^s}$$

and

$$\left(1 + \frac{4\varepsilon^{-\beta}}{\omega}\right)\|P_\varepsilon Z\|^2_{H^s} \le (1 + 20^2\varepsilon^{-\beta})\varepsilon^{-2\alpha}\|P_\varepsilon Z\|^2_{H^{s-1}} \le \frac{\varepsilon^{2\beta-1}}{4}\|Z\|^2_{H^{s-1}}$$

as soon as $\beta > 0$ and $1 - 3\beta - 2\alpha > 0$.

From now on to simplify we thus take $(\alpha, \beta) = (1/8, 1/8)$. Setting

$$\begin{aligned}X_s(t) &= \frac{\mathrm{Re}}{2}\|W(t)\|^2_{H^s} + \frac{\varepsilon}{4\omega}\|\tau(t)\|^2_{H^s} + \frac{\varepsilon^{2\beta}}{2}\|Z\|^2_{H^{s-1}}\\ &\quad + \int_0^t \frac{(1-\omega)}{4}\|\nabla W\|^2_{H^s} + \frac{1}{8}\|Q_\varepsilon\tau\|^2_{H^s} + \frac{\varepsilon^\beta}{8}\|P_\varepsilon\tau\|^2_{H^s} + \frac{\varepsilon^{2\beta-1}}{4}\|Z\|^2_{H^{s-1}}\,ds,\end{aligned}$$

we infer that X_s satisfies the following differential inequality:

$$\begin{aligned}\frac{d}{dt}X_s &\le 8\omega^2\|Q_\varepsilon\nabla v\|^2_{H^s} + 8\omega^2\varepsilon^\beta\|P_\varepsilon\nabla v\|^2_{H^s} + C\frac{\varepsilon^{2\beta}}{\mathrm{Re}\,(1-\omega)}\|\mathcal{P}f\|^2_{H^{s-1}}\\ &\quad + C(\ \mathrm{Re}\,,\delta)\Big[\varepsilon^{2\beta}\|\tau\|^2_{H^s} + \|\nabla u\|^2_{H^s} + \|\nabla v\|^2_{H^s}\Big]X_s\\ &\quad + C\frac{\varepsilon^\beta}{(1-\omega)}(\mathrm{Re}\varepsilon^\beta\|\nabla u\|^2_{H^s} + \varepsilon^\beta\|\tau\|^2_{H^s})\|v\|^2_{H^s},\end{aligned}\tag{36}$$

where we rewrite u as $W+v$ and use the triangle inequality when necessary. Hence, Gronwall's inequality leads to

$$\begin{aligned} X_s(t) \le \exp\Big[C(\,\mathrm{Re}\,,\delta)\Big(\varepsilon^{2\beta}\|\tau\|^2_{L^2_tH^s} + \|\nabla u\|^2_{L^2_tH^s} + \|\nabla v\|^2_{L^2_tH^s}\Big)\Big] \\ \Big[X_s(0) + 8\omega^2\|Q_\varepsilon \nabla v\|^2_{L^2_tH^s} + 8\varepsilon^\beta\|\nabla v\|^2_{L^2_tH^s} + C\frac{\varepsilon^{2\beta}}{\mathrm{Re}\,\delta}\|\mathcal{P}f\|^2_{L^2_tH^{s-1}} \\ +\frac{\varepsilon^\beta}{\delta}(\mathrm{Re}\,\varepsilon^\beta\|\nabla u\|^2_{L^2_tH^s} + \varepsilon^\beta\|\tau\|^2_{L^2_tH^s})\|v\|^2_{L^\infty_tH^s}\Big]. \end{aligned} \tag{37}$$

Rewriting u as $v+W$, we finally obtain

$$\begin{aligned} X_s(t) \le \exp\Big[C(\,\mathrm{Re}\,,\delta)\Big(\|\nabla v\|^2_{L^2_tH^s} + X_s(t)\Big)\Big] \\ \Big[X_s(0) + 8\omega^2\|Q_\varepsilon \nabla v\|^2_{L^2_tH^s} + 8\varepsilon^\beta\|\nabla v\|^2_{L^2_tH^s} + C\frac{\varepsilon^{2\beta}}{\mathrm{Re}\ \delta}\|\mathcal{P}f\|^2_{L^2_tH^{s-1}} \\ + C\frac{(1+\mathrm{Re})\varepsilon^\beta}{\delta}\|v\|^2_{L^\infty_tH^s}X_s(t) + C\frac{\mathrm{Re}\,\varepsilon^{2\beta}}{\delta}\|v\|^2_{L^\infty_tH^s}\|\nabla v\|^2_{L^2_tH^s}\Big], \end{aligned} \tag{38}$$

where

$$X_s(0) = \frac{\varepsilon}{4\omega}\|\tau_0\|^2_{H^s} + \frac{\varepsilon^{2\beta}}{2}\|\tau_0 - 2\omega\mathbf{D}[u_0]\|^2_{H^{s-2}}. \tag{39}$$

Let us now assume that $T^*_\varepsilon \le T_0$. Since (38) holds for any $N/2 < s' < s$, noticing that

$$\|Q_\varepsilon \nabla v\|_{L^2_tH^{s'}} \le \varepsilon^{\alpha(s-s')}\|\nabla v\|_{L^2_tH^s},$$

we deduce from (38), (39), and the continuity of $t \mapsto X_{s'}(t)$ that there exists

$$\varepsilon_0(s, \|\tau_0\|_{H^s}, \|u_0\|_{H^s}, \|\nabla v\|_{L^2_{T_0}H^s}, \|v\|_{L^\infty_{T_0}H^s}) > 0$$

such that for any $0<\varepsilon<\varepsilon_0$ and any $0<t<T^*_\varepsilon$,

$$X_{s'}(t) \le C\,\varepsilon^{\min(\beta, 2\alpha(s-s'))} \le C\varepsilon^{\min(1,s-s')/8} \tag{40}$$

which contradicts (21) of Theorem 2.1. This ensures that $T^*_\varepsilon > T_0$. Now, since by Lebesgue monotone convergence theorem

$$\|Q_\varepsilon \nabla v\|_{L^2_{T_0}H^s} \underset{\varepsilon\to 0}{\longrightarrow} 0, \tag{41}$$

it follows from (38) and (39) that $X_s(T_0) \to 0$ as $\varepsilon \to 0$. This proves (13) and (15). To prove (14) we observe that from this last limit and (41), $\|Q_\varepsilon(\tau - 2\omega\mathbf{D}[u])\|_{L^2_{T_0}H^s} \to 0$ and $\varepsilon^{2\beta-1}\|P_\varepsilon(\tau - 2\omega\mathbf{D}[u])\|^2_{L^2_{T_0}H^{s-1}} \to 0$. This yields the result by Bernstein inequality since $2\beta - 1 + 2\alpha < 0$.

2.4. The periodic setting. Let us give here the modifications needed to handle the case $\Omega = \mathbb{T}^N$, $N = 2,3$. It is worth noticing that Lemma 2.2 holds also with $\Omega = \mathbb{T}^N$. On the other hand, the Sobolev inequality (2.3) does not hold for general functions in $\mathbb{T}^N$ but holds, for instance, for zero mean-value functions. Note that if $f(t)$ has mean value zero for all time $t \ge 0$, then using the invariance by Galilean transformations, $u \mapsto u(t, x - z\,t) + z$ with $z \in \mathbb{R}^3$, we can assume that u has zero mean value for all time and we are done. Otherwise, we have only to care about the treatment of the nonlinear term $(W.\nabla)v$ in (28). Denoting by $\underline{W}$ the L^2-projection of W on zero mean-value functions, we rewrite $(W.\nabla)v$ as

$$(W.\nabla)v = (\underline{W}.\nabla)v + \left(\int_\Omega W\right)\nabla v. \tag{42}$$

We take the H^s-scalar product of (28) with $\underline{W}$ and add with the L^2-scalar product of (28) with W. The H^s-scalar product coming from the first term of the right-hand side of (42) can be treated as in $\mathbb{R}^N$. For the second term, we observe that

$$\begin{aligned}\left|\left(\left(\left(\int_\Omega W\right)\nabla v, \underline{W}\right)\right)_{H^s}\right| &= \left|\int_\Omega W\right| \left|((\nabla v, \underline{W}))_{H^s}\right| \\ &\lesssim \|W\|_{L^2}\|\nabla v\|_{H^s}\|\nabla \underline{W}\|_{H^{s-1}}.\end{aligned}$$

On the other hand, concerning the L^2-scalar product we notice that

$$\begin{aligned}\left|\left(\left((W.\nabla)v, W\right)\right)_{L^2}\right| &= \left|\left(\left((\underline{W}.\nabla)v, W\right)\right)_{L^2} + \left(\int_\Omega W\right)\left(\left(\nabla v, \underline{W}\right)\right)_{L^2}\right| \\ &\lesssim \|\underline{W}\|_{L^6}\|\nabla v\|_{L^2}\|W\|_{L^3} + \left|\int_\Omega W\right|\|\nabla v\|_{L^2}\|\underline{W}\|_{L^2} \\ &\lesssim \|W\|_{H^s}\|\nabla W\|_{L^2}\|\nabla v\|_{L^2}.\end{aligned}$$

We thus obtain, exactly as in (30),

$$\begin{aligned}\frac{\mathrm{Re}}{2}\frac{d}{dt}\Big(\|\underline{W}\|_{H^s}^2 + \|W\|_{L^2}^2\Big) + \frac{3}{4}\|P_\varepsilon \nabla W\|_{H^s}^2 + \frac{(1-\omega)}{2}\|Q_{>\varepsilon}\nabla W\|_{H^s}^2 \\ \le ((Q_\varepsilon \operatorname{div}\tau, \underline{W}))_{H^s} + \|P_\varepsilon Z\|_{H^s}^2 + \frac{\omega^2}{4(1-\omega)}\|Q_\varepsilon \nabla v\|_{H^s}^2 \\ + \frac{C\ \mathrm{Re}}{(1-\omega)^2}\left(\|\nabla u\|_{H^s}^2 + \|\nabla v\|_{H^s}^2\right)\|W\|_{H^s}^2.\end{aligned}$$

The remainder of the analysis is now exactly the same as in $\mathbb{R}^N$.

3. Proof of Theorems 1.2 and 1.3. In this section we prove a convergence result in the Besov spaces $B_2^{s-1,1}$, $s \ge N/2$. It will require a smallness assumption on the retardation parameter ω but, on the other hand, will enable us to reach the critical regularity space for (5). Note that our smallness assumption on the retardation parameter is the same as the one in [1] to get the stability of the null solution in such function spaces.

Let us recall the following well-posedness result derived in [1].

THEOREM 3.1. *Let $(u_0, \tau_0) \in B^{s-1}(\mathbb{R}^N) \cap B^{N/2-1}(\mathbb{R}^N) \times B^s(\mathbb{R}^{N^2}) \cap B^{N/2}(\mathbb{R}^{N^2})$ with $s \ge N/2$. Then there exist a unique positive maximal time T^* and a unique solution*

$$(u,\tau) \in C([0,T^*[; B^{s-1}\cap B^{N/2-1}) \cap L^1_{\mathrm{loc}}(0,T^*; B^{s+1}\cap B^{N/2+1}) \times C([0,T^*[; B^s \cap B^{N/2}).$$

Moreover, if $T^* < \infty$, then

$$\limsup_{t \nearrow T^*}\Big(\|u(t)\|_{B^{N/2-1}} + \|\tau(t)\|_{B^{N/2}}\Big) = +\infty. \tag{43}$$

We will make use of the following classical commutator and product estimates (see, for instance, [1], [2], and [11]).

LEMMA 3.2. *For all* $s \in]1-N/2, 1+N/2[$ *we have*

$$\|\tilde{\Delta}_j[(a.\nabla), \Delta_j]b\|_2 \lesssim 2^{-j(s-1)}\gamma_j \|\nabla a\|_{B^{N/2+1}} \|b\|_{B^{s-1}} \tag{44}$$

with $\|\gamma_j\|_{L^1(\mathbb{Z})} \lesssim 1$.

For all $s_1, s_2 \le N/2$ with $s_1 + s_2 > 0$ it holds that

$$\|ab\|_{B^{s_1+s_2-N/2}} \lesssim \|a\|_{B^{s_1}} \|b\|_{B^{s_2}} \,. \tag{45}$$

For any $\varepsilon > 0$ we divide $\mathbb{Z}$ into the three following subsets:

$$I := \mathbb{Z}_-^* = \{j \in \mathbb{Z}, 0 < 2^j < 1\}, \quad J_\varepsilon := \{j \in \mathbb{Z}, 1 \le 2^j \le \varepsilon^{-\alpha}\}, \quad \text{and}$$
$$K_\varepsilon := \{j \in \mathbb{Z}, 2^j > \varepsilon^{-\alpha}\},$$

and for any subset $N \subset \mathbb{Z}$ we denote by $\|\cdot\|_{B^s_N}$ the seminorm

$$\|u\|_{B^s_N} = \sum_{j \in N} 2^{js} \|\Delta_j u\|_{L^2} \,.$$

3.1. Estimate on W and $\varepsilon\tau$.

LEMMA 3.3. *The solution* $(u, \varepsilon\tau)$ *of* (5) *satisfies, for all* $0 < t < T^*$,

$$\begin{aligned}
&\frac{d}{dt}\Big(\mathrm{Re}\|W\|_{B^{s-1}} + 4\varepsilon\|\tau\|_{B^s}\Big) \\
&\quad + [(1-\omega)/2 - 16\omega]\|W\|_{B^{s+1}_{K_\varepsilon}} + \|W\|_{B^{s+1}_{I\cup J_\varepsilon}} + 2\|\tau\|_{B^s_{K_\varepsilon}} + 2\varepsilon^\beta \|\tau\|_{B^s_{I\cup J_\varepsilon}} \\
&\quad \le 5\|Z\|_{B^s_{I\cup J_\varepsilon}} + 16\omega\|v\|_{B^{s+1}_{K_\varepsilon}} + 16\omega\varepsilon^\beta(\|W\|_{B^{s+1}_{I\cup J_\varepsilon}} + \|v\|_{B^{s+1}_{I\cup J_\varepsilon}}) \\
&\quad + C\,\varepsilon\mu_1 \|u\|_{B^{N/2+1}} \|\tau\|_{B^s} + C\,(\|u\|_{B^{N/2+1}} + \|v\|_{B^{N/2+1}})\|W\|_{B^{s-1}}.
\end{aligned} \tag{46}$$

Proof. Applying Δ_j to (28) we have for $j \in J_\varepsilon$,

$$\begin{aligned}
&\mathrm{Re}\Big(\partial_t \Delta_j W + \mathcal{P}(u.\nabla)\Delta_j W\Big) - (1-\omega)\Delta_j \Delta W \\
&\quad = -\omega \Delta_j \Delta v + \Delta_j \mathcal{P}\mathrm{div}\,\tau + \mathrm{Re}\,\tilde{\Delta}_j \mathcal{P}[(u.\nabla), \Delta_j]W + \mathrm{Re}\,\Delta_j \mathcal{P}(W.\nabla)v
\end{aligned} \tag{47}$$

and for $j \in I$,

$$\begin{aligned}
&\mathrm{Re}\Big(\partial_t \Delta_j W + \mathcal{P}(u.\nabla)\Delta_j W\Big) - \Delta_j \Delta W \\
&\quad = \Delta_j Z + \mathrm{Re}\,\tilde{\Delta}_j \mathcal{P}[(u.\nabla), \Delta_j]W + \mathrm{Re}\,\Delta_j \mathcal{P}(W.\nabla)v \,.
\end{aligned} \tag{48}$$

Taking the scalar product in $L^2(\mathbb{R}^N)$ of (47) with $\Delta_j W$, using that W is divergence free and using the Cauchy–Schwarz inequality we get

$$\begin{aligned}
&\frac{1}{2}\,\mathrm{Re}\,\frac{d}{dt}\|\Delta_j W\|_2^2 + (1-\omega)\|\nabla \Delta_j W\|_2^2 \\
&\qquad \le \|\Delta_j W\|_2 \Big(\omega\|\Delta_j \Delta v\|_2 + \|\Delta_j \mathrm{div}\,\tau\|_2 \\
&\qquad\qquad + \mathrm{Re}\,\|\tilde{\Delta}_j P[(u.\nabla), \Delta_j]W\|_2 + \mathrm{Re}\,\|\Delta_j P(W.\nabla)v\|_2\Big).
\end{aligned} \tag{49}$$

We use now that, according to Bernstein inequality, $\|\nabla \Delta_j W\|_2 \ge 2^{j-1}\|\Delta_j W\|_2$ and divide (49) by $\|\Delta_j W\|_2$. Then, estimating the commutator term, thanks to (44), and the last term, thanks to (45), with $s_1 = s - 1$ and $s_2 = N/2$, using Bernstein inequalities, it follows that

$$\begin{aligned} \mathrm{Re}\, \frac{d}{dt}\|\Delta_j W\|_2 + \frac{(1-\omega)}{2} 2^{2j}\|\Delta_j W\|_2 &\le 2\|\Delta\Delta_j v\|_{L^2} + 2\|\Delta_j \mathrm{div}\,\tau\|_2 \\ &\quad + \gamma_j 2^{-j(s-1)}(\|u\|_{B^{N/2+1}} + \|v\|_{B^{N/2+1}})\|W\|_{B^{s-1}} \end{aligned} \tag{50}$$

with $\|(\gamma_j)\|_{l^1(\mathbb{Z})} \lesssim 1$. Multiplying by $2^{j(s-1)}$ and summing in $j \in K_\varepsilon$, it follows that

$$\begin{aligned} \mathrm{Re}\, \frac{d}{dt}\|W\|_{B^{s-1}_{K_\varepsilon}} + \frac{(1-\omega)}{2}\|W\|_{B^{s+1}_{K_\varepsilon}} - 2\|\tau\|_{B^s_{K_\varepsilon}} \\ \le (\|u\|_{B^{N/2+1}} + \|v\|_{B^{N/2+1}})\|W\|_{B^{s-1}}\,. \end{aligned} \tag{51}$$

Proceeding in the same way with (48) but summing in $j \in I \cup J_\varepsilon$, we obtain

$$\begin{aligned} \mathrm{Re}\, \frac{d}{dt}\|W\|_{B^{s-1}_{I\cup J_\varepsilon}} + \frac{1}{2}\|W\|_{B^{s+1}_{I\cup J_\varepsilon}} - \|Z\|_{B^s_{I\cup J_\varepsilon}} \\ \le (\|u\|_{B^{N/2+1}} + \|v\|_{B^{N/2+1}})\|W\|_{B^{s-1}}\,. \end{aligned} \tag{52}$$

Now, for $j \in \mathbb{Z}$, we infer from (5) that

$$\begin{aligned} &\varepsilon\partial_t\Delta_j\tau + \varepsilon(u.\nabla)\Delta_j\tau + \Delta_j\tau \\ &\quad = 2\omega\Delta_j\mathbf{D}[u] - \varepsilon[(u.\nabla),\Delta_j]\tau + \varepsilon\Delta_j\mathbf{g}(\nabla u,\tau)\,. \end{aligned} \tag{53}$$

Rewriting $\Delta_j(\tau - 2\omega\mathbf{D}[u])$ as $\Delta_j(\tau - 2\omega\mathbf{D}[W] - 2\omega\mathbf{D}[v])$ for $j \in K_\varepsilon$ and as

$$\varepsilon^\beta\Delta_j\tau - 2\omega\varepsilon^\beta\Delta_j(\mathbf{D}[W] + \mathbf{D}[v]) + (1-\varepsilon^\beta)\Delta_j Z$$

for $j \in I \cup J_\varepsilon$, similar considerations as above lead to the two following inequalities:

$$\begin{aligned} \varepsilon\frac{d}{dt}\|\tau\|_{B^s_{K_\varepsilon}} + \|\tau\|_{B^s_{K_\varepsilon}} &\le 4\omega\|W\|_{B^{s+1}_{K_\varepsilon}} \\ &\quad + 4\omega\|v\|_{B^{s+1}_{K_\varepsilon}} + C\,\varepsilon\,\|u\|_{B^{N/2+1}}\|\tau\|_{B^s} \end{aligned} \tag{54}$$

and

$$\begin{aligned} \varepsilon\frac{d}{dt}\|\tau\|_{B^s_{I\cup J_\varepsilon}} + \varepsilon^\beta\|\tau\|_{B^s_{I\cup J_\varepsilon}} &\le \|Z\|_{B^s_{I\cup J_\varepsilon}} + 4\omega\varepsilon^\beta\|W\|_{B^{s+1}_{I\cup J_\varepsilon}} \\ &\quad + 4\omega\varepsilon^\beta\|v\|_{B^{s+1}_{I\cup J_\varepsilon}} + C\,\varepsilon\,\|u\|_{B^{N/2+1}}\|\tau\|_{B^s}. \end{aligned} \tag{55}$$

Adding (51) + (52) + 4((55) + (54)), (46) follows. ∎

3.2. Estimate on $\boldsymbol{\tau - 2\omega\mathbf{D}[u]}$.

LEMMA 3.4.

$$\begin{aligned} \frac{d}{dt}\|Z\|_{B^{s-2}_{J_\varepsilon}} + \frac{1}{\varepsilon}\|Z\|_{B^{s-2}_{J_\varepsilon}} &\le \frac{(1+\omega)}{\mathrm{Re}}\|\tau\|_{B^s_{J_\varepsilon}} + \|\mathcal{P}f\|_{B^{s-1}_{J_\varepsilon}} \\ &\quad + C\,\alpha\ln(\varepsilon^{-1})\Big(\|u\|_{B^{N/2+1}} + \|\tau\|_{B^{N/2}}\Big)\|u\|_{B^{s-1}}, \end{aligned} \tag{56}$$

$$\begin{aligned} \frac{d}{dt}\|Z\|_{B^s_I} + \frac{1}{\varepsilon}\|Z\|_{B^s_I} &\le \frac{(1+\omega)}{\mathrm{Re}}\|\tau\|_{B^s_I} + \|\mathcal{P}f\|_{B^{s-1}_I} \\ &\quad + C\Big(\|u\|_{B^{N/2+1}} + \|\tau\|_{B^{N/2}}\Big)\|u\|_{B^{s-1}}. \end{aligned} \tag{57}$$

Proof. Applying Δ_j to (33) and taking the L^2-scalar product with $\Delta_j Z$ we get

$$\frac{d}{dt}\|\Delta_j Z\|_{L^2} + \frac{(1-\omega)}{2\ \mathrm{Re}}2^{2j}\|\Delta_j Z\|_{L^2} + \frac{1}{\varepsilon}\|\Delta_j Z\|_{L^2} \lesssim \|\Delta_j f_1\|_{L^2} + \|\Delta_j f_2\|_{L^2},$$

where

$$f_1 = \frac{2\omega}{\mathrm{Re}}\mathbf{D}[\mathcal{P}\mathrm{div}\,\tau] - \frac{(1-\omega)}{\mathrm{Re}}\Delta\tau + \frac{2\omega}{\mathrm{Re}}\mathbf{D}[\mathcal{P}f] - 2\omega\mathbf{D}[\mathcal{P}(u.\nabla)u]$$

and

$$f_2 = \mathcal{P}(u.\nabla)\tau + g(\nabla u, \tau).$$

Multiplying this inequality by $2^{j(s-2)}$ and summing in $j \in J_\varepsilon$, we infer that

$$\begin{aligned}\frac{d}{dt}\|Z\|_{B^{s-2}_{J_\varepsilon}} + \frac{(1-\omega)}{2\ \mathrm{Re}}\|Z\|_{B^{s}_{J_\varepsilon}} + \frac{1}{\varepsilon}\|Z\|_{B^{s-2}_{J_\varepsilon}} \le \frac{(1+\omega)}{\mathrm{Re}}\|\tau\|_{B^{s}_{J_\varepsilon}} + \|\mathcal{P}f\|_{B^{s-1}_{J_\varepsilon}} \\ +\|(u.\nabla)u\|_{B^{s-1}_{J_\varepsilon}} + \|(u.\nabla)\tau\|_{B^{s-2}_{J_\varepsilon}} + \|\mathbf{g}(\nabla u,\tau)\|_{B^{s-2}_{J_\varepsilon}}.\end{aligned} \tag{58}$$

For $s > 1$ we estimate the nonlinear term thanks to (45) with, respectively, $(s_1, s_2) = (s-1, N/2)$, $(s-1, N/2-1)$, and $(s-2, N/2)$. For $s = 1$ (of course $N = 2$) we estimate the first nonlinear term in the same way and use the following lemma to estimate the two last ones. This lemma follows directly from the definitions of I and J_ε and the fact that, for $|s| \le N/2$, the usual product maps continuously[1] $B^{-s,1} \times B^{s,1}$ into $B^{-N/2,\infty}$ (see, for instance, [11]). Note, in particular, that $|J_\varepsilon| \lesssim \alpha \ln(\varepsilon^{-1})$. $\square$

LEMMA 3.5. *For all $s_1, s_2 \le N/2$ with $s_1 + s_2 = 0$ it holds that*

$$\|a\,b\|_{B^{-N/2}_{J_\varepsilon}} \lesssim \alpha \ln(\varepsilon^{-1})\|a\|_{B^{s_1}}\|b\|_{B^{s_2}} \tag{59}$$

and

$$\|a\,b\|_{B^{-N/2+2}_{I}} \lesssim \|a\|_{B^{s_1}}\|b\|_{B^{s_2}}. \tag{60}$$

We apply this lemma with $(s_1, s_2) = (0, 0)$ and $(-1, 1)$ for, respectively, the second and the third nonlinear term of (58) to complete the proof of (56). Finally (57) can be easily obtained in the same way by using that $\|a\|_{B^s_I} \le \|a\|_{B^{s'}_I}$ for $s' \le s$ and (60).

3.3. Convergence to the Newtonian flow. From now on we set $\gamma(\omega) = (1-\omega)/2 - 16\omega$ and assume that $0 \le \omega \le \omega_0$ with $\gamma(\omega_0) > 0$.

We proceed as in section 2.3. For $0 < \beta < 1$, we add (46) and $\varepsilon^{2\beta}((56) + (57))$ to get

$$\begin{aligned}&\frac{d}{dt}\Big(\mathrm{Re}\|W\|_{B^{s-1}} + 4\varepsilon\|\tau\|_{B^s} + \varepsilon^{2\beta}(\|Z\|_{B^{s-2}_{J_\varepsilon}} + \|Z\|_{B^s_I})\Big) + \gamma(\omega_0)\|W\|_{B^{s+1}_{K_\varepsilon}} \\ &\quad + \|W\|_{B^{s+1}_{I\cup J_\varepsilon}} + 2\|\tau\|_{B^s_{K_\varepsilon}} + \varepsilon^{\beta}\|\tau\|_{B^s_{I\cup J_\varepsilon}} + \frac{\varepsilon^{2\beta-1}}{2}(\|Z\|_{B^{s-2}_{J_\varepsilon}} + \|Z\|_{B^s_I}) \\ &\le 16\omega\|v\|_{B^{s+1}_{K_\varepsilon}} + 16\omega\varepsilon^{\beta}\|v\|_{B^{s+1}_{I\cup J_\varepsilon}} + \varepsilon^{2\beta}\|\mathcal{P}f\|_{B^{s-1}_{I\cup J_\varepsilon}} \\ &\quad + C\,\varepsilon\|u\|_{B^{N/2+1}}\|\tau\|_{B^s} + C\,(\|u\|_{B^{N/2+1}} + \|v\|_{B^{N/2+1}})\|W\|_{B^{s-1}} \\ &\quad + C\,\alpha\varepsilon^{2\beta}\ln(\varepsilon^{-1})(\|u\|_{B^{N/2+1}} + \|\tau\|_{B^{N/2}})\,\|v\|_{B^{s-1}}.\end{aligned} \tag{61}$$

[1]For $1 \le p \le \infty$, $\|f\|_{B^{s,p}} = \big\|\{2^{js}\|\Delta_j(f)\|_{L^2}\}\big\|_{l^p(\mathbb{Z})}$.

Here we used that for ε small enough, $\varepsilon^\beta \le \min(16\gamma(\omega_0), \frac{\mathrm{Re}}{4})$, $\varepsilon^{2\beta-1}/2 \ge 5$, and

$$5\|Z\|_{B^s_{J_\varepsilon}} \lesssim \varepsilon^{-2\alpha}\|Z\|_{B^{s-2}_{J_\varepsilon}} \le \frac{\varepsilon^{2\beta-1}}{2}\|Z\|_{B^{s-2}_{J_\varepsilon}}$$

as soon as

$$0<\alpha<1/2 \quad \text{and} \quad 0<2\beta<1-2\alpha\,. \tag{62}$$

From now on we set $(\alpha,\beta)=(1/8,1/8)$ so that (62) is satisfied. Setting

$$\begin{aligned}X_s(t) =\ & \mathrm{Re}\,\|W(t)\|_{B^{s-1}} + 4\varepsilon\|\tau\|_{B^s} + \varepsilon^{2\beta}\|Z\|_{B^{s-2}_{I\cup J_\varepsilon}}\\ &+\int_0^t \frac{\gamma(\omega_0)}{2}\|W\|_{B^{s+1}} + \Big(\|\tau\|_{B^s_{K_\varepsilon}} + \varepsilon^\beta\|\tau\|_{B^s_{I\cup J_\varepsilon}}\Big) + \frac{\varepsilon^{2\beta-1}}{2}\Big(\|Z\|_{B^{s-2}_{J_\varepsilon}} + \|Z\|_{B^s_I}\Big)\,ds,\end{aligned}$$

we infer that

$$\begin{aligned}\frac{d}{dt}X_s(t) \le\ & 16\omega\|v\|_{B^{s+1}_{K_\varepsilon}} + 16\omega\varepsilon^\beta\|v\|_{B^{s+1}_{I\cup J_\varepsilon}} + \varepsilon^{2\beta}\|\mathcal{P}f\|_{B^{s-1}_{I\cup J_\varepsilon}}\\ &+C(\mathrm{Re}\,,\omega)\Big[\|W\|_{B^{N/2+1}} + \|v\|_{B^{N/2+1}} + \varepsilon^{2\beta}\ln(\varepsilon^{-1})\|\tau\|_{B^{N/2}}\Big]X_s\\ &+C\,\alpha\varepsilon^{2\beta}\ln(\varepsilon^{-1})(\|v\|_{B^{N/2+1}} + \|W\|_{B^{N/2+1}} + \|\tau\|_{B^{N/2}})\,\|v\|_{B^{s-1}}.\end{aligned}$$

By Gronwall's lemma we infer that

$$\begin{aligned}X_s(t) \le\ & \exp\Big(C(\omega,\ \mathrm{Re}\,)\Big(\|v\|_{L^1_tB^{N/2+1}} + X_{N/2}(t)\Big)\Big)\\ &\Big[X_s(0) + 16\omega\|v\|_{L^1_tB^{s+1}_{K_\varepsilon}} + 16\omega\varepsilon^\beta\|v\|_{L^1_tB^{s+1}_{I\cup J_\varepsilon}} + \varepsilon^{2\beta}\|\mathcal{P}f\|_{L^1_tB^{s-1}_{I\cup J_\varepsilon}}\\ &+C\,\alpha\ln(\varepsilon^{-1})\varepsilon^\beta\Big(X_{N/2}(t)\|v\|_{L^\infty_tB^{s-1}} + \varepsilon^\beta\|v\|_{L^1_tB^{s-1}}\|v\|_{L^\infty_tB^{s-1}}\Big)\Big],\end{aligned} \tag{63}$$

where

$$X_s(0) = 4\varepsilon\|\tau_0\|_{B^s} + \varepsilon^{2\beta}\Big(\|\tau_0 - 2\omega\mathbf{D}[u_0]\|_{B^s_I} + \|\tau_0 - 2\omega\mathbf{D}[u_0]\|_{B^{s-2}_{J_\varepsilon}}\Big)\,. \tag{64}$$

Assuming that $T^*_\varepsilon \le T_0$ and noticing that

$$\|v\|_{L^1_{T_0}B^{N/2+1}_{K_\varepsilon}} \to 0 \text{ as } \varepsilon \to 0, \tag{65}$$

we deduce from (63) and (64) and the continuity of $t \mapsto X_{N/2}(t)$ that there exists $\varepsilon_0 = \varepsilon_0(N, \|\tau_0\|_{B^{N/2}}, \mathcal{P}f, u_0)$ such that for any $0<\varepsilon<\varepsilon_0$ and any $0<t<T^*_\varepsilon$,

$$X_{N/2}(t) \le \Lambda(\varepsilon)$$

with $\Lambda(\varepsilon) \searrow 0$ as $\varepsilon \to 0$. This contradicts (43) of Theorem 3.1 and thus ensures that $T^*_\varepsilon > T_0$. The convergence results (16) and (18) follow as well. To prove (17) we notice that from this last limit and (65), $\|\tau - 2\omega\mathbf{D}[u]\|_{L^1_{T_0}B^{N/2}_{K_\varepsilon}} \to 0$, $\varepsilon^{2\beta-1}\|\tau - 2\omega\mathbf{D}[u]\|_{L^1_{T_0}B^{N/2}_I} \to 0$, and $\varepsilon^{2\beta-1}\|\tau - 2\omega\mathbf{D}[u]\|_{L^1_{T_0}B^{N/2-2}_{J_\varepsilon}} \to 0$. This gives the result since $2\beta - 1 + 2\alpha \le 0$ and thus

$$\|\tau - 2\omega\mathbf{D}[u]\|_{L^1_{T_0}B^s_{J_\varepsilon}} \lesssim \varepsilon^{-2\alpha}\|\tau - 2\omega\mathbf{D}[u]\|_{L^1_{T_0}B^{s-2}_{J_\varepsilon}} \lesssim \varepsilon^{2\beta-1}\|\tau - 2\omega\mathbf{D}[u]\|_{L^1_{T_0}B^{s-2}_{J_\varepsilon}}.$$

Finally, for $s > N/2$, the proof follows the same lines using that

$$\|v\|_{L^1_{T_0} B^{N/2+1}_{K_\varepsilon}} \le \varepsilon^{\alpha(s-N/2)} \|v\|_{L^1_{T_0} B^{s+1}_{K_\varepsilon}} \tag{66}$$

and thus with

$$\varepsilon_0 = \varepsilon_0(N, \|\tau_0\|_{B^{N/2}}, \|u_0\|_{B^{N/2-1}}, \|v\|_{L^1_{T_0} B^{s+1}_{K_\varepsilon}}, \|\mathcal{P} f\|_{L^1_{T_0} B^{N/2-1}}).$$

This completes the proof of Theorems 1.2 and 1.3. ∎

Acknowledgment. The authors are grateful to the referees for useful remarks.

REFERENCES

[1] J.-Y. Chemin and N. Masmoudi, *About lifespan of regular solutions of equations related to viscoelastic fluids*, SIAM J. Math. Anal., 33 (2001), pp. 84–112.

[2] R. Danchin, *Zero Mach number limit in critical spaces for compressible Navier–Stokes equations*, Ann. Sci. École Norm. Sup. (4), 35 (2002), pp. 27–75.

[3] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.

[4] E. Fernandez-Cara, F. Guillen, and R. R. Ortega, *Somme theoretical results concerning non-Newtonian fluids of Oldroyd kind*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 26 (1998), pp. 1–29.

[5] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics 24, Pitman, Boston, MA, 1985.

[6] C. Guillopé and J.C. Saut, *Existence results for flow of viscoelastic fluids with a differential constitutive law*, Nonlinear Anal., 15 (1990), pp. 849–869.

[7] C. E. Kenig, G. Ponce, and L. Vega, *Well-posedness and scattering results for the generalized Korteweg-de Vries equation via contraction principle*, Comm. Pure Appl. Math., 46 (1993), pp. 527–620.

[8] L. Molinet and R. Talhouk, *On the global and periodic regular flows of viscoelastic fluids with a differential constitutive law*, NoDEA Nonlinear Differential Equations Appl., 11 (2004), pp. 349–359.

[9] J. G. Oldroyd, *On the formation of rheological equations of state*, Proc. Roy. Soc. London Ser. A, 200 (1950), pp. 523–541.

[10] M. Renardy, W. J. Hrusa, and J. A. Nohel, *Mathematical Problems in Viscoelasticity*, Longman, London, 1987.

[11] M. E. Taylor, *Tools for PDE. Pseudodifferential Operators, Paradifferential Operators, and Layer Potentials*, Math. Surveys Monogr. 81, AMS, Providence, RI, 2000.

[12] M. Tom, *Smoothing properties of some weak solutions of the Benjamin-Ono equation*, Differential Integral Equations, 3 (1990), pp. 683–694.

SIAM J. Math. Anal.
Vol. 39, No. 5, pp. 1595–1626

INVERSE SCATTERING FOR SCHRÖDINGER-TYPE OPERATORS WITH INTERFACE CONDITIONS ACROSS SMOOTH SURFACES*

STEPHEN O'DELL†

Abstract. We consider direct and inverse scattering for the Laplace–Beltrami operator with electromagnetic potentials in domains with smooth surfaces upon which we impose interface conditions. The boundary conditions used encompass physical models of imperfect transmission arising in acoustics, quantum scattering, semiconductors, and geophysics. We prove uniqueness of the location of the surfaces and the interface conditions from the fixed-energy scattering amplitude. If the surface encloses a compact region, we also prove uniqueness of the Dirichlet-to-Neumann operator at the boundary of the obstacle.

Key words. inverse scattering, interface conditions, transmission problems

AMS subject classifications. 81U40, 35J25

DOI. 10.1137/060675198

1. Introduction. Consider the Laplace–Beltrami operator with electromagnetic potentials in $\mathbb{R}^n$:

$$L = \left(\sum_{j,k} \frac{1}{\sqrt{g}} \left(-i\frac{\partial}{\partial x_j} + A_j(x) \right) \sqrt{g} g^{jk}(x) \left(-i \frac{\partial}{\partial x_k} + A_k(x) \right) \right) + V(x), \tag{1.1}$$

where all the coefficients are real-valued and $L = -\triangle$ for $|x| \geq R$. In its full form, this operator is the Hamiltonian for a quantum particle in an electromagnetic field constrained to a Riemannian manifold. It also can be used to model wave propagation in anisotropic media. Here, $A(x) = (A_1(x), \ldots, A_n(x))$ is the magnetic potential, $V(x)$ is the electric potential, and $\sum_{j,k} g_{jk}(x)\, dx^j \otimes dx^k$ is a metric with determinant $g(x)$. In this paper, we study inverse scattering for this operator when it is defined with interface conditions across smooth embedded surfaces.

To explain this in detail, we fix some bounded open connected domain Ω with smooth boundary $\partial\Omega$. Define the boundary operators,

$$\gamma^0_\pm u(x) = \lim_{x\to\partial\Omega^\pm} u(x), \tag{1.2}$$

$$\gamma^1_\pm u(x) = \lim_{x\to\partial\Omega^\pm} \sum_{j,k} g^{jk} \left(\frac{\partial u}{\partial x_j} + iA_j u \right) \nu_k(x) \left(\sum_{p,r} g^{pr}(x)\nu_p(x)\nu_r(x) \right)^{-\frac{1}{2}}, \tag{1.3}$$

where the positive limit denotes that it is taken from the exterior and ν is the outward pointing unit normal (with respect to the Euclidean metric). Let L^e and L^i be operators of the form (1.1) with smooth coefficients $\{g_e^{jk}, A_j^e, V^e\}$ and $\{g_i^{jk}, A_j^i, V^i\}$, respectively. Now consider the time-independent scattering problem:

$$(L^e - k^2)(e^{ik\omega\cdot x} + w(x,k\omega)) = 0 \text{ in } \mathbb{R}^n\backslash\Omega, \tag{1.4}$$

$$(L^i - k^2)w(x,k\omega) = 0 \text{ in } \Omega, \tag{1.5}$$

*Received by the editors November 16, 2006; accepted for publication (in revised form) October 4, 2007; published electronically January 30, 2008.
http://www.siam.org/journals/sima/39-5/67519.html

†Department of Mathematics, University of California, Los Angeles, 520 Portola Plaza, Box 951555, Los Angeles, CA 90095-1555. Current address: Trafelet Delta Funds, 590 Madison Avenue, 39th Floor, New York, NY 10022 (sodell@trafelet.com).

with interface conditions,

$$\begin{pmatrix} \gamma_-^0 w \\ \gamma_-^{1,i} w \end{pmatrix} = \begin{pmatrix} a(x) & b(x) \\ c(x) & d(x) \end{pmatrix} \begin{pmatrix} \gamma_+^0 (e^{ik\omega\cdot x} + w) \\ \gamma_+^{1,e} (e^{ik\omega\cdot x} + w) \end{pmatrix}, \tag{1.6}$$

and outgoing radiation conditions,

$$\frac{\partial w}{\partial r}(x, k\omega) - ikw(x, k\omega) = o\left(\frac{1}{r^{\frac{n-1}{2}}}\right), \quad w(x, k\omega) = O\left(\frac{1}{r^{\frac{n-1}{2}}}\right). \tag{1.7}$$

In what follows, we denote the matrix in (1.6) as $T(x)$ and refer to it as the interface matrix, or transfer matrix. It is worthwhile to remark that we make no assumptions here on the coefficients or the interface to guarantee there is in fact a boundary. This allows us to use the above problem to also study *transmission cracks* which occur if $L^e = L^i$ everywhere and $T = I$ on some subset of the boundary. Similarly this boundary value problem also covers the case when there may just be a "tear" in the coefficients, i.e., L^i and L^e smoothly connect and T is trivial on some subset of the boundary.

Let us briefly formulate the scattering problem. The direct problem is to show there exists a solution $w(x, k\omega)$ solving (1.4)–(1.7) that has the asymptotic form,

$$w(x, k\omega) = e^{ik\omega\cdot x} + a\left(\frac{x}{|x|}, \omega, k\right)\frac{e^{ik|x|}}{|x|^{\frac{n-1}{2}}} + O\left(\frac{1}{|x|^{\frac{n+1}{2}}}\right), \tag{1.8}$$

where $a(\theta, \omega, k)$ is the scattering amplitude for some $k > 0$. The inverse problem is to show that the scattering amplitude uniquely characterizes the perturbations to the exterior coefficients. For the transmission obstacles given above, this entails showing that the location of the boundary and the transfer matrix is uniquely determined everywhere that at least one of the following holds: (1) The coefficients are not smooth across $\partial\Omega$, or (2) $T \neq I$. Also, we want to prove that the Dirichlet-to-Neumann operator is uniquely determined on $\partial\Omega$.

Without *any* further restrictions on the transfer matrix or the coefficients, there are some constant factors by which we can change the interface and interior coefficients which do not affect the fixed energy scattering amplitude. Moreover, altering the coefficients and interfaces in this manner will produce a different obstacle. Indeed, we prove that the scattering operator can distinguish two such obstacles, whereas the fixed-energy scattering amplitude cannot. In most practical applications, however, the boundary conditions (1.6) are subject to constraints which remove these degeneracies. This is discussed in detail in section 2.2. In the interest of obtaining the most general results, we prove the uniqueness of the inverse problem up to these constants and then interpret the results in physically interesting cases afterward.

Our main theorems, as well as some background, are given in section 2. In section 2.1, we formulate our result on the forward problem and in section 2.2 cite physical examples in which these boundary conditions arise. In section 2.3 the results on the inverse problem are given, and in section 2.4 we discuss previous work on scattering from transmission obstacles and interfaces.

The plan of the remainder of the paper is as follows. After reviewing preliminary material on semigeodesic coordinates, layer potentials, and wave front sets in section 3, we solve the direct problem in section 4 by reducing the problem to a system of pseudodifferential equations on $\partial\Omega$. In section 5, we solve the inverse problem when there are only electric potentials. In this case the analysis of certain boundary operators is simplified since the first two terms of the symbol of the Dirichlet-to-Neumann operator are known. In section 6, the full symbol of the Dirichlet-to-Neumann operator is recalled and we formulate a theorem on boundary determination in the presence of anisotropic media and electromagnetic potentials. This prepares us to solve the general inverse problem in section 7 in which case a much more exhaustive analysis is required to prove the uniqueness.

2. Results and background.

2.1. The direct problem.

2.1.1. Self-adjointness. Consider the scattering problem (1.4)–(1.7). Let $\sqrt{\tilde{g}^e}$ and $\sqrt{\tilde{g}^i}$ denote the determinant of the restriction of the metrics to the boundary $\partial\Omega$. We say the boundary conditions are *self-adjoint* if:

$$a(x)d(x) - b(x)c(x) = \frac{\sqrt{\tilde{g}^e}(x)}{\sqrt{\tilde{g}^i}(x)}. \tag{2.1}$$

The reason for this definition is that if we define L to be an operator equal to L^i on Ω and L^e on $\mathbb{R}^n\backslash\Omega$, and impose interface conditions of the form (1.6) on $\partial\Omega$, then the adjoint of L, as defined in the L^2-inner product, will be the same as L if and only if (2.1) holds. This is proved in section 4.2. We also show there that (2.1) is equivalent to the energy flux being equal on each side of the boundary.

The distinction between self-adjoint and nonself-adjoint conditions is relevant since it affects the solvability of the forward problem. In particular, it is in general not possible to prove the uniqueness with nonself-adjoint boundary conditions. Nonetheless, we prove solvability off of a discrete set of energies in this case.

2.1.2. Results on the direct problem. We always assume that

$$a > 0,\ \ d > 0,\ \text{ and } \det T \neq 0, \tag{2.2}$$

and either

$$b \neq 0 \text{ on all of } \partial\Omega \text{ or } b = 0 \text{ on all of } \partial\Omega, \tag{2.3}$$

or

$$\exists \Delta \subset \partial\Omega \text{ open with } \partial\Delta \text{ smooth s.t. } \operatorname{Re} b \neq 0 \text{ on } \Delta \text{ and } b = 0 \text{ on } \partial\Omega\backslash\overline{\Delta}. \tag{2.4}$$

The term $b(x)$ requires special consideration since it contributes to the leading term of the system of pseudodifferential equations to which the direct problem is reduced. The conditions in (2.4) describe an analogue of partially coated obstacles for transmission obstacles, which we introduce to model transmission obstacles with a resistive coating on a subset of the boundary. Finally, assume a, c, and d are smooth everywhere and b is smooth on $\overline{\Delta}$ and $\partial\Omega\backslash\Delta$. Unless otherwise stated, we always assume that (2.2) holds with either (2.3) or (2.4) for every transfer matrix.

Our main result on the direct scattering problem is the following.

THEOREM 2.1. *Say T satisfies* (2.2) *and* (2.3) *or* (2.4). *Then for all but a discrete set of $k \in (0,\infty)$ there exists a solution to* (1.4)–(1.7) *of the form*

$$w(x, k\omega) = a(\theta, \omega, k)\frac{e^{ik|x|}}{|x|^{\frac{n-1}{2}}} + O\left(\frac{1}{|x|^{\frac{n+1}{2}}}\right) \tag{2.5}$$

as $|x| \to \infty, \theta = \frac{x}{|x|}$. *Moreover, if also*

$$\mathrm{Im} b \le 0, \quad \mathrm{Im} c \ge 0, \qquad \textit{and} \qquad ad - \bar{b}c = \frac{\sqrt{\tilde{g}^e}}{\sqrt{\tilde{g}^i}}, \tag{2.6}$$

then the discrete set is empty. In particular, (1.4)–(1.7) *is solvable for all* $k > 0$ *in the self-adjoint case.*

These results are proved in section 4.

2.2. Physical background. It is worthwhile to briefly relate our problem explicitly to the case of acoustic scattering in anisotropic media. In this situation, to model the inhomogeneities, one uses an operator of the general form

$$L' = -\sum_{j,k=1}^{n} \frac{\partial}{\partial x_j} \gamma^{jk}(x) \frac{\partial}{\partial x_k} + \eta(x), \tag{2.7}$$

where we assume $\eta(x) = 0$ and $\gamma^{jk}(x) = \delta_{jk}$ for $|x|$ large. For $n \ge 3$ we can set $g^{jk} = (\det\{\gamma^{jk}\})^{\frac{1}{2-n}} \gamma^{jk}$ and multiply through by $\sqrt{g}$ in (1.4) to obtain an operator of the form (2.7). Moreover, it is straightforward to check that (if $A_j = 0$),

$$\sum_{j,k} \gamma^{jk}(x) \left(\frac{\partial u}{\partial x_k} \right)_+ \nu^k = \gamma_+^{1,L} u(x) \sqrt{\tilde{g}^e}. \tag{2.8}$$

Thus, an interface for operators of the form (2.7) has the form

$$\begin{pmatrix} \gamma_-^0 u \\ \sum_{j,k} \gamma_i^{jk}(x)(\frac{\partial u}{\partial x_k})_+ \nu^k \end{pmatrix} = \begin{pmatrix} a'(x) & b'(x) \\ c'(x) & d'(x) \end{pmatrix} \begin{pmatrix} u_+ \\ \sum_{j,k} \gamma_e^{jk}(x)(\frac{\partial u}{\partial x_k})_- \nu^k \end{pmatrix}, \tag{2.9}$$

where for g^{jk} and γ^{jk} related as above,

$$\begin{pmatrix} a'(x) & b'(x) \\ c'(x) & d'(x) \end{pmatrix} = \begin{pmatrix} a(x) & b(x)(\sqrt{\tilde{g}^e}) \\ c(x)(\sqrt{\tilde{g}^i})^{-1} & d(x)(\sqrt{\tilde{g}^e})(\sqrt{\tilde{g}^i})^{-1} \end{pmatrix}. \tag{2.10}$$

Despite the simpler form of the transfer matrix, we will always use the Laplace–Beltrami operator even when discussing acoustic wave propagation in anisotropic media since using the Laplace–Beltrami considerably simplifies some of the calculations, especially when it is necessary to make coordinate changes.

We now discuss some physical examples. Recall that the natural transmission conditions at an interface have the form

$$\gamma_+^0 u - \gamma_-^0 u = 0, \tag{2.11}$$

$$\gamma_+^{1,e} u \sqrt{\tilde{g}^e} - \gamma_-^{1,i} u \sqrt{\tilde{g}^i} = 0. \tag{2.12}$$

A simple application of the Green's function yields these conditions. Note there are no density terms for operators of the form (2.7).

Imperfect interfaces occur originally in the study of elastodynamic waves passing between two different elastic media which are not in perfect contact. Interpreted in the context of acoustic waves, the boundary conditions have the form

$$\gamma_+^0 u - \gamma_-^0 u = \alpha(x) \left[\gamma_-^{1,i} u \sqrt{\tilde{g}^i} + \gamma_+^{1,e} u \sqrt{\tilde{g}^e} \right], \tag{2.13}$$

$$\gamma_+^{1,e} u \sqrt{\tilde{g}^e} - \gamma_-^{1,i} u \sqrt{\tilde{g}^i} = \beta(x) \left[\gamma_-^0 u + \gamma_+^0 u \right]. \tag{2.14}$$

See [3] and [34] for background and references. Physically, we can interpret α and β as the relative compressibility and permeability of a thin membrane separating the two regions [34]. These interface conditions are easily converted into boundary conditions of the form (1.6) and the resultant interface matrix looks like

$$T = \frac{1}{1-\alpha\beta}\begin{pmatrix} 1+\alpha\beta & -2\alpha(\sqrt{\tilde{g}^e}) \\ -2\beta(\sqrt{\tilde{g}^i})^{-1} & (1+\alpha\beta)(\sqrt{\tilde{g}^e})(\sqrt{\tilde{g}^i})^{-1} \end{pmatrix}. \tag{2.15}$$

Of course, for operators of the form (2.7), the interface matrix does not have the density terms.

In the case $\beta = 0$, the boundary conditions are known as spring-contact boundary conditions. For background on their use, see the review given in [3]. In this paper the authors also derive a more accurate mathematical model (than the spring-contact boundary conditions) for 2-dimensional sound harmonic waves separated by a thin layer which turn out to be of the form (2.11)–(2.12), except that (2.12) has an additional correction term dependent on second order tangential derivatives at the surface.

The resistive and conductive transmission boundary conditions, which arise in geophysics, are also of this form (see [1] and the references therein). Resistive conditions arise if α is complex-valued and $\beta = 0$ in (2.13)–(2.14), and conductive conditions are given by assuming β is complex-valued and $\alpha = 0$ in (2.13)–(2.14).

In quantum scattering the simplest example of boundary conditions of the form (1.6) arises in the model of an electron wave passing through a δ-like potential. Assuming the media is Euclidean for simplicity, then if $V(x) = \kappa\delta_{\partial\Omega}$, it is easy to derive the transfer matrix will have the form (see [13])

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

In general, the boundary conditions describing scattering from electric potentials of the above form are given by (2.13)–(2.14) with $\alpha = 0$.

Finally, transfer matrix boundary conditions have also been in use for many years in the effective mass method which is used to model the electronic states of semiconducting materials containing abrupt interfaces between materials of different compositions. Transfer matrix heterojunctions, as they are known in the literature, are used in this context to connect wave functions, known as envelopes, across regions of different chemical compositions and through barriers of yet a third material. See [2] and [32] for a review of their use as well as various physical interpretations of the matrix elements.

2.3. Results on the inverse problem.

2.3.1. Nonuniqueness. Consider (1.4)–(1.7). There are two types of alterations as follows which do not affect the scattering amplitude:

1. $T \to \tau T$ for $\tau > 0$.
2.

$$\begin{pmatrix} \Omega \\ T \\ g^{jk} \\ A_j \\ V \end{pmatrix} \to \begin{pmatrix} \Omega \\ \begin{pmatrix} \rho^{-\frac{n}{2}} & 0 \\ 0 & \rho^{1-\frac{n}{2}} \end{pmatrix} T \\ \rho^2 g^{jk} \\ A_j \\ \rho^2 V + k^2(1-\rho^2) \end{pmatrix} \quad \text{for } \rho > 0.$$

When describing physical phenomena, the model of a transmission obstacle will generally be selected in such a fashion that these constants do not arise. Shortly we will discuss a number of restrictions that remove these constants.

The first type of nonuniqueness results from the fact that if we multiply the interface by a constant, we can multiply the solution on the interior by the inverse of that constant without affecting the solution on the exterior. Of course, this type of degeneracy will not affect the full scattering operator and is somewhat trivial.

The second type of nonuniqueness is obtained by replacing (L^i-k^2) by $\rho^2(L^i-k^2)$ in (1.5) and then changing the interface matrix accordingly. The precise powers of ρ are selected in such a way that self-adjoint boundary conditions remain self-adjoint. Note that as a result of this type of nonuniqueness, there are always an infinite number of obstacles which are not identical to the background but which nonetheless are "invisible" to the scattering amplitude at a given fixed energy. Such obstacles, moreover, are distinguishable from the background by the scattering operator. It is worthwhile to formulate some restrictions as follows which force the constants to be one:

$\tau = 1$ if any of the following hold:
- The boundary conditions are self-adjoint.
- Any nonzero element of the interface matrix is fixed.

$\rho = 1$ if any of the following hold:
- $d(x) = (\sqrt{\tilde{g}^e})(\sqrt{\tilde{g}^i})^{-1}a(x)$.
- $g_e^{jk} = g_i^{jk} = \delta_{jk}$.
- $V^e = V^i = 0$.
- Any nonzero element of the interface matrix is fixed.

Note that for imperfect interfaces, the boundary conditions (2.15) are self-adjoint and $d(x) = (\sqrt{\tilde{g}^e})(\sqrt{\tilde{g}^i})^{-1}a(x)$ so that both τ and ρ are 1. The resistive and conductive boundary conditions, as well as the boundary conditions describing scattering from delta-like potentials, are a subset of imperfect interface boundary conditions (2.15).

In order to simplify notation in the future, we say $(L_1 - k^2)$ and $\rho^2(L_2 - k^2)$ *smoothly connect in the normal direction at* $x_0 \in \partial\Omega$ if for all p,

$$\left(\frac{\partial}{\partial\nu}\right)^p g_1^{jk}(x_0) = \rho^2 \left(\frac{\partial}{\partial\nu}\right))^p g_2^{jk}(x_0, \quad \left(\frac{\partial}{\partial\nu}\right)^p A_j^1(x_0) = \left(\frac{\partial}{\partial\nu}\right)^p A_j^2(x_0), \text{ and}$$

$$\left(\frac{\partial}{\partial\nu}\right)^p V_1(x_0) = \rho^2 \left(\frac{\partial}{\partial\nu}\right)^p (V_2(x_0) + k^2(1-\rho^2)). \tag{2.16}$$

2.3.2. Inverse problem. Our main result is that the fixed-energy scattering amplitude uniquely determines the obstacles up to the constant factors described above. We interpret these results in the case of imperfect interfaces afterwards. First, consider the location and let an obstacle be defined by three elements $\{\Omega, T, L^i\}$. Ω is a subset of $\mathbb{R}^n$, T is the interface matrix, and L^i is an operator of the form (1.1) which describes the physical properties of the interior of the obstacle.

THEOREM 2.2. *Consider scattering from an obstacle* $\{\Omega, T, L^i\}$ *as in problem* (1.4)–(1.7)*, where* T *satisfies* (2.2) *and* (2.3) *or* (2.4)*. Then* $a(\theta,\omega,k)$ *for* $\theta,\omega \in S^{n-1}$ *and some fixed* $k>0$ *uniquely determines the location of all points* $x_0 \in \partial\Omega$ *at which there does not exist* $\tau > 0, \rho > 0$ *such that*

1. $T(x_0) = \tau \begin{pmatrix} \rho^{-\frac{n}{2}} & 0 \\ 0 & \rho^{1-\frac{n}{2}} \end{pmatrix}$ *and*
2. $\rho^2(L^i - k^2)$ *and* $(L^e - k^2)$ *smoothly connect in the normal direction at* x_0.

When the constants are one, this means we uniquely determine all points on the boundary, where either $T \neq I$ or the coefficients do not smoothly connect. Now consider the boundary conditions. Below, Λ_j denotes the Dirichlet-to-Neumann operator on $\partial\Omega$ for the operator $L_j^i - k^2$ and $\sigma(\Lambda_j)$ is its full symbol (see section 6).

THEOREM 2.3. *Assume we are given two obstacles* $\{\Omega, T_1, L_1^i\}$ *and* $\{\Omega, T_2, L_2^i\}$ *with* T_1 *and* T_2 *satisfying* (2.2) *and* (2.3) *or* (2.4) *along with their respective scattering data* $a_1(\theta,\omega,k)$ *and* $a_2(\theta,\omega,k)$*. Then* $a_1(\theta,\omega,k) = a_2(\theta,\omega,k)$ *for* $\theta,\omega \in S^{n-1}$ *and some* $k > 0$ *implies there exists* $\tau > 0$, $\rho > 0$ *such that,*

$$T_1(x) = \tau \begin{pmatrix} \rho^{-\frac{n}{2}} & 0 \\ 0 & \rho^{1-\frac{n}{2}} \end{pmatrix} T_2(x) \text{ and } \sigma(\Lambda_1) = \rho\sigma(\Lambda_2).$$

Also, if k^2 *is not a Dirichlet eigenvalue for either* L_1^i *or* L_2^i *on* Ω*, then* $\Lambda_1 = \rho\Lambda_2$.

These results are proved in sections 5 and 7. As is proved in section 6 the equality of the symbols of the Dirichlet-to-Neumann operators implies that $(L_1^i - k^2)$ and $\rho^2(L_2^i - k^2)$ smoothly connect on the boundary in the sense of (2.16).

For imperfect interfaces (2.13) and (2.14), the results are as follows.

THEOREM 2.4. $a(\theta,\omega,k)$ *for* $\theta,\omega \in S^{n-1}$ *and some fixed* $k > 0$ *uniquely determines the location of all points* $x_0 \in \partial\Omega$ *at which it is not the case that*

1. $\alpha(x_0) = 0$, $\beta(x_0) = 0$ *and*
2. L^i *and* L^e *smoothly connect in the normal direction at* x_0.

THEOREM 2.5. *Assume* Ω *is known. Then* $a_1(\theta,\omega,k) = a_2(\theta,\omega,k)$ *for* $\theta,\omega \in S^{n-1}$ *and some* $k > 0$ *implies,*

$$\alpha_1(x) = \alpha_2(x),\ \beta_1(x) = \beta_2(x),\ \sigma(\Lambda_1) = \sigma(\Lambda_2).$$

Also, if k^2 *is not a Dirichlet eigenvalue for either* L_1 *or* L_2 *on* Ω*, then* $\Lambda_1 = \Lambda_2$.

These theorems are proved using Theorems 2.2 and 2.3. Since Theorems 2.2 and 2.3 are the more general results, we focus our attention on them throughout the paper.

2.4. Historical remarks and overview of approach. Already there has been extensive study of direct and inverse scattering from transmission obstacles in the case of perfectly transmitting interfaces and constant isotropic media on the interior and exterior (see the review article [5] and books [6] and [7] for a full history). The most general results in this case are due to Kirsch and Päivärinta [21], in which the authors consider the scattering problem for transmission obstacles embedded in known electric inhomogeneities with boundary conditions of the form

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

(with τ constant) and containing unknown electric potentials. Here τ represents the inverse of the density of the interior constant isotropic media so that these boundary conditions are in fact self-adjoint. They prove that the fixed energy scattering data uniquely determine the location of the obstacle, the boundary conditions, and the interior electric potentials.

Recently there has also been greater focus on the scattering problem when the obstacle contains anisotropic media (see [4], [7], [9], [15], [27], and [28]). The anisotropic media is modeled in these papers using an operator of the form $\nabla \circ A(x)\nabla$ and the inverse scattering problem of determining the location is solved under various restrictions on $A(x)$. The recovery of the interior media is not considered in these papers.

As is well known, for $n \geq 3$ this problem is equivalent to the one we consider with a Laplace–Beltrami operator. A closely related problem is obtained by allowing inhomogeneous media on the interior (i.e., media with variable, isotropic density). The related direct scattering problem is considered by Martin [24] and Werner [35] and the inverse problem by Isakov [20]. The main result of [20] is that the fixed energy scattering data uniquely determine the location of the obstacle as well as the media at the boundary of the obstacle. Inhomogeneous media are covered by our work for $n \geq 3$ by using the metric $g_{jk}(x) = \rho(x)^{\frac{2-n}{2}}\delta_{jk}$, where ρ is the inhomogeneous density.

There has been little previous work on general boundary conditions of the type we consider. The only results we are aware of with off-diagonal elements in the transfer matrix are for resistive and conductive boundary conditions (see [1], [14], [16], and [33]) with the Helmholtz equation on the interior and exterior. In these papers the direct scattering problem is solved and in [16] and [33] it is shown the fixed-energy scattering data uniquely determine the boundary conditions and the location of an obstacle with conductive interfaces. Finally, we are unaware of previous work with nonself-adjoint boundary conditions.

To solve the inverse problem we investigate the behavior of the singularity of the Green's function near the boundary of obstacles. The idea to consider the behavior of the Green's function is originally due to Isakov [20]. We obtain a precise formula for the singularity of the Green's function in terms of the interface matrix and the symbol of the Dirichlet-to-Neumann operator as the pole approaches some point on the boundary. This is different approach than the one used by Isakov [20], Kirsch and Päivärinta [21], and Hettlich [16] in that we focus on the singularity itself, rather than on the behavior of the solution on the interior or on a subset of the boundary as the pole approaches the obstacle. See also [8] for a further discussion on the use of the fundamental solution in inverse scattering problems.

A similar formula for the singularity of the Green's function has been obtained by Potthast and Stratis [29] for a transmission obstacle with boundary conditions of the form

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

without exterior or interior potentials and with the same wave number on the interior as the exterior.

For the recovery of the boundary conditions and the Dirichlet-to-Neumann operator, the main idea is to analyze the symbol of the boundary operator,

$$(\Lambda^i \circ b - d)^{-1}(\Lambda^i \circ a - c),$$

on $\partial\Omega$ which we prove is uniquely determined from the singularity as well. Here $(\Lambda^i \circ b - d)^{-1}$ really represents a parametrix of the pseudodifferential operator $\Lambda^i \circ b - d$ (the invertibility does not necessarily hold). We show that one can extract all the necessary information, i.e., the interface conditions and the symbol of the Dirichlet-to-Neumann operator, from this operator, where to explicitly calculate the full symbol we use the full symbol of the Dirichlet-to-Neumann as derived in [12] and [23].

3. Preliminaries. In this section we collect some basic facts about semigeodesic coordinates, layer potentials, and wave front sets that will be fundamental for our analysis.

First, we recall the semigeodesic (i.e., boundary normal) coordinates at the surface $\partial\Omega$. In semigeodesic coordinates around some $p \in \partial\Omega$, $\partial\Omega$ has coordinates $(x', 0)$ and the outward pointing normal to the surface points in the positive x_n direction. Let us briefly recall their construction. For each $q \in \partial\Omega$, define $\gamma_q : (-\epsilon, \epsilon) \to \mathbb{R}^n$ to be the unit-speed geodesic with q at $\gamma_q(0)$ and normal to $\partial\Omega$ there and such that $t < 0$ implies $\gamma_q(0) \in \Omega$. Now let $(\hat{x}_1, \ldots, \hat{x}_{n-1})$ be any local coordinates for $\partial\Omega$ near some $p \in \partial\Omega$. We can extend the coordinates smoothly into $\mathbb{R}^n\backslash\Omega$ and into Ω by having them be constant along each normal geodesic γ_q. If we take $\hat{x}_n$ to be the parameter along each γ_q, then one can check that $(\hat{x}_1, \ldots, \hat{x}_{n-1}, \hat{x}_n)$ form coordinates in $\mathbb{R}^n$ in some neighborhood of p. Note that in these coordinates, $\hat{x}_n = 0$ implies $\hat{x} \in \partial\Omega$ and the metric will have the form

$$\hat{g} = \sum_{j,k=1}^{n-1} \hat{g}_{jk}(\hat{x})\, d\hat{x}^j \otimes d\hat{x}^k + d\hat{x}^n \otimes d\hat{x}^n.$$

Often when there is no chance of confusion we will use $(x_1, \ldots, x_n)$ also as the semi-geodesic coordinates (e.g., $\partial\Omega$ will have coordinates $(x', 0)$). Also, we shall denote a function f in the semi-geodesic coordinates $\hat{f}$.

It is particularly important to note that the selection of the coordinates on $\partial\Omega$ is arbitrary since often we will want to compare different or unknown metrics on $\partial\Omega$. In particular, assume g^e and g^i are metrics as above and let $(\hat{x}_1, \ldots, \hat{x}_{n-1})$ be any local coordinates on $\partial\Omega$. Then we can form separate semigeodesic coordinates for g^e and g^i in a neighborhood of some point $p \in \partial\Omega$ of the form $(\hat{x}_1, \ldots, \hat{x}_{n-1}, \hat{x}_{n^e})$ and $(\hat{x}_1, \ldots, \hat{x}_{n-1}, \hat{x}_{n^i})$.

We will use the following notation when using pseudodifferential operators. Below let $\mathscr{S}$ be the class of Schwartz functions and H^s the standard Sobolev spaces. See, for example, [10] or [30]. Let X be an open set in $\mathbb{R}^n$ and recall the symbol classes $S^m(X \times \mathbb{R}^n)$ of Hörmander [19].

Definition 3.1. *$a \in S^m(X \times \mathbb{R}^n)$ if $a \in C^\infty(X \times \mathbb{R}^n)$ and*

$$\left|\partial_\xi^\alpha \partial_x^\beta a(x, \xi)\right| \le C_{\alpha,\beta}(1 + |\xi|)^{m-|\alpha|}, \qquad x \in X, \quad \xi \in \mathbb{R}^n.$$

S^m is called the space of symbols of order m.

Definition 3.2. *If $a(x, \xi) \in S^m(X \times \mathbb{R}^n)$, the operator $a(x, D) \in L^m(X)$ is defined on $\mathscr{S}(X)$ by*

$$a(x, D)u(x) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{ix\cdot\xi} a(x, \xi)\tilde{u}(\xi)\, d\xi.$$

We say $a(x, D)$ is a pseudodifferential operator of order m. Moreover, $a(x, D)$ can be extended to a continuous operator from $H^s_{comp}(X) \to H^{s-m}_{loc}(X)$.

Finally, recall that the operator $a(x, D)$ is *properly supported* if its kernel,

$$K(x, y) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{i(x-y)\cdot\xi} a(x, \xi)\, d\xi$$

is properly supported. We say a kernel $K(x, y)$ is properly supported if, for every compact $M \subseteq X$, $\{(x, y) \in \text{supp } K : x \in M \text{ or } y \in M\}$ is compact.

Now let L be an operator of the form (1.1) and let $G(x, y, k)$ denote the fundamental solution for $L - k^2$ with the outgoing radiation conditions (as constructed in [26], for example). Then the single layer potential S on $\partial\Omega$ is defined by

$$S\phi(x) = \int_{\partial\Omega_j} G(x, y)\phi(y)\, dS_y. \tag{3.1}$$

This is smoothing from $\partial\Omega$ to $\mathbb{R}^n\backslash\partial\Omega$ which means it is of order $-\infty$. We are particularly interested in its behavior as a map from functions on the boundary into itself as obtained when x is restricted to the boundary. Let S_+, respectively, S_-, denote the operator obtained when x is restricted to $\partial\Omega$ limiting from above, respectively, below. Both are the sum of a properly supported elliptic pseudodifferential operator on $\partial\Omega$ of order -1 and a term with a smoothing kernel. The latter condition means the kernel is in $C^\infty(X\times X)$.

Although this is already well known, a brief outline of this fact is given below since it will often be necessary to refer to the construction of layer potentials. For simplicity assume we are on a half-space and that we are a given an operator of the form $L-k^2$ that is equal to zero outside some large ball with fundamental solution $G(x,y)$. Modulo smoothing terms, this is the situation obtained by restricting to a local coordinate chart using the semigeodesic coordinates. Now,

$$G(x,y)=\frac{1}{(2\pi)^n}\int_{\mathbb{R}^n}e^{i(x-y)\cdot\xi}r(x,\xi)\,d\xi+S(x,y), \tag{3.2}$$

where $S(x,y)$ is $C^\infty(\mathbb{R}^n\times\mathbb{R}^n)$, $r(x,\xi)$ is the symbol of the parametrix for $L-k^2$, and the integral is defined as an oscillatory integral. Let $\phi(x')\in C_0^\infty(\mathbb{R}^{n-1})$. Then the single layer potential is defined as follows:

$$\begin{aligned}\lim_{x_n\downarrow 0}&\int_{\mathbb{R}^{n-1}}G(x,y',0)\phi(y')\,dy'\\
&=\lim_{x_n\downarrow 0}\frac{1}{(2\pi)^n}\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}e^{i(x-y)\cdot\xi}r(x,\xi)(\phi(y')\otimes\delta(y_n))\,d\xi\,dy\\
&=\frac{1}{(2\pi)^{n-1}}\int_{\mathbb{R}^{n-1}}\int_{\mathbb{R}^{n-1}}e^{i(x'-y')\cdot\xi'}\left(\lim_{x_n\downarrow 0}\frac{1}{2\pi}\int_{\mathbb{R}}e^{ix_n\cdot\xi_n}r(x,\xi)\,d\xi_n\right)\phi(y')\,d\xi'\,dy'\\
&=\frac{1}{(2\pi)^{n-1}}\int_{\mathbb{R}^{n-1}}e^{i(x'-y')\cdot\xi'}\left(\int^+r(x',0,\xi)\,d\xi_n\right)\phi(y')\,dy',\end{aligned}$$

where $\int^+$ is equal to i times the sum of residues in the upper half-plane (see Chapter 18 of [19] for a more detailed account). Using that

$$r(x',0,\xi)=\frac{1}{\xi_n^2+\sum^{n-1}\hat{g}^{jk}(x')\xi_j\xi_k}+\frac{(\frac{\partial_{x_n}\sqrt{\hat{g}}}{\sqrt{\hat{g}}}(x')-2\hat{A}_n(x'))\xi_n}{(\xi_n^2+\sum^{n-1}\hat{g}^{jk}(x')\xi_j\xi_k)^2}+O\left(\frac{1}{\xi_n^4}\right),$$

where the $\hat{g}$ denotes we are in semigeodesic coordinates, it is easy to derive the principal symbol of the layer potential (as well as its jump relations). Let S_+ (respectively, S_-) denote the operator obtained from (3.1) by taking the limit of x from above (respectively, below). Also, let $(\frac{\partial S}{\partial\nu_x})_\pm$ denote the conormal derivative of the single layer potentials, i.e., (1.3) without the terms due to the magnetic potential.

Theorem 3.3. *$S_\pm$ and $(\frac{\partial S}{\partial\nu_x})_\pm$ are each equal to the sum of a properly supported elliptic pseudodifferential operator and a term with a smooth kernel. The principal symbols are*

$$\sigma_{-1}(S_\pm)=\frac{1}{2\sqrt{\sum_{j,k}^{n-1}\hat{g}^{jk}(x',0)\xi_j\xi_k}},$$

$$\sigma_0\left(\left(\frac{\partial S}{\partial\nu_x}\right)_\pm\right)=\mp\frac{1}{2}.$$

Also,

$$S_+ - S_- = 0, \qquad \left(\frac{\partial S}{\partial \nu_x}\right)_+ - \left(\frac{\partial S}{\partial \nu_x}\right)_- = -I. \tag{3.3}$$

We now recall some basic facts about wave front sets of distributions from [18] and [19]. Given a compactly supported distribution $v \in \mathcal{E}'(\mathbb{R}^n)$, let $\tilde{v}(\xi)$ be its Fourier transform and define the cone $\Sigma(v)$ to be all $\eta \in \mathbb{R}^n \backslash \{0\}$ having no conic neighborhood V such that

$$|\tilde{v}(\xi)| \le C_N(1+|\xi|)^{-N}, \qquad N \in N, \quad \xi \in V.$$

For X an open set in $\mathbb{R}^n$ and $u \in \mathcal{D}'(X)$, define for $x \in X$,

$$\Sigma_x(u) = \bigcap_{\substack{\phi \in C_0^\infty(X) \\ \phi(x) \ne 0}} \Sigma(\phi u).$$

Then,

$$WF(u) = \{(x,\xi) \in X \times (\mathbb{R}^n \backslash 0); \xi \in \Sigma_x(u)\}.$$

As a specific example, note $WF(\delta(x-y)) = y \times \mathbb{R}^n \backslash \{0\}$, since $\tilde{\delta} = 1$.

Now we consider characteristic sets and wave front sets of pseudodifferential operators. Let $A \in L^m$ be an operator with principal symbol $a(x,\xi) \in S^m(T^*(X))$. We say A is noncharacteristic at $(x_0,\xi_0) \in T^*(X)\backslash 0$ if $ab - 1 \in S^{-1}$ in a conic neighborhood of (x_0,ξ_0) for some $b \in S^{-m}$. Denote by Char A the set of characteristic points of A. In particular, note Char $P = \emptyset$ for any elliptic differential operator. This leads to an alternate definition of the wave front set:

$$WF(u) = \bigcap_{\substack{A \in L^m(X),\, m \in \mathbb{R} \\ A\, prop.\, supp. \\ Au \in C^\infty(X)}} \mathrm{Char}\ A,$$

from which it is easy to prove the following theorem.

THEOREM 3.4. *If $A \in L^m(X)$ is properly supported and $u \in \mathcal{D}'(X)$, then*

$$WF(u) \subset WF(Au) \cup \mathrm{Char} A.$$

As an example relevant to us, if $G(x,y)$ is the fundamental solution for $L - k^2$, then $WF(G(x,y)) = y \times \mathbb{R}^n \backslash \{0\}$ since $L - k^2$ is elliptic and $WF(\delta(x-y)) = y \times \mathbb{R}^n \backslash \{0\}$.

The last part of the theory we need to recall is the definition of the wave front set of a pseudodifferential operator. Let $A \in L^m(X)$ be a properly supported pseudodifferential operator with symbol $a(x,\xi) \in S^m$ as above. Then its kernel is

$$K_A(x,y) = \int_{R^n} e^{i(x-y)\cdot\xi} a(x,\xi)\, d\xi,$$

where the integral is defined as an oscillatory integral. Suppose Char $A = X \times \Gamma_A(x)$ for some $\Gamma_A(x) \subseteq \mathbb{R}^n \backslash \{0\}$. Then one can check that for any fixed $y \in X$,

$$WF(K_A(x,y)) = y \times \Gamma_A(y). \tag{3.4}$$

See, for example, Proposition 18.1.26 in [19]. In light of (3.2), this gives an alternate proof that $WF(G(x,y)) = y \times \mathbb{R}^n \backslash \{0\}$.

Our interest in wave front sets is due to the next two propositions.

PROPOSITION 3.5. $WF(\lim_{y \downarrow x^* \in \partial\Omega} G(x,y)|_{x \in \partial\Omega}) = x^* \times S^{n-2}$.

Proof. By a discussion analogous to that preceding Theorem 3.3, it is easy to see that modulo a smoothing kernel, $\lim_{y \downarrow x^* \partial\Omega} G(x,y)|_{x \in \partial\Omega}$ is the kernel of an elliptic pseudodifferential operator (in the variables x and x^*). Therefore the claim follows from (3.4). ∎

PROPOSITION 3.6. *Say $P_1(x,D)$ and $P_2(x,D)$ are two elliptic pseudodifferential operators on $\mathbb{R}^n$ and $u \in \mathcal{D}'(\mathbb{R}^n)$ is a distribution with wave front set $WF(u) = 0 \times \mathbb{R}^n \backslash \{0\}$. Then $P_1 u = P_2 u + f$ for some $f \in C^\infty(\mathbb{R}^n)$ implies that $\sigma(P_1)(0,\xi) = \sigma(P_2)(0,\xi)$.*

Proof. This is an immediate consequence of Theorem 3.4. ∎

4. The direct problem. Consider scattering from a general transmission obstacle as in (1.4)–(1.7). Because it significantly complicates the problem when b is both zero and nonzero, we initially assume that either

$$b(x) = 0 \text{ for all } x \in \partial\Omega \tag{4.1}$$

or

$$b(x) \neq 0 \text{ for all } x \in \partial\Omega. \tag{4.2}$$

In section 4.1, we will solve the direct problem, assuming T satisfies (2.2), (2.6), and (4.1), or (4.2), and then in section 4.2 we consider the nonself-adjoint problem when just (2.2) and (4.1) or (4.2) is satisfied. Finally, in section 4.3 we solve the problem for general scatterers when only (2.2) and (2.3) or (2.4) hold by considering a mixed boundary value problem and using the results of the previous sections.

First, recall the following lemma from [7] and [26].

LEMMA 4.1. *If u satisfies $(-\triangle - k^2)u = 0$ outside of B_S and u satisfies the outgoing radiation conditions*

$$\frac{\partial}{\partial r} u(x) - iku(x) = o\left(\frac{1}{r^{\frac{n-1}{2}}}\right), \quad u = O\left(\frac{1}{r^{\frac{n-1}{2}}}\right), \tag{4.3}$$

then

$$u = a(\theta)\frac{e^{ik|x|}}{|x|^{\frac{n-1}{2}}} + O\left(\frac{1}{|x|^{\frac{n+1}{2}}}\right) \tag{4.4}$$

when $|x| \to \infty$, $\theta = \frac{x}{|x|}$.

Now consider the boundary value problem,

$$(L^e - k^2)u(x) = f^e(x) \text{ in } \mathbb{R}^n \backslash \Omega, \tag{4.5}$$

$$(L^i - k^2)u(x) = f^i(x) \text{ in } \Omega, \tag{4.6}$$

$$\begin{pmatrix} u_- \\ \gamma_-^{1,i} u \end{pmatrix} = \begin{pmatrix} a(x) & b(x) \\ c(x) & d(x) \end{pmatrix} \begin{pmatrix} u_+ \\ \gamma_+^{1,e} u \end{pmatrix}, \tag{4.7}$$

$$\frac{\partial}{\partial r} u(x) - iku(x) = o\left(\frac{1}{r^{\frac{n-1}{2}}}\right), \quad u = O\left(\frac{1}{r^{\frac{n-1}{2}}}\right), \tag{4.8}$$

where $f^e \in C_0^\infty(\mathbb{R}^n \backslash \Omega)$ and $f^i \in C^\infty(\Omega)$.

4.1. Dissipative conditions. We begin with the uniqueness.

LEMMA 4.2. *Say u satisfies* (4.5)–(4.8) *with $f^e=0$, $f^i=0$ and T satisfies* (2.2) *and* (2.6)*. Then u is equal to* 0 *in $\mathbb{R}^n$.*

Proof. Applying Green's formula to the identity,

$$0=\int_{B_R\backslash\Omega}(L^e-k^2)u\overline{u}\,\sqrt{g^e}\,dx-\int_{\Omega}(L^i-k^2)u\overline{u}\,\sqrt{g^i}\,dx,$$

taking the imaginary part of both sides, plugging the expansion (4.4) into the integral over ∂B_R, and letting R tend to infinity, we obtain

$$\int_{S^{n-1}}k\,|a(\theta)|^2\,d\theta=\ \mathrm{Im}\ \left[\int_{\partial\Omega}\gamma_+^{1,e}u\overline{u_+}\,\sqrt{\tilde{g}^e}dS_x-\int_{\partial\Omega}\gamma_-^{1,i}u\overline{u_-}\,\sqrt{\tilde{g}^i}dS_x\right].\tag{4.9}$$

Using the boundary conditions (4.7) and also (2.2) and (2.6),

$$\begin{aligned}\gamma_-^{1,i}u\overline{u_-}\sqrt{\tilde{g}^i}&=\left[\left(cu_++d\gamma_+^{1,e}u\right)\overline{\left(au_++b\gamma_+^{1,e}u\right)}\right]\sqrt{\tilde{g}^i}\\&=\left[ac\,|u_+|^2+\overline{b}d\left|\gamma_+^{1,e}u\right|^2+ad\overline{u_+}\gamma_+^{1,e}u+\overline{b}cu_+\overline{\gamma_+^{1,e}u}\right]\sqrt{\tilde{g}^i}\\&=ac\,|u_+|^2\sqrt{\tilde{g}^i}+\overline{b}d\left|\gamma_+^{1,e}u\right|^2\sqrt{\tilde{g}^i}+\overline{u_+}\gamma_+^{1,e}u\sqrt{\tilde{g}^e}\\&\qquad+2\overline{b}c\ \mathrm{Re}\ u_+\overline{\gamma_+^{1,e}u}\sqrt{\tilde{g}^i}.\end{aligned}$$

Plugging this into (4.9) and utilizing that $-\overline{b}d$ and $-ac$ are each less than zero by (2.2) and (2.6), we get

$$\begin{aligned}\int_{S^{n-1}}k\,|a(\theta)|^2\,d\theta&\leq\ \mathrm{Im}\ \int_{\partial\Omega}-2\overline{b}c\ \mathrm{Re}\ u_+\overline{\gamma_+^{1,e}u}\sqrt{\tilde{g}^i}\,dS_x\\&\leq 0.\end{aligned}$$

Note $\overline{b}c$ has no imaginary part as a result of (2.6). Therefore $u=O(\frac{1}{|x|^{\frac{n+1}{2}}})$ and so Rellich's lemma and unique continuation imply $u=0$ in $\mathbb{R}^n\backslash\Omega$ (see, e.g., [26]). The boundary conditions give us that $u_-|_{\partial\Omega}=0$ and $\gamma_-^{1,i}u=0$, which by the uniqueness of the Cauchy problem implies $u=0$ in Ω (see, for example, [17]). ∎

To prove Theorem 2.1 assuming (2.2), (2.6), and (4.1) or (4.2), we will construct the Green's function for (4.5)–(4.8).

THEOREM 4.3. *Assume T satisfies* (2.2), (2.6)*, and* (4.1) *or* (4.2)*. Then there exists unique $G^e(x,y,k)$ and $G^i(x,y,k)$ that solve* (4.5)–(4.8) *with $f^e(x)=\delta(x-y)$ and $f^i(x)=\delta(x-y)$.*

Proof. Let $\tilde{G}^e(x,y)$ and $\tilde{G}^i(x,y)$ be the outgoing the Green's functions for (L^e-k^2) and (L^i-k^2) in $\mathbb{R}^n$ as constructed in [26]. Note $\tilde{G}^e(x,y)=E_+(x-y)$ for x,y large, where $E_+(x-y)$ is the outgoing fundamental solution of the Helmholtz equation. We look for G^e and G^i in the form

$$\begin{aligned}G^e(x,y)&=\tilde{G}^e(x,y)+S^e\phi^e(x,y),\\G^i(x,y)&=\tilde{G}^i(x,y)+S^i\phi^i(x,y),\end{aligned}$$

where S^e and S^i are single layer potentials for $\tilde{G}^e$ and $\tilde{G}^i$ on $\partial\Omega$ as defined in section 3. Fix $y\in\mathbb{R}^n\backslash\Omega$. Then the boundary conditions (4.7) give us the following system of pseudodifferential equations on $\partial\Omega$:

$$\begin{pmatrix} S^i_- & -aS^e_+ - b\gamma^{1,e}_+ S^e \\ \gamma^{1,i}_- S^i & -cS^e_+ - d\gamma^{1,e}_+ S^e \end{pmatrix} \begin{pmatrix} \phi^i(x,y) \\ \phi^e(x,y) \end{pmatrix} = \begin{pmatrix} -\tilde{G}^i(x,y)|_{\partial\Omega} + a\tilde{G}^e(x,y)|_{\partial\Omega} + b\gamma^{1,e}_+ \tilde{G}^e \\ -\gamma^{1,i}_- \tilde{G}^i(x,y) + c\tilde{G}^e(x,y)|_{\partial\Omega} + d\gamma^{1,e}_+ \tilde{G}^e \end{pmatrix}. \tag{4.10}$$

We want to show that the matrix B of pseudodifferential operators on the left-hand side is uniquely invertible. Note B is bounded $L^2(\partial\Omega)\otimes L^2(\partial\Omega) \to L^2(\partial\Omega)\otimes L^2(\partial\Omega)$. In Corollary 4.6 we show that B is Fredholm with index 0.

Assuming this for now, we have only to show that B is injective. Say first that k^2 is not a Dirichlet eigenvalue of L^e on Ω. In this case, suppose the right-hand side of (4.10) is equal to zero and let $y^e(x) = S^e_+\phi^e(x)$ and $y^i(x) = S^i_-\phi^i(x)$. Then y^e, y^i solve the transmission problem (4.5)–(4.8) with $f^e = 0$ and $f^i = 0$. Lemma 4.2 implies

$$\begin{aligned} y^e(x) &= 0 \text{ in } \mathbb{R}^n\backslash\Omega, \\ y^i(x) &= 0 \text{ in } \Omega. \end{aligned}$$

Using the jump relations given in Theorem 3.3, we see that y^i solves $(L^i - k^2)y^i = 0$ in $\mathbb{R}^n\backslash\Omega$ and $y^i|_{\partial\Omega} = 0$. Therefore it is easy to see that $y^i = 0$ in $\mathbb{R}^n\backslash\Omega$ (see, for example, [26]) and so another application of the jump relations implies $\phi^i = 0$. Similarly, y^e solves $(L^e - k^2)y^e = 0$ in Ω and $y^e|_{\partial\Omega} = 0$. Therefore since we are assuming k^2 is not a Dirichlet eigenvalue for L^e on Ω, y^e must be identically zero in Ω and another application of the jump relations proves that $\phi^e = 0$.

In the case that k^2 is a Dirichlet eigenvalue for L^e on Ω there are a number of methods we could use to solve the problem. One would be to use different layer potentials (like the ones used in [26]) or use a sum of single and double layer potentials as is utilized in [6]. Instead, here we will alter the equation on the interior of Ω following the approach taken in [11]. In particular, pick some constant q such that k^2 is not a Dirichlet eigenvalue of $L^e + q$ on Ω. Let χ_Ω be the distribution equal to 1 on $\overline{\Omega}$ and 0 everywhere else. One can construct the Green's function $G^e(x,y)$ to solve

$$(L^e + \chi_\Omega q - k^2)G^e(x,y) = \delta(x-y)$$

as in [11] and it will have the same jump relations as those given in Theorem 3.3. Therefore the preceding argument with this alteration to the exterior Green's function can be used. □

Define the Green's function $G(x,y)$ for $y \notin \partial\Omega$ by

$$G(x,y) = \begin{cases} G^e(x,y), & x \in \mathbb{R}^n\backslash\Omega; \\ G^i(x,y), & x \in \Omega. \end{cases} \tag{4.11}$$

In particular, for $f(x) \in C_0^\infty(\mathbb{R}^n)$,

$$Gf(x) = \begin{cases} \int_{\mathbb{R}^n} G^e(x,y)f(y)\,dy, & x \in \mathbb{R}^n\backslash\Omega; \\ \int_{\mathbb{R}^n} G^i(x,y)f(y)\,dy, & x \in \Omega, \end{cases} \tag{4.12}$$

where to avoid $y \in \partial\Omega$, we can interpret the integrals as being over $\mathbb{R}^n\backslash\partial\Omega$ since $\partial\Omega$ is a set of Lebesgue measure zero.

Proof of Theorem 2.1 (*in the dissipative case and assuming either* (4.1) *or* (4.2)). Letting $v(x,k\omega) = G[-(L^e - k^2)e^{ik\omega\cdot x}]$, we look for a solution to (1.4)–(1.7) of the form

$$w(x,k\omega) = \begin{cases} v(x,k\omega) + S^e\psi^e(x), & x \in \mathbb{R}^n\backslash\Omega; \\ v(x,k\omega) + S^i\psi^i(x), & x \in \Omega, \end{cases}$$

for two smooth functions $\psi^e(x)$ and $\psi^i(x)$ on $\partial\Omega$. Plugging $w(x,k\omega)$ into (1.6) leads to a system of pseudodifferential equations on $\partial\Omega$ of the form (4.10) with, of course, different functions on the right-hand side. This is solvable as in the proof of the previous theorem. That $w(x,k\omega)$ has the correct asymptotics follows from Lemma 4.1. $\square$

4.2. General nonself-adjoint problem. First, we find the boundary conditions of the adjoint problem.

THEOREM 4.4. *The adjoint boundary value problem to* (4.5)–(4.8) *has transfer matrix*

$$T^*(x) = \frac{\sqrt{\tilde{g}^e}}{\sqrt{\tilde{g}^i}} \frac{1}{ad - \overline{bc}} \begin{pmatrix} a(x) & \overline{b(x)} \\ \overline{c(x)} & d(x) \end{pmatrix}$$

in place of T and the solution v must satisfy the incoming radiation conditions

$$\frac{\partial}{\partial r}v(x) + ikv(x) = o\left(\frac{1}{r^{\frac{n-1}{2}}}\right),\ v = O\left(\frac{1}{r^{\frac{n-1}{2}}}\right) \tag{4.13}$$

instead of the outgoing radiation conditions (4.8).

Note $(T^*)^* = T$.

Proof. Assume u is a solution to (4.5)–(4.8) for some f^e and f^i. Then the adjoint boundary conditions are, by definition, the conditions that a function v in the domain of $L^e - k^2$ on $\mathbb{R}^n\backslash\Omega$ and in the domain of $L^i - k^2$ on Ω must satisfy so that

$$\begin{aligned} &\int_{\mathbb{R}^n\backslash\Omega}(L^e - k^2)u\bar{v}\sqrt{g^e}\,dx + \int_\Omega (L^i - k^2)u\bar{v}\sqrt{g^i}\,dx \\ &= \int_{\mathbb{R}^n\backslash\Omega} u\overline{(L^e - k^2)v}\sqrt{g^e}\,dx + \int_\Omega u\overline{(L^i - k^2)v}\sqrt{g^i}\,dx. \end{aligned} \tag{4.14}$$

Here $(L^e - k^2)$ and $(L^i - k^2)$ are self-adjoint on $L^2(\mathbb{R}^n, \sqrt{g^e}\,dx)$ and $L^2(\mathbb{R}^n, \sqrt{g^i}\,dx)$, respectively.

Integrating over some large ball B_R instead of $\mathbb{R}^n$ in (4.14) and applying Green's formula, we obtain

$$\begin{aligned} &-\int_{\partial B_R}\frac{\partial u}{\partial r}\bar{v}\,dS_x + \int_{\partial\Omega}\gamma_+^{1,e}u\bar{v}\sqrt{\tilde{g}^e}\,dS_x - \int_{\partial\Omega}\gamma_-^{1,i}u\bar{v}\sqrt{\tilde{g}^i}\,dS_x \\ &= -\int_{\partial B_R} u\overline{\frac{\partial v}{\partial r}}\,dS_x + \int_{\partial\Omega} u\overline{\gamma_+^{1,e}v}\sqrt{\tilde{g}^e}\,dS_x - \int_{\partial\Omega} u\overline{\gamma_-^{1,i}v}\sqrt{\tilde{g}^i}\,dS_x. \end{aligned}$$

It is easy to see that in order for the integrals over ∂B_R to vanish, v must satisfy the incoming radiation conditions. Assuming this to be the case and taking the limit as $R \to \infty$, the remaining integrals over $\partial\Omega$ will give us the form of T^*. In particular, we must have

$$\left\langle \gamma_+^{1,e}u, v_+\right\rangle_{\sqrt{\tilde{g}^e}} - \left\langle \gamma_-^{1,i}u, v_-\right\rangle_{\sqrt{\tilde{g}^i}} = \left\langle u_+, \gamma_+^{1,e}v\right\rangle_{\sqrt{\tilde{g}^e}} - \left\langle u_-, \gamma_-^{1,i}v\right\rangle_{\sqrt{\tilde{g}^i}},$$

where we use the notation $\langle f, g\rangle_h = \int_{\partial\Omega} f\overline{g}\, h\, dS_x$. Plugging in the boundary conditions (4.7) this becomes

$$\left\langle \frac{1}{ad-bc}(-cu_- + a\gamma_-^{1,i}u), v_+ \right\rangle_{\sqrt{\tilde{g}^e}} - \left\langle \gamma_-^{1,i}u, v_- \right\rangle_{\sqrt{\tilde{g}^i}}$$
$$= \left\langle \frac{1}{ad-bc}(du_- - b\gamma_-^{1,i}u), \gamma_+^{1,e}v \right\rangle_{\sqrt{\tilde{g}^e}} - \left\langle u_-, \gamma_-^{1,i}v \right\rangle_{\sqrt{\tilde{g}^i}}$$

and reorganizing,

$$\left\langle u_-, \frac{1}{\overline{ad-bc}}(-\overline{c}v_+ - d\gamma_+^{1,e}v^e) \right\rangle_{\sqrt{\tilde{g}^e}} + \left\langle \gamma_-^{1,i}u, \frac{1}{\overline{ad-bc}}(av_+ + \overline{b}\gamma_+^{1,e}v) \right\rangle_{\sqrt{\tilde{g}^e}}$$
$$= -\left\langle u_-, \gamma_-^{1,i}v \right\rangle_{\sqrt{\tilde{g}^i}} + \left\langle \gamma_-^{1,i}u, v_- \right\rangle_{\sqrt{\tilde{g}^i}}.$$

It follows that

$$\begin{pmatrix} v_- \\ \gamma_-^{1,i}v \end{pmatrix} = \frac{\sqrt{\tilde{g}^e}}{\sqrt{\tilde{g}^i}} \frac{1}{\overline{ad-bc}} \begin{pmatrix} a(x) & \overline{b(x)} \\ \overline{c(x)} & d(x) \end{pmatrix} \begin{pmatrix} v_+ \\ \gamma_+^{1,e}v \end{pmatrix}$$

which completes the proof. □

It follows from this result that the problem is self-adjoint when both (2.2) and (2.6) are satisfied and b and c have no imaginary component. This is not surprising since Im $\overline{u_+}\gamma_+^{1,e}u$ and Im $\overline{u_-}\gamma_-^{1,i}u$ physically represent the exterior and interior flux at the boundary and, as shown in the proof of Lemma 4.2, they are equal precisely when (2.2) and (2.6) hold and the imaginary components of the off-diagonal terms are zero.

We now finish the proof of Theorem 2.1 by constructing the Green's function for (4.5)–(4.8) when the transfer matrix is assumed only to satisfy (2.2) and (4.1) or (4.2).

THEOREM 4.5. *Assume T satisfies* (2.2) *and* (4.1) *or* (4.2). *Then for all but a discrete set of $k > 0$, there exists $G^e(x, y, k)$ and $G^i(x, y, k)$ that solve* (4.5)–(4.8) *with $f^e = \delta(x - y)$ and $f^i = \delta(x - y)$.*

Proof. We will prove this by constructing $G(x, y, z)$ for all but a discrete set of z in $\mathbb{C}$. Note that for all but a discrete set of $z \in \mathbb{C}$, there exists fundamental solutions for $L^e - z^2$ and $L^i - z^2$ in $\mathbb{R}^n$, $G^e(x, y, z)$, and $G^i(x, y, z)$, where for $z > 0$ these fundamental solutions have the outgoing radiation conditions (see [26]). In fact we know the poles must occur on the imaginary axis. Define $\mathcal{K}$ to be this discrete subset of $\mathbb{C}$. In order to solve the transmission problem for z outside of this discrete set, as in the proof of Theorem 4.3 the problem is reduced to showing the invertibility of the matrix of pseudodifferential operators on $\partial\Omega$:

$$M = \begin{pmatrix} S_-^i & -aS_+^e - b\gamma_+^{1,e}S^e \\ \gamma_-^{1,i}S^i & -cS_+^e - d\gamma_+^{1,e}S^e \end{pmatrix}. \tag{4.15}$$

Here the layer potentials are with respect to the operators $G^e(x, y, z)$ and $G^i(x, y, z)$ on $\partial\Omega$. Let

$$\Lambda = \left\{ z \in \mathbb{C} : \frac{\pi}{2} - \epsilon < \arg z < \frac{\pi}{2} + \epsilon \right\}$$

and let $(x_1, \ldots, x_{n-1}, 0)$ be local coordinates for $\partial\Omega$ around some point $y \in \partial\Omega$. Furthermore, let $\hat{g}_i^{jk}$ and $\hat{g}_e^{jk}$ be the restriction of the interior and exterior metrics to $\partial\Omega$ with respect to this coordinate system. Then in some neighborhood of y, the layer potentials have the following symbols in the calculus of the pseudodifferential operators with a parameter for $z \in \Lambda\backslash\mathcal{K}$ (see [26] and [30]):

$$\begin{aligned}
S^i_-(x', \xi', z) &= \frac{1}{2\sqrt{\sum_{j,k=1}^{n-1} \hat{g}_i^{jk}(x', 0)\xi_j\xi_k - z^2}} + R^i_{-2}(x', \xi', z),\\
S^e_+(x', \xi', z) &= \frac{1}{2\sqrt{\sum_{j,k=1}^{n-1} \hat{g}_e^{jk}(x', 0)\xi_j\xi_k - z^2}} + R^e_{-2}(x', \xi', z),\\
\gamma_+^{1,e} S^e(x', \xi', z) &= \frac{1}{2} + T^i_{-1}(x', \xi', z),\\
\gamma_-^{1,i} S^i(x', \xi', z) &= -\frac{1}{2} + T^e_{-1}(x', \xi', z).
\end{aligned}$$

Here R^i_{-2} and R^e_{-2} are bounded pseudodifferential operators (quantized in the calculus with a parameter) on $\partial\Omega$ of order -2, and T^e_{-1} and T^i_{-1} represent operators of order -1.

Therefore,

$$M(x', \xi', z) = \begin{pmatrix} \frac{1}{2\ell_1^i(z)} & -\frac{a}{2\ell_1^e(z)} + \frac{b}{2} - bT^e_{-1}(z) \\ \frac{1}{2} + T^i_{-1}(z) & -\frac{c}{2\ell_1^e(z)} + \frac{d}{2} + dT^e_{-1}(z) \end{pmatrix} + R_{-2}(z),$$

where $\ell_1^i(x', \xi', z) = (\sum_{j,k=1}^{n-1} \hat{g}_i^{jk}(x', 0)\xi_j\xi_k - z^2)^{\frac{1}{2}}$ and $\ell_1^e(x', \xi', z)$ is defined similarly. Also, $R_{-2}(x', \xi', z) \in S^{-2}(\partial\Omega \times \partial\Omega \times \mathbb{R}^{n-1}, \Lambda)$.

Assume $z \in \Lambda\backslash\mathcal{K}$ and $|z| \geq R$ for some large R. Since we are assuming either (4.1) or (4.2), it is not difficult to construct an operator $B(z) \in L^{-2}(\partial\Omega \times \partial\Omega, \Lambda)$ with symbol $B(x', \xi', z) \in S^{-2}(\partial\Omega \times \partial\Omega \times \mathbb{R}^{2n}, \Lambda)$ such that

$$M(z)B(z) = I + T(z), \tag{4.16}$$

where

$$\|T(z)\|_{L^2(\partial\Omega)\times L^2(\partial\Omega)} \leq \frac{C}{1 + |z|}$$

and T is compact on the same space. Thus, $(I+T(z))^{-1}$, and hence $M(z)^{-1}$, exists for $z \in (\Lambda\backslash\mathcal{K}) \cap \{|z| \geq R\}$ for some large $R > 0$. Now let $B'(z)$ be a parametrix for $M(z)$ constructed using the regular pseudodifferential calculus (i.e., the Kohn–Nirenberg quantization). Then, $M(z)B'(z) = I + T'(z)$ for some compact $T'(z)$. Moreover, $(I + T'(z))^{-1}$ exists for $z \in (\Lambda\backslash\mathcal{K}) \cap \{|z| \geq R\}$ since $M(z)^{-1}$ is invertible there. Since $T'(z)$ is an operator-valued meromorphic function of z on $L^2(\partial\Omega) \times L^2(\partial\Omega)$, $(I + T'(z))^{-1}$, and hence $M(z)^{-1}$, exists for all but a discrete set of $z \in \mathbb{C}$ (see [31]). Note it was necessary to consider separately $T(z), B(z)$ and $T'(z), B'(z)$ since the construction of $B(z)$ and $T(z)$ becomes problematic on the positive real axis (because the denominator of the principal symbol of the single layer potential may be zero there in the parameter-dependent calculus). ∎

COROLLARY 4.6. *$M(z)$ is Fredholm with index 0 for all $z \in \mathbb{C}\backslash\mathcal{K}$.*

Proof. This is true for $M(z)$ with $z \in (\Lambda\backslash\mathcal{K}) \cap \{|z| \geq R\}$ since $M(z)$ is Fredholm by (4.16) and invertible. Moreover, $M(z_1) - M(z_2)$ is compact on $L^2(\partial\Omega) \times L^2(\partial\Omega)$ since z appears only on terms of order -1. ∎

The following theorem relates the Green's function and its adjoint. Note that by arguments similar to those given in [26], the set of $k > 0$ on which $G(x, y, k)$ and $G^*(x, y, k)$ exist coincides.

THEOREM 4.7. *Say $k > 0$ is such that $G(x, y, k)$ and $G^*(x, y, k)$ exist and assume $x, y \notin \partial\Omega$. Then $G^*(x, y, k) = \overline{G(y, x, k)}$.*

Proof. Define L to be the operator equal to L^e in $\mathbb{R}^n \backslash \Omega$ and equal to L^i in Ω. Let $f, g \in C_0^\infty(\mathbb{R}^n)$. Then

$$\begin{aligned}
&\langle f, G^* g\rangle_{L^2(B_R)} \\
&\quad = \left\langle (L - k^2) G f, G^* g\right\rangle_{L^2(B_R)} \\
&\quad = \left\langle G f, (L^* - k^2) G^* g\right\rangle_{L^2(B_R)} - \int_{\partial B_R} \frac{\partial G f}{\partial r} \overline{G^* g} - G f \frac{\overline{\partial G^* g}}{\partial r}\, dS_x.
\end{aligned}$$

As R tends to infinity the integral over ∂B_R tends to zero since Gf and $\overline{G^* g}$ satisfy the outgoing radiation conditions. Thus,

$$\langle f, G^* g\rangle_{L^2(\mathbb{R}^n \backslash \Omega)} = \langle G f, g\rangle_{L^2(\mathbb{R}^n \backslash \Omega)}$$

for all $f, g \in C_0^\infty(\mathbb{R}^n \backslash \Omega)$ from which it follows that $G^*(x, y) = \overline{G(y, x)}$. ◻

Note in the self-adjoint case this proves $G(x, y, k) = \overline{G(y, x, k)}$ for all $k > 0$.

4.3. Direct problem for general obstacles. We now prove the full result for obstacles satisfying only (2.2) and (2.3) or (2.4).

Proof of Theorem 2.1 *for general transmission obstacles.* Let T be a transfer matrix satisfying (2.2) and (2.3) or (2.4). It suffices to solve (4.10). Let $b'(x)$ be a smooth extension of $b(x)$ onto all $\partial\Omega$ which equals $b(x)$ on Δ. Then define

$$M = \begin{pmatrix} S_-^i & -a S_+^e - b' \gamma_+^{1,e} S^e \\ \gamma_-^{1,i} S^i & -c S_+^e - d \gamma_+^{1,e} S^e \end{pmatrix} \tag{4.17}$$

and

$$M_0 = \begin{pmatrix} S_-^i & -a S_+^e \\ \gamma_-^{1,i} S^i & -c S_+^e - d \gamma_+^{1,e} S^e \end{pmatrix}. \tag{4.18}$$

As in [26] (see also [10]), solving (4.10) reduces to finding $\vec{g} \in \mathring{H}^{\frac{1}{2}}(\Delta)$, solving

$$\rho_\Delta (M M_0^{-1}) \vec{g} = \vec{f} \tag{4.19}$$

for $\vec{f} \in H^{-\frac{1}{2}}(\Delta)$. Here $\mathring{H}^s(\Delta)$ denotes the subspace of $H^s(\Delta)$ consisting of functions with support in $\overline{\Delta}$. It is easy to see that

$$M M_0^{-1} = \begin{pmatrix} I + Q_1 & P_0 \\ 0 & I \end{pmatrix},$$

where P_0 and Q_1 are pseudodifferential operators of orders 0 and 1 on $\partial\Omega$. In fact, after a trivial but lengthy calculation we find (in semigeodesic coordinates)

$$\sigma(Q_1)(x', \xi') = -b(x') \left(\frac{a(x')}{\ell_1^e(x', \xi')} + \frac{d(x')}{\ell_1^i(x', \xi')} \right)^{-1} + \cdots.$$

Since Q_1 has a strongly elliptic principal symbol by (2.4), it follows that $(I+Q_1)$ is Fredholm with index 0. Therefore $\rho_\Delta(MM_0^{-1})$ is also Fredholm with index 0. In the self-adjoint case, or more generally whenever (2.6) holds, the necessary uniqueness follows by Lemma 4.2 and the proof in this case is finished.

For the nonself-adjoint case, since Q_1 is elliptic, we can find a matrix of pseudo differential operators B such that

$$\rho_\Delta(MM_0^{-1})Bf = \rho_\Delta(I+T_{-1})f.$$

Therefore using the parameter-dependent pseudodifferential calculus as in section 4.2, we can prove invertibility for all but a discrete set of $k>0$.

We remark that the ϕ^e and ϕ^i obtained by the above argument are both in $H^{\frac{1}{2}}(\partial\Omega)$, which implies the Green's functions $G^e(x,y)$ and $G^i(x,y)$ are in $H^2(\mathbb{R}^n\backslash\Omega)$ and $H^2(\Omega)$ away from the pole. This means the application of Green's formula is valid in the uniqueness proofs, or in establishing the scattering amplitude uniquely determines the Green's function (as below). See [26] for a more detailed discussion. $\square$

4.4. The Green's function and the scattering amplitude. Assume we are given two obstacles, $\{\Omega_1, T_1(x), L_1\}$ and $\{\Omega_2, T_2(x)\, L_2\}$, along with their respective scattering data, $a_1(\theta,\omega,k)$ and $a_2(\theta,\omega,k)$. We recall the following theorem from [26] (where Theorem 4.4 establishes that the adjoint the Green's function exists).

THEOREM 4.8. *$a_1(\theta,\omega,k)=a_2(\theta,\omega,k)$ for $\theta,\omega\in S^{n-1}$ implies $G_1(x,y,k)=G_2(x,y,k)$ for $x,y\in\mathbb{R}^n\backslash\overline{(\Omega_1\cup\Omega_2)}$ and $x\neq y$.*

Strictly speaking the proof of the theorem requires all the obstacles to be impenetrable. However, the alterations for the case of transmission obstacles are minor. Note this theorem holds even when the boundary conditions are nonself-adjoint.

This also establishes the equivalence of the scattering problem and the inverse boundary value problem on the boundary of a ball large enough to contain all the scatterers. See [26].

5. Inverse problem for constant isotropic media with electric potentials. For this entire section assume $L^e=-\triangle+V^e(x)-k^2$ and $L^i=-\triangle+V^i-k^2$. By assuming the metrics are the same (and known) at the surface of the scatterer and that there are no magnetic potentials, we avoid some difficult technicalities while still preserving the general idea of all of the proofs. After recalling the symbol of the Dirichlet-to-Neumann operator in section 6, as well as formulating some basic properties on boundary determination, we will be prepared for the more exhaustive analysis required in the case of a nontrivial metric and electromagnetic potentials given in section 7.

5.1. Location. First, we analyze the behavior of the singularity of the Green's function in a neighborhood of the boundary of an obstacle.

We say the obstacle Ω is *well defined at* x_0 if $x_0\notin\partial\Delta$ and at least one of the following holds:

1. $T(x_0)\neq\tau I$ for some $\tau>0$.
2. V^e and V^i do not smoothly connect along the normal direction at x_0.

The following lemma proves a singularity develops in $G^e(x,y)-\tilde{G}^e(x,y)$ as y approaches $x\in\partial\Omega$, which is well defined, and the subsequent corollary gives an exact formula for the principal symbol of the single layer potential of $G^e(x,y)$ obtained as $y\downarrow x^*\in\partial\Omega$.

LEMMA 5.1. *$\partial\Omega$ is well defined at x_0 if and only if*

$$\lim_{y\downarrow x^*\in\partial\Omega}\big(G^e(x,y)-\tilde{G}^e(x,y)\big)=\lim_{y\downarrow x^*\in\partial\Omega}S^e_+\phi^e(x,y)\notin C^\infty(B_\epsilon(x^*)\cap\partial\Omega).$$

Proof. Say first that k^2 is not a Dirichlet eigenvalue of either $-\triangle+V^e$ or $-\triangle+V^i$ on Ω. This restriction will be removed at the end of the proof. Note the Dirichlet-to-Neumann operator for $-\triangle+V^i(x)-k^2$ on Ω, $\Lambda^i(k)$ can be written $\gamma_-^{1,i}S^i(S_-^i)^{-1}$. Fix $y\in\mathbb{R}^n\backslash\overline{\Omega}$. We rewrite (4.10) in the form

$$\begin{aligned}&\Lambda^i\circ\big[aS_+^e\phi^e(x,y)+b\gamma_+^{1,e}S^e\phi^e(x,y)-\tilde{G}^i+a\tilde{G}^e(x,y)+b\gamma_+^{1,e}\tilde{G}^e(x,y)\big]\\&\quad=cS_+^e\phi^e(x,y)+d\gamma_+^{1,e}S^e\phi^e(x,y)-\gamma_-^{1,i}\tilde{G}^i(x,y)+c\tilde{G}^e(x,y)+d\gamma_+^{1,e}\tilde{G}^e(x,y).\end{aligned}$$

Define $\Lambda_e^e=-\gamma_+^{1,e}S^e(S_+^e)^{-1}$ to be the Dirichlet-to-Neumann operator of the exterior problem for $-\triangle+V^e(x)-k^2$ (with outgoing radiation conditions) and Λ^e the Dirichlet-to-Neumann operator for $-\triangle+V^e(x)-k^2$ on Ω. Using that $\Lambda^i\tilde{G}^i(x,y)=\gamma_-^{1,i}\tilde{G}^i(x,y)$, we can rewrite the above equation as

$$\begin{aligned}&\big[\Lambda^i\circ a-\Lambda^i\circ b\circ\Lambda_e^e-c+d\circ\Lambda_e^e\big](S_+^e\phi^e)(x,y)\\&\quad=-\big[\Lambda^i\circ a+\Lambda^i\circ b\circ\Lambda^e-c-d\circ\Lambda^e\big]\tilde{G}^e(x,y).\end{aligned}$$

Also since $\Lambda_e^e=-\Lambda^e+(S_\pm^e)^{-1}$,

$$\begin{aligned}&\big[\Lambda^i\circ a-\Lambda^i\circ b\circ\Lambda_e^e-c+d\circ\Lambda_e^e\big](S_+^e\phi^e)(x,y)\\&\quad=\big[d\circ(S_\pm^e)^{-1}-\Lambda^i\circ b\circ(S_\pm^e)^{-1}\big]\tilde{G}^e(x,y)\\&\qquad-\big[\Lambda^i\circ c-\Lambda^i\circ b\circ\Lambda_e^e-c+d\circ\Lambda_e^e\big]\tilde{G}^e(x,y).\end{aligned}$$

We want to invert the operator on the left-hand side, at least, at the symbol level, in order to obtain a formula for the singularity of $S_+^e\phi^e(x,y^*)$ as $y\downarrow x^*\in\partial\Omega$. Now,

$$\begin{aligned}&\sigma\big[\Lambda^i\circ a-\Lambda^i\circ b\circ\Lambda_e^e-c+d\circ\Lambda_e^e\big]\\&\quad=a\ell_1^e-b\ell_1^e\ell_1^i+d\ell_1+bm_{-1}(x,\xi)+T_{-2}(x,\xi),\end{aligned}$$

where $m_{-1}(x,\xi)$ is some operator of order -1. Let P be a local parametrix of $\big[\Lambda^i\circ a-\Lambda^i\circ b\circ\Lambda_e^e-c+d\circ\Lambda_e^e\big]$ which exists since we are assuming $x^*\notin\partial\Delta$. Then, in an appropriate neighborhood of x^*,

$$\begin{aligned}G^e(x,y)-\tilde{G}^e(x,y)&=S_+^e\phi^e(x,y)\\&\sim-\tilde{G}^e(x,y)+P\big[d\circ(S_\pm^e)^{-1}-\Lambda^i\circ b\circ(S_+^e)^{-1}\big]\tilde{G}^e(x,y),\end{aligned}\tag{5.1}$$

where $\sim$ denotes equality modulo C^∞. In the following corollary we use the above formula to find the principal symbol of the single layer potential of $G^e(x,y)$ obtained by taking the limit as $y\downarrow x^*\in\partial\Omega$. For now, it suffices to show that a singularity necessarily develops in (5.1) as $y\downarrow x^*\in\partial\Omega$, which will be the case unless

$$\sigma(P\big[d\circ(S_\pm^e)^{-1}-\Lambda^i\circ b\circ(S_+^e)^{-1}\big])=1\tag{5.2}$$

for all x in a neighborhood of x^*. This condition is sufficient by Proposition 3.6.

For all $x_0\in\partial\Omega$ near x^*, we see that equality of both sides of (5.2) is equivalent to

$$\sigma\big[\Lambda^i\circ a+\Lambda^i\circ b\circ\Lambda^e-c-d\circ\Lambda^e\big](x_0,\xi)=0$$

or

$$\sigma\big[\Lambda^i\circ a+\Lambda^i\circ b\circ\Lambda^e\big](x_0,\xi)=\sigma\big[c+d\circ\Lambda^e\big](x_0,\xi).$$

However, for this to be true we need $b(x_0) = 0$ and $\sigma(\Lambda^i \circ a)(x_0) = \sigma(d \circ \Lambda^e + c)(x_0)$. Using the composition formula for pseudodifferential operators, this becomes

$$\ell_1 a + \ell_0 a + \sum_j \partial_{\xi_j} \ell_1 D_{x_j} a + \cdots = d\ell_1 + d\ell_0 + c + \cdots .$$

From the coefficient of ℓ_1 we see that $a = d$ and it follows that $\sum \partial_{\xi_j} \ell_1 D_{x_j} a$ and c must equal zero. The former implies $a = \tau$ for some constant τ. Of course if the boundary conditions are self-adjoint, $\tau = 1$. Therefore $\sigma(\Lambda^i) = \sigma(\Lambda^e)$. It is well-known that the symbol of the Dirichlet-to-Neumann operator uniquely determines the electric potential, along with all its normal derivatives, at the surface (see [22] or the following section). This contradicts that the obstacle is well defined at x^* and proves a singularity must form in (5.1) as $y \downarrow x^* \in \partial\Omega$.

It still remains to consider the case when k^2 is a Dirichlet eigenvalue of $-\triangle + V^i(x)$ on Ω. In this case, let $\phi^i(x), \phi^e(x)$ be solutions to the equation

$$\begin{aligned} &\begin{pmatrix} S^i_- & -aS^e_+ - b\gamma^{1,e}_+ S^e \\ \gamma^{1,i}_- S^i & -cS^e_+ - d\gamma^{1,e}_+ S^e \end{pmatrix} \begin{pmatrix} \phi^i(x,y) \\ \phi^e(x,y) \end{pmatrix} \\ &\sim \begin{pmatrix} -\tilde{G}^i(x,y)|_{\partial\Omega} + a\tilde{G}^e(x,y)|_{\partial\Omega} + b\gamma^{1,e}_+ \tilde{G}^e \\ -\gamma^{1,i}_- \tilde{G}^i(x,y) + c\tilde{G}^e(x,y)|_{\partial\Omega} + d\gamma^{1,e}_+ \tilde{G}^e \end{pmatrix} \end{aligned} \tag{5.3}$$

in some neighborhood of x^*. Specifically, as discussed in Theorem 4.5, there exists a parametrix of the matrix of pseudodifferential operators on the left-hand side. We want ϕ^i and ϕ^e to be the functions obtained by applying this parametrix to the right-hand side. Since neither $(S^i_-)^{-1}$ nor $(S^e_+)^{-1}$ necessarily exists, we understand these terms to refer to parametrices too. The proof then proceeds in the same fashion. ∎

The following corollary of (5.1) gives the principal symbol of the transpose of the layer potential of $G^e(x,y)$ on $\partial\Omega$.

Corollary 5.2. *Let $(S^e_+)^t$ refer to the transpose of the single layer potential of $G^e(x,y)$ on $\partial\Omega$. Then if $b \neq 0$ at x_0,*

$$\sigma_{-1}((S^e_+)^t)(x_0,\xi') = \frac{1}{\sqrt{\sum_{j,k}^{n-1} \hat{g}^{jk}_e(x',0)\xi_j\xi_k}},$$

and if $b(x_0) = 0$,

$$\sigma_{-1}((S^e_+)^t)(x_0,\xi') = \frac{1}{(\frac{a}{d}+1)\sqrt{\sum_{j,k}^{n-1} \hat{g}^{jk}_e(x',0)\xi_j\xi_k}}$$

where we are in local, semigeodesic coordinates.

Note that if $T = I$, we recover the principal symbol of the usual single layer potential on $\partial\Omega$ for $\tilde{G}^e(x,y)$ as given in Theorem 3.3.

It is now easy to show that the location of the obstacle is uniquely determined everywhere it is well defined. Assume $\{\Omega_1, T_1(x), L_1\}$ and $\{\Omega_2, T_2(x), L_2\}$ are two obstacles well defined on $\Gamma_1 \subseteq \partial\Omega_1$ and $\Gamma_2 \subseteq \partial\Omega_2$. Let $G_1(x,y)$ and $G_2(x,y)$ be their respective Green's functions.

Theorem 5.3. *If $G^e_1(x,y) = G^e_2(x,y)$ for all $x,y \in \mathbb{R}^n \backslash \overline{(\Omega_1 \cup \Omega_2)}$ and $x \neq y$, then $\Gamma_1 = \Gamma_2$.*

Proof. Say there exists a point $x^* \in \Gamma_2$ such that $x^* \in \mathbb{R}^n \backslash (\Omega_1 \cup \overline{\Gamma_1})$, where the closure of Γ_1 is taken with respect to $\partial\Omega$. Fix some neighborhood Γ such that

$x^* \in \Gamma \subset \Gamma_2$ and $\Gamma \cap \overline{\Gamma_1} = \emptyset$. Then, by Lemma 5.1,

$$\lim_{y \downarrow x^*} G_2^e(x,y) - \tilde{G}^e(x,y) \notin C^\infty(\Gamma).$$

On the other hand, since x^* is a positive distance away from Γ_1, it is easy to see that

$$\lim_{y \downarrow x^*} G_1^e(x,y) - \tilde{G}^e(x,y) \in C^\infty(\Gamma).$$

This is a contradiction and proves the theorem. □

5.2. Transfer matrix and Dirichlet-to-Neumann operator. The following theorem proves the transfer matrix and the full symbol of the Dirichlet-to-Neumann operator are uniquely determined on the boundary. Also, if the Dirichlet-to-Neumann operator on the boundary exists, we show it is uniquely determined.

THEOREM 5.4. *Assume $G^e(x,y)$ is known for all x and y in $\mathbb{R}^n \backslash \Omega$. This uniquely determines $T(x)$ (up to the constant τ), $\sigma(\Lambda^i)$, and, if k^2 is not a Dirichlet eigenvalue of $-\triangle + V^i$ on Ω, $\Lambda^i(k)$ itself is uniquely determined.*

Proof. First, we show the transfer matrix is uniquely determined. Assume that k^2 is not a Dirichlet eigenvalue of $-\triangle + V^i(x)$ on Ω so that Λ^i exists. This restriction is easily removed, as discussed at the end of the proof of Lemma 5.1.

The proof will rely on an analysis of the symbol of the operator

$$(d - \Lambda^i \circ b)^{-1}(\Lambda^i \circ a - c),$$

where $(d - \Lambda^i \circ b)^{-1}$ refers to the parametrix for $d - \Lambda^i \circ b$. We first need to show that this pseudodifferential operator on $\partial\Omega$ is uniquely determined from $G^e(x,y)$.

We know that

$$\begin{pmatrix} \gamma_-^0 G^i(x,y) \\ \gamma_-^{1,i} G^i(x,y) \end{pmatrix} = \begin{pmatrix} a(x) & b(x) \\ c(x) & d(x) \end{pmatrix} \begin{pmatrix} \gamma_+^0 G^e(x,y) \\ \gamma_+^{1,e} G^e(x,y) \end{pmatrix}. \tag{5.4}$$

Rewriting this by using that $\gamma_-^{1,i} G^i(x,y) = \Lambda^i G^i(x,y)$, we obtain for $y \in \mathbb{R}^n \backslash \overline{\Omega}$ and $x \in \partial\Omega$ that

$$(\Lambda^i \circ a - c) G^e(x,y) = (d - \Lambda^i \circ b)\gamma_+^{1,e} G^e(x,y).$$

Since $(d - \Lambda^i \circ b)$ is elliptic, the previous equality implies

$$\Big[(\Lambda^i \circ b(x) - d(x))^{-1}(\Lambda^i \circ a(x) - c(x))\Big] G^e(x,y) \sim -\gamma_+^{1,e} G^e(x,y). \tag{5.5}$$

Note this holds only in a local neighborhood of any point on the boundary that is not in $\partial\Delta$.

Fix some $x_0 \in \partial\Omega \backslash \partial\Delta$. Since $\gamma_+^{1,e} G^e(x,y)$ is known for all $y \in \mathbb{R}^n \backslash \overline{\Omega}$, so is the limit as $y \downarrow x_0 \in \partial\Omega$. Now, by Corollary 5.2 and Proposition 3.5, $WF(G^e(x,x_0)|_{\partial\Omega}) = x_0 \times S^{n-2}$ since this is true of $\tilde{G}^e(x,x_0)$. Therefore the symbol of the operator acting on $G^e(x,x_0)$ is uniquely determined by Proposition 3.6.

This reduces the problem to showing that

$$\mathcal{B}(x_0,\xi) = \sigma\Big[(\Lambda^i \circ b - d)^{-1}(\Lambda^i \circ a - c)\Big](x_0,\xi) \tag{5.6}$$

uniquely determines $a(x_0), b(x_0), c(x_0)$, and $d(x_0)$ which will be accomplished by analyzing the symbol.

Recall that the symbol of the Dirichlet-to-Neumann operator on $\partial\Omega$ has the form (in local semigeodesic coordinates around x_0)

$$\sigma(\Lambda^i)(x',\xi') = \ell_1(x',\xi') + \ell_0(x',\xi') + \ell_{-1}(x',\xi') + \cdots, \tag{5.7}$$

where $\ell_1 = (\sum_{j,k}^{n-1} \hat{g}_i^{jk}(x',0)\xi_j\xi_k)^{\frac{1}{2}}$ and ℓ_0 does not depend on $V(x)$ (see [22] or section 6). It will considerably simplify the notation in the future to define the product,

$$\Pi_j\big(p,q\big)(x,\xi) = \sum_{|\alpha|=j} \partial_\xi^\alpha p(x,\xi) D_x^\alpha q(x,\xi)/\alpha!. \tag{5.8}$$

Also, in analyzing the symbol, we separate the proof into the cases where $b = 0$ and $b \neq 0$. Note we can immediately tell whether $b = 0$ in some neighborhood of x_0 or whether $b \neq 0$ there by the asymptotics of $G^e(x,y)$ given in Corollary 5.2. Finally, it is worth mentioning that if we consider only the self-adjoint case, the analysis of the symbol can be simplified.

First, consider the simpler case in which $b(x) = 0$ for all x in some neighborhood of x_0. Then at x_0 and in local semigeodesic coordinates around $x_0 \in \partial\Omega$, (5.6) looks like

$$\begin{aligned} \mathcal{B}(x_0,\xi') &= \sigma\Big[\frac{1}{d}\circ\Lambda^i\circ a - \frac{c}{d}\Big](x_0,\xi') \\ &= \frac{a}{d}\ell_1 + \frac{1}{d}\Pi_1(\ell_1,a) + \frac{a}{d}\ell_0 + \frac{c}{d} + \cdots. \end{aligned} \tag{5.9}$$

Thus, at any x_0, we uniquely determine $\frac{a}{d}$, $\frac{1}{d}\Pi_1(\ell_1,a)$, and $\frac{c}{d}$. We claim this uniquely determines the transfer matrix (up to a constant). Say we had two transfer matrices with elements a_1, c_1, d_1 and a_2, c_2, d_2. Then

$$\frac{a_1}{d_1} = \frac{a_2}{d_2}, \tag{5.10}$$

$$\Pi_1(\ell_1,a_1) = \frac{d_1}{d_2}\Pi_1(\ell_2,a_2), \tag{5.11}$$

$$\frac{c_1}{d_1} = \frac{c_2}{d_2}. \tag{5.12}$$

Thus by (5.10),

$$\Pi_1(\ell_1,a_1) = \Pi_1\left(\ell_1,\frac{d_1}{d_2}a_2\right) = \frac{d_1}{d_2}\Pi_1(\ell_1,a_2) + a_2\Pi_1\left(\ell_1,\frac{d_1}{d_2}\right)$$

which combined with (5.11) implies that $\frac{d_1}{d_2} = \tau > 0$ is a constant everywhere. Therefore, (5.10) and (5.12) imply that $a_1 = \tau a_2$ and $c_1 = \tau c_2$.

A much more detailed analysis is needed when $b(x) \neq 0$. Let $t = \det T$. We claim, in local semigeodesic coordinates around any $x_0 \in \partial\Omega$ (and in a sufficiently small neighborhood such that $b \neq 0$),

$$\begin{aligned} \mathcal{B}(x',\xi') = \frac{a}{b} &+ \frac{1}{\ell_1}\Big[\frac{t}{b^2}\Big] \\ &+ \frac{1}{\ell_1^2}\Big[\frac{t}{b^2}\left(\frac{d}{b} - \ell_0 - \frac{1}{\ell_1}\Pi_1(\ell_1,\ell_1)\right) + \frac{1}{b}\Pi_1\left(\ell_1,\frac{t}{b}\right)\Big] + \cdots. \end{aligned} \tag{5.13}$$

Before obtaining this formula, we show this finishes the proof that the transfer matrix is uniquely determined.

Given two transfer matrices, we know

$$\frac{a_1}{b_1} = \frac{a_2}{b_2}, \tag{5.14}$$

$$\frac{t_1}{b_1^2} = \frac{t_2}{b_2^2}, \tag{5.15}$$

$$\frac{d_1}{b_1} = \frac{d_2}{b_2}, \tag{5.16}$$

$$\Pi_1\left(\ell_1, \frac{t_1}{b_1}\right) = \frac{b_1}{b_2}\Pi_1\left(\ell_1, \frac{t_2}{b_2}\right). \tag{5.17}$$

Now, (5.15) implies

$$\Pi_1\left(\ell_1, \frac{t_1}{b_1}\right) = \Pi_1\left(\ell_1, \frac{b_1}{b_2}\frac{t_2}{b_2}\right) = \frac{b_1}{b_2}\Pi_1\left(\ell_1, \frac{t_2}{b_2}\right) + \frac{t_2}{b_2}\Pi_1\left(\ell_1, \frac{b_1}{b_2}\right)$$

which in view of (5.17) implies that there exists a constant $\tau > 0$ such that $\frac{b_1}{b_2} = \tau$. Therefore (5.14), (5.15), and (5.16) finish the claim. That this is the same constant as it is on the set where $b = 0$ follows by the continuity and positivity of a and d on all of $\partial\Omega$.

To prove the symbol has the form (5.13) we will solve for $\mathcal{B}(x', \xi') = \sum_{j=0}^{\infty} r_{-j}(x', \xi')$ using the composition formula for pseudodifferential operators. By (5.6),

$$\sigma(\Lambda^i \circ b - d) \circ \sigma(\mathcal{B}) = \sigma(\Lambda^i \circ a - c).$$

Plugging in the symbol of the Dirichlet-to-Neumann operator (5.7) we obtain (up to terms of order -2)

$$\begin{aligned}\big[\ell_1 b + \Pi_1(\ell_1, b) + \ell_0 b - d + \Pi_2(\ell_1, b) + \Pi_1(\ell_0, b) + \ell_{-1} b\big] \circ \big[r_0 + r_{-1} + r_{-2}\big] \\ = \ell_1 a + \Pi_1(\ell_1, a) + \ell_0 a - c + \Pi_2(\ell_1, a) + \Pi_1(\ell_0, a) + \ell_{-1} a + \cdots .\end{aligned}$$

Equating terms by order of homogeneity, this becomes

$$(\ell_1 b) r_0 = \ell_1 a, \tag{5.18}$$

$$(\ell_1 b) r_{-1} + \big[\Pi_1(\ell_1, b) + \ell_0 b - d\big] r_0 + \Pi_1(\ell_1 b, r_0) = \Pi_1(\ell_1, a) + \ell_0 a - c, \tag{5.19}$$

$$\begin{aligned}&(\ell_1 b) r_{-2} + \big[\Pi_1(\ell_1, b) + \ell_0 b - d\big] r_{-1} + \big[\Pi_2(\ell_1, b) + \Pi_1(\ell_0, b) + \ell_{-1} b\big] r_0 \\ &\quad + \Pi_2(\ell_1 b, r_0) + \Pi_1(\ell_1 b, r_{-1}) + \Pi_1([\Pi_1(\ell_1, b) + \ell_0 b - d], r_0) \\ &\qquad = \Pi_2(\ell_1, a) + \Pi_1(\ell_0, a) + \ell_{-1} a.\end{aligned} \tag{5.20}$$

From (5.18) we see that $r_0 = \frac{a}{b}$. Thus, (5.19) can be rewritten

$$(\ell_1 b) r_{-1} + \Pi(\ell_1, b)\frac{a}{b} + \ell_0 a - \frac{ad}{b} + \Pi_1\left(\ell_1 b, \frac{a}{b}\right) = \Pi_1(\ell_1, a) + \ell_0 a - c.$$

A simple calculation shows that

$$\Pi_1\left(\ell_1 b, \frac{a}{b}\right) = b\Pi_1\left(\ell_1, \frac{a}{b}\right) = \Pi_1(\ell_1, a) - \frac{a}{b}\Pi_1(\ell_1, b),$$

so that (5.19) can be simplified to,

$$(\ell_1 b) r_{-1} - \frac{ad}{b} = -c.$$

Therefore

$$r_{-1} = \frac{1}{\ell_1}\left(\frac{ad}{b^2} - \frac{c}{b}\right) = \frac{1}{\ell_1}\left(\frac{ad-bc}{b^2}\right).$$

Plugging in the formulas for r_0 and r_{-1} and expanding the term $\Pi_1(\ell_0 b, \frac{a}{b})$ as above, (5.20) becomes

$$\begin{aligned}(\ell_1 b)r_{-2} &+ \frac{a}{b}\Pi_2\left(\ell_1, b\right) + \Pi_2\left(\ell_1 b, \frac{a}{b}\right) + \Pi_1\left(\Pi_1(\ell_1, b), \frac{a}{b}\right) \\ &+ \frac{\det T}{\ell_1 b^2}\left[\Pi_1(\ell_1, b) + \ell_0 b - d\right] + \Pi_1\left(\ell_1 b, \frac{\det T}{\ell_1 b^2}\right) = \Pi_2(\ell_1, a).\end{aligned}$$

A trivial, but lengthy, calculation shows that

$$\Pi_2\left(\ell_1 b, \frac{a}{b}\right) = \Pi_2(\ell_1, a) - \frac{a}{b}\Pi_2(\ell_1, b) - \Pi_1\left(\Pi_1(\ell_1, b), \frac{a}{b}\right).$$

Plugging this into the previous equation, expanding $\Pi_1(\ell_1 b, \frac{\det T}{\ell_1 b^2})$ as previously, and solving for r_{-2} we finish the proof of (5.13).

Finally, we show that for k^2 not a Dirichlet eigenvalue of Ω, the Dirichlet-to-Neumann operator Λ^i on $\partial\Omega$ is uniquely determined. By (5.4), we know

$$\Lambda^i\Big(a(x)G^e(x,y) + b(x)\gamma_+^{1,e}G^e(x,y)\Big) = c(x)G^e(x,y) + d(x)\gamma_+^{1,e}G^e(x,y)$$

for $y \in \mathbb{R}^n\backslash\Omega$. Therefore it suffices to note that $a(x)G^e(x,y) + b(x)\gamma_+^{1,e}G^e(x,y)$ for $y \in \mathbb{R}^n\backslash\Omega$ is dense in $L^2(\partial\Omega)$, which is proved in the following lemma. ∎

LEMMA 5.5. *If k^2 is not a Dirichlet eigenvalue of $-\triangle + V^i(x)$ on Ω, then for any open set $\mathcal{O} \subset \mathbb{R}^n\backslash\Omega$,*

$$\{a(x)G^e(x,y) + b(x)\gamma_+^{1,e}G^e(x,y) : y \in \mathcal{O}\}$$

is dense in $L^2(\partial\Omega)$.

Proof. Say

$$u(y) = \int_{\partial\Omega} a(x)G^e(x,y) + b(x)\gamma_+^{1,e}G^e(x,y)\phi(x)\, dS_x = 0 \text{ for all } y \in \mathcal{O}.$$

Then, using the boundary conditions, it follows that

$$u(y) = \int_{\partial\Omega} G^i(x,y)\phi(x)\, dS_x = 0 \text{ for all } y \in \mathcal{O}$$

and so by unique continuation, $u = 0$ in $\mathbb{R}^n\backslash\Omega$. Therefore $u_+(y) = 0$ on $\partial\Omega$ and $\gamma_+^{1,i}u(y) = 0$ on $\partial\Omega$. Since k^2 is not a Dirichlet eigenvalue of $-\triangle + V^i(y) - k^2$ on Ω and, by the jump relations, $u_-(y) = 0$, it follows that $\gamma_-^{1,i}u = 0$. Another application of the jump relations then shows $\phi(y) = 0$. ∎

Theorems 5.3 and 5.4, along with Theorem 4.8, finishes the proof of Theorems 2.2 and 2.3 when there are no magnetic potentials and the media is Euclidean.

6. The symbol of the Dirichlet-to-Neumann operator and boundary determination. In this section we recall the symbol of the Dirichlet-to-Neumann operator and prove that it uniquely determines g, A, and V at the boundary of the obstacle modulo gauge transformations and diffeomorphisms of the metric that do not affect the operator at the boundary. Note these results hold for complex-valued electromagnetic potentials as well.

Let L be a Schrödinger operator of the form

$$Lu(x) = \left(\sum_{j,k} \frac{1}{\sqrt{g}}\big(D_j + A_j(x)\big)\sqrt{g}g^{jk}(x)\big(D_k + A_k(x)\big)\right) u(x) + q(x)u(x)$$

with $(A_1(x), \ldots, A_n)$ and $q(x)$ infinitely smooth, and fix some compact, connected subset $\Omega \subset \mathbb{R}^n$. Consider the boundary value problem for $f \in H^{\frac{1}{2}}(\partial\Omega)$,

$$Lu = 0 \text{ in } \Omega,$$
$$u|_{\partial\Omega} = f.$$

Then the Dirichlet-to-Neumann map, $\Lambda : H^{\frac{1}{2}}(\partial\Omega) \to H^{-\frac{1}{2}}(\partial\Omega)$, is defined by

$$\Lambda f = \sum_{j,k} g^{jk}\left(\frac{\partial u}{\partial x_j} + iA_j u\right)\nu_k(x)\left(\sum_{p,r} g^{pr}(x)\nu_p(x)\nu_r(x)\right)^{-\frac{1}{2}}, \tag{6.1}$$

where ν is the outward pointing normal in Euclidean coordinates.

Before introducing the symbol, it will help to simplify the operator. In a neighborhood of an arbitrary point $x_0 \in \partial\Omega$ introduce semigeodesic coordinates (which we shall continue to denote by $x = (x', x_n)$) so that $x_n = 0$ is the equation of $\partial\Omega$. In these coordinates,

$$\begin{aligned}\hat{L}(x, D)\hat{u} &= D_n^2\hat{u}(x) + L_n(x, D)\hat{u}(x)\\ &\quad + \sum_{j,k}^{n-1} \frac{1}{\sqrt{\hat{g}}}(D_j + \hat{A}_j(x))\sqrt{\hat{g}}\hat{g}^{jk}(x)(D_k + \hat{A}_k(x))\hat{u}(x) + \hat{q}(x)\hat{u}(x),\end{aligned}$$

where

$$L_n(x, D) = (E_n(x) + \hat{A}_n(x))D_n + \hat{A}_n D_n + (E_n + \hat{A}_n)\hat{A}_n + D_n\hat{A}_n).$$

Above, as well as in what follows, $E_j(x) = \frac{D_j(\sqrt{g})}{\sqrt{g}} = \frac{D_j(\sqrt{\hat{g}})}{\sqrt{\hat{g}}}$ $(1 \le j \le n)$. We now select a gauge $\kappa(x)$ such that $\kappa(x', 0) = 1$ and

$$A_n'(x) = \hat{A}_n(x) + \frac{D_n\kappa(x)}{\kappa(x)} = 0.$$

Note κ is easily found by assuming $\kappa(x) = e^{-i\psi(x)}$ for some $\psi(x)$ equal to zero on the boundary. We then have

$$L' = D_n^2 + E_nD_n + \sum_{j,k}^{n-1} \frac{1}{\sqrt{\hat{g}}}(D_j + \hat{A}_j)\sqrt{\hat{g}}\hat{g}^{jk}(D_k + \hat{A}_k)\hat{u} + \hat{q},$$

which we will rewrite in the form

$$\begin{aligned}L'(x, D) &= D_n^2 + E_n(x)D_n + q_2(x, D') + q_1(x, D')\\ &\quad + 2\sum_{j,k}^{n-1} A_j'(x)\hat{g}^{jk}(x)D_k + G(x),\end{aligned}$$

where

$$q_2(x, D') = \sum_{j,k}^{n-1} \hat{g}^{jk}(x) D_j D_k,$$

$$q_1(x, D') = \sum_{j,k}^{n-1} (\hat{g}^{jk}(x) E_j(x) + D_j \hat{g}^{jk}(x)) D_k,$$

and

$$G(x) = \sum_{j,k}^{n-1} (E_j(x) + A'_j + D_j)(\hat{g}^{jk} A_k) + \hat{q}(x).$$

If we define Λ' to be the Dirichlet-to-Neumann operator corresponding to the operator L', then $\Lambda' f = \frac{\partial \hat{u}(x',0)}{\partial x_n}$ and $\Lambda' f = \Lambda f$.

We want to show the symbol of the Dirichlet-to-Neumann operator uniquely determines the value of $\hat{g}$, A', and $\hat{q}$, as well as all their normal derivatives, on the boundary. Here $\hat{g}|_{\partial\Omega}$ is independent of a diffeomorphism of the metric and the $A'_j(x)\,(1 \le j \le n-1)$ are the result of specifying a particular gauge (i.e., one that makes $\hat{A_n}(x) = 0$). Therefore the metric and potentials, along with all powers of their normal derivatives, are uniquely defined at the boundary. Boundary determination from the symbol of the Dirichlet-to-Neumann operator has been proven by Kohn and Vogelius [22] for the case $g_{jk} = \delta_{jk}$ and $A = 0$; by Lee and Uhlmann [23] for arbitrary g and $A = 0, V = 0$; and by Nakamura, Sun, and Uhlmann [25] and Eskin [12] for $g_{jk} = \delta_{jk}$ and real-valued A, V. We slightly improve these results to show the uniqueness for arbitrary $\{g, A, V\}$ by investigating not only terms of a specific order of homogeneity in the symbol but also terms of a specific type of homogeneity. To explain what is meant by type, note both ξ_j and $(\xi_j \xi_l)((\sum_{p,r} g^{pr} \xi_p \xi_r)^{\frac{1}{2}})^{-\frac{1}{2}}$ are of order 1 but are of different type.

The following theorem gives the exact symbol of the Dirichlet-to-Neumann operator. For the proof see [12], [22], [23], or [25].

THEOREM 6.1. *The symbol of the Dirichlet-to-Neumann operator is*

$$\sigma(\Lambda')(x', \xi') = \ell_1(x', \xi') + \ell_0(x', \xi') + \ell_{-1}(x', \xi') + \ell_{-2}(x', \xi') + \cdots,$$

where

$$\ell_1(x', \xi') = \sqrt{q_2(x, \xi)} = \sqrt{\sum_{j,k}^{n-1} \hat{g}^{jk}(x', 0) \xi_j \xi_k},$$

$$\ell_0(x', \xi') = -\frac{1}{2\ell_1} \left(\Pi_1(\ell_1, \ell_1) - q_1 - \partial_{x_n} \ell_1 + E_n \ell_1 + \sum_{j,k}^{n-1} A'_j \hat{g}^{jk}(x', 0) \xi_k \right),$$

$$\ell_{-1}(x', \xi') = -\frac{1}{2\ell_1} \left(\sum_{\substack{0 \le p,r, \le 1 \\ K = p+r}} \Pi_K(\ell_p, \ell_r) + \partial_{x_n} \ell_0 - E_n \ell_0 - G \right),$$

and for $m<0$,

$$\ell_{m-1}(x',\xi') = -\frac{1}{2\ell_1}\left(\sum_{\substack{m\le p,r\le 1\\ K=p+r-m}} \Pi_K(\ell_p,\ell_r) + \partial_{x_n}\ell_m - E_n\ell_m\right).$$

The expansion gives us an immediate corollary on boundary determination.

Corollary 6.2. $\sigma(\Lambda')$ *uniquely determines* (*for all* $p>0$ *and* $1\le j\le n-1$) $(\frac{\partial}{\partial x_n})^p\hat{g}|_{\{x_n=0\}}$, $(\frac{\partial}{\partial x_n})^p A'_j|_{\{x_n=0\}}$, $(\frac{\partial}{\partial x_n})^p\hat{q}(x)|_{\{x_n=0\}}$.

Proof. In ℓ_1^2, $\hat{g}^{\alpha\beta}$ is the coefficient of $\xi_\alpha\xi_\beta$ so that $\hat{g}^{\alpha\beta}$, as well as all their derivatives in directions tangent to the boundary, are uniquely determined.

Define $\tilde{A} = (\{\hat{g}\}_{\alpha,\beta=1}^{n-1})^{-1}A'$. The only new information in ℓ_0 is the appearance of $\frac{\partial\hat{g}^{\alpha\beta}}{\partial x_n}$ for $1\le\alpha,\beta\le n-1$ and $\tilde{A}_j$ for $1\le j\le n-1$. These occur in the following manner:

$$\frac{\partial\hat{g}^{\alpha\beta}(x)}{\partial x_n} \text{ is part of the coefficient of } \frac{\xi_\alpha\xi_\beta}{\sum\hat{g}^{pr}\xi_p\xi_r},$$
$$\tilde{A}_j(x) \text{ is part of the coefficient of } \frac{\xi_j}{(\sum\hat{g}^{pr}\xi_p\xi_r)^{\frac{1}{2}}}.$$

Note the other terms in the expansion are known from ℓ_0. Thus, from ℓ_0 we obtain $\frac{\partial\hat{g}^{\alpha\beta}}{\partial x_n}$ for $1\le\alpha,\beta\le n-1$, and $\tilde{A}_j$. It is now easy to see that new information occurs in a manner which uniquely determines the desired values. Specifically, on the term ℓ_{1-k},

$$\frac{\partial^k\hat{g}^{\alpha\beta}(x',0)}{\partial(x_n)^k} \text{ is part of the coefficient of } \frac{\xi_\alpha\xi_\beta}{(\sum\hat{g}^{pr}\xi_p\xi_r)^{\frac{k+1}{2}}},$$
$$\frac{\partial^{k-1}\tilde{A}_j(x',0)}{\partial(x_n)^{k-1}} \text{ is part of the coefficient of } \frac{\xi_j}{(\sum\hat{g}^{pr}\xi_p\xi_r)^{\frac{k}{2}}},$$
$$\frac{\partial^{k-2}\hat{q}(x',0)}{\partial(x_n)^{k-2}} \text{ is part of the coefficient of } \frac{1}{(\sum\hat{g}^{pr}\xi_p\xi_r)^{\frac{k-1}{2}}}.$$

This implies the claim (note we can recover A' from $\tilde{A}$). ∎

7. Inverse problem for anisotropic media and electromagnetic potentials. Our arguments will closely follow those in section 5. In order to avoid excessive indices, we denote $p_j = \ell_1^{(j)}$ below.

Lemma 7.1. *Say* $p_1^{-2}\Pi_1(p_1,p_1) = p_2^{-2}\Pi_1(p_2,p_2)$ *and* $p_1 = fp_2$. *Then* f *is constant.*

Proof. Substituting the latter into the former, we obtain,

$$\Pi_1(p_1,p_1) = f^2\Pi_1(p_2,p_2). \tag{7.1}$$

On the other hand,

$$\Pi_1(p_1,p_1) = \Pi_1(fp_2,fp_2) = f^2\Pi_1(p_2,p_2) + fp_2\Pi_1(p_2,f). \tag{7.2}$$

Therefore (7.1) and (7.2) imply that $\Pi_1(p_2,f)=0$, from which it follows that f must be constant. ∎

First, let us show the location is uniquely determined. We say an obstacle is *well defined at* x_0 if conditions (i) and (ii) in Theorem 2.2 do not hold for any pair of $\sigma > 0$ and $\rho > 0$. Analogously to Lemma 5.1, we obtain the following.

LEMMA 7.2. $\{\Omega, T, L^i\}$ *is well defined at* x^* *if and only if*

$$\lim_{y\downarrow x^*\in\partial\Omega}\big(G^e(x,y)-\tilde{G}^e(x,y)\big)=\lim_{y\downarrow x^*\in\partial\Omega}S^e_+\phi^e(x,y)\notin C^\infty(\partial\Omega).$$

Proof. By the same reasoning as in the proof of Lemma 5.1, we know a singularity in $\tilde{G}^e(x,y)-G^e(x,y)$ will form as $y\downarrow x\in\partial\Omega$ unless

$$\sigma\big[\Lambda^i\circ a+\Lambda^i\circ b\circ\Lambda^e\big](x_0,\xi)=\sigma\big[c+d\circ\Lambda^e\big](x_0,\xi).$$

Assume this is true. Then $b=0$ and $\sigma(\Lambda^i\circ a)=\sigma(c+d\circ\Lambda^e)$. Therefore

$$\ell^i_1 a+\Pi_1(\ell^i_1,a)+\ell^i_0 a=d\ell^e_1+d\ell^e_0+c+\cdots$$

and using the expansion of ℓ_0,

$$\begin{aligned}\ell^i_1 a+\Pi_1(\ell^i_1,a)-a\frac{1}{2\ell^i_1}&\left(\Pi_1(\ell^i_1,\ell_1)-q^i_1-\partial_{x_{n_i}}\ell^i_1+E^i_n\ell^i_1+\sum_{j,k}^{n-1}(A^i)'_j\hat{g}^{jk}_i(x',0)\xi_k\right)\\ =d\ell^e_1-d\frac{1}{2\ell^e_1}&\left(\Pi_1(\ell^e_1,\ell^e_1)-q^e_1-\partial_{x_{n_e}}\ell^e_1+E^e_n\ell^e_1+\sum_{j,k}^{n-1}(A^e)'_j\hat{g}^{jk}_e(x',0)\xi_k\right)+c.\end{aligned}$$

Immediately we see that $\ell^e_1 a=\ell^i_1 d$ and $\frac{a}{\ell^i_1}\Pi_1(\ell^i_1,\ell^i_1)=\frac{d}{\ell^e_1}\Pi_1(\ell^e_1,\ell^e_1)$ which implies, by Lemma 7.1, there exists a constant ρ such that $\frac{a}{d}=\rho$ and $\frac{\ell^e_1}{\ell^i_1}=\rho$. Since the magnetic potentials are real and the terms $\frac{q^e_1 a}{\ell^e_1}$ and $\frac{q^i_1 d}{\ell^i_1}$ are equal by the previous statement, it follows that $\Pi_1(\ell_1,a)=0$ (since these terms are of a specific homogeneity type). Thus there exists τ such that $a=\tau\rho^{-\frac{n}{2}}$. Therefore $d=\tau\rho^{1-\frac{n}{2}}$ and so $c=0$. We leave it to the reader to verify that in the self-adjoint case, $\tau=1$. $\square$

Note we have proved the following corollary which is of interest in its own right.

COROLLARY 7.3. $\sigma(\Lambda_1\circ f)=\sigma(g\circ\Lambda_2)$ *implies that* f/g *is constant.*

The principal symbol of the single layer potentials can now be calculated.

COROLLARY 7.4. *For* $b\neq 0$,

$$\sigma_{-1}((S^e_+)^t)=\frac{1}{\sqrt{\sum_{j,k}^{n-1}\hat{g}^{jk}_e(x',0)\xi_j\xi_k}},$$

and for $b=0$,

$$\sigma_{-1}((S^e_+)^t)=\frac{1}{(\frac{a}{d}\frac{\ell^i_1}{\ell^e_1}+1)\sqrt{\sum_{j,k}^{n-1}\hat{g}^{jk}_e(x',0)\xi_j\xi_k}},$$

where we are in local, semigeodesic coordinates.

Analogously to Theorem 5.4, we have the following.

THEOREM 7.5. *Given* $G_1(x,y)$ *and* $G_2(x,y)$ *such that* $G^e_1(x,y)=G^e_2(x,y)$ *for all* x *and* y *in* $\mathbb{R}^n\backslash\Omega$, *it follows that there exists a constant* $\rho>0$ *such that*

$$T_1(x)=\begin{pmatrix}\rho^{-\frac{n}{2}} & 0\\ 0 & \rho^{1-\frac{n}{2}}\end{pmatrix}T_2(x),\ \sigma(\Lambda^i_1)=\rho\sigma(\Lambda^i_2),$$

and if k^2 is not a Dirichlet eigenvalue for either $L_1^i - k^2$ or $L_2^i - k^2$ on Ω, then $\Lambda_1^i = \rho\Lambda_2^i$.

Proof. As in the proof of Theorem 5.4, we know that $G_j^e(x,y)$ $(j=1,2)$ uniquely determines the operator

$$\mathcal{B}_j(x,\xi) = \sigma\big[(\Lambda^i \circ b - d)^{-1}(\Lambda^i \circ a - c)\big](x_0,\xi).$$

Also, as mentioned in the proof of Theorem 5.4, we can tell whether $b=0$ or $b \neq 0$ in the neighborhood of any $x_0 \notin \partial\Delta$ by the asymptotics of $G_j^e(x,y)$ given in the previous corollary. Therefore $b_1(x) = 0$ when $b_2(x) = 0$.

We will consider only the case that $b_j(x) \neq 0$ since the other case follows similarly. Using (x',ξ') to denote we are in local semigeodesic coordinates and restricting to a sufficiently small neighborhood,

$$\begin{aligned}\mathcal{B}_j(x',\xi') = \frac{a_j}{b_j} + \frac{1}{\ell_1^{(j)}}\left[\frac{t_j}{b_j^2}\right] \\ + \frac{1}{(\ell_1^{(j)})^2}\left[\frac{t_j}{b_j^2}\left(\frac{d_j}{b_j} - \ell_0^{(j)} - \frac{1}{\ell_1^{(j)}}\Pi_1(\ell_1^{(j)},\ell_1^{(j)})\right) + \frac{1}{b_j}\Pi_1\left(\ell_1^{(j)},\frac{t_j}{b_j}\right)\right] + \cdots .\end{aligned}$$

Equating terms by order and type of homogeneity in the equality $\mathcal{B}_1(x',\xi') = \mathcal{B}_2(x',\xi')$, we get

$$\frac{a_1}{b_1} = \frac{a_2}{b_2}, \tag{7.3}$$

$$\frac{1}{p_1}\frac{t_1}{b_1^2} = \frac{1}{p_2}\frac{t_2}{b_2^2}, \tag{7.4}$$

$$\frac{1}{p_1^2}\frac{t_1}{b_1^2}\left(\frac{d_1}{b_1} + \frac{D_{n_1}(\sqrt{\hat g_1})}{\sqrt{\hat g_1}} - \frac{\partial_{x_{n_1}}p_1}{p_1}\right) = \frac{1}{p_2^2}\frac{t_2}{b_2^2}\left(\frac{d_2}{b_2} + \frac{D_{n_2}(\sqrt{\hat g_2})}{\sqrt{\hat g_2}} - \frac{\partial_{x_{n_2}}p_2}{p_2}\right), \tag{7.5}$$

$$\frac{1}{p_1^2}\Pi_1(p_1,p_1) = \frac{1}{p_2^2}\Pi_1(p_2,p_2), \tag{7.6}$$

$$\frac{1}{p_1^2 b_1}\Pi_1\left(p_1,\frac{t_1}{b_1}\right) + \frac{t_1}{b_1^2}\frac{1}{p_1^2}\sum A_j^{1'}\hat g_1^{jk}\xi_k = \frac{1}{p_2^2 b_2}\Pi_1\left(p_2,\frac{t_2}{b_2}\right) + \frac{t_2}{b_2^2}\frac{1}{p_2^2}\sum A_j^{2'}\hat g_2^{jk}\xi_k. \tag{7.7}$$

Note in (7.7) we really should have included $\frac{t_1}{b_1^2}\frac{1}{p_1^2}\frac{q_1^{(1)}}{p_1}$ and $\frac{t_2}{b_2^2}\frac{1}{p_2^2}\frac{q_2^{(1)}}{p_2}$. However, we will not need (7.7) until it already will have been proven that these terms are equal. From (7.6) and (7.4), we see that there exists a constant ρ such that

$$\frac{p_1}{p_2} = \frac{t_1}{t_2}\left(\frac{b_2}{b_1}\right)^2 = \rho. \tag{7.8}$$

Now,

$$\begin{aligned}\Pi_1\left(p_1,\frac{t_1}{b_1}\right) &= \Pi_1\left(\rho p_2, \rho\frac{b_1}{b_2}\frac{t_2}{b_2}\right) \\ &= \rho^2\Pi_1\left(p_2,\frac{b_1}{b_2}\frac{t_2}{b_2}\right) \\ &= \rho^2\frac{b_1}{b_2}\Pi_1\left(p_2,\frac{t_2}{b_2}\right) + \rho^2\frac{t_2}{b_2}\Pi_1\left(p_2,\frac{b_1}{b_2}\right).\end{aligned}$$

Plugging this result into (7.7), we obtain

$$\frac{c^2}{p_1^2}\frac{b_2}{b_1}\frac{t_2}{b_2^2}\Pi_1\left(p_2,\frac{b_1}{b_2}\right)+\frac{t_1}{b_1^2}\frac{1}{p_1^2}\sum A_j^{1'}\hat{g}_1^{jk}\xi_k=\frac{t_2}{b_2^2}\frac{1}{p_2^2}\sum A_j^{2'}\hat{g}_2^{jk}\xi_k.$$

Using that the magnetic potentials are real-valued and that $\frac{t_1}{b_1^2}$ and $\frac{t_2}{b_2^2}$ have a real constant proportionality, it follows that $\Pi_1(p_2,\frac{b_1}{b_2})=0$. Thus there exists a constant ρ such that $\frac{b_1}{b_2}=\tau\rho^{-\frac{n}{2}}$ and, hence, $\frac{t_1}{t_2}=\tau^2\rho^{n-1}$. To finish the unique determination, we need to analyze (7.5) in order to show that $\frac{d_1}{b_1}=\rho\frac{d_2}{b_2}$. There are two possibilities for the terms of the form $\frac{\partial_{x_n}p}{p}$: either (1) $\partial_{x_n}\hat{g}^{jk}=\tau\hat{g}^{jk}$ for all j,k and some $\tau>0$, or (2) this is not the case. In case (2), the term has homogeneity type $\frac{\xi_a}{p}$ which is distinguishable from homogeneity type 1. Note that if case (2) holds for $\hat{g}_1$, then it also holds for $\hat{g}_2$ since the symbols are equal. In this case, $\partial_{x_{n_1}}\hat{g}_1^{jk}$ and $\partial_{x_{n_2}}\hat{g}_2^{jk}$ are uniquely determined (and proportional to one another). Therefore so are the terms involving the determinants. Consequently, the above relation between $\frac{d_1}{b_1}$ and $\frac{d_2}{b_2}$ must hold. On the other hand, if case (1) holds for both g_1 and g_2, then a calculation shows that

$$\frac{D_{n_1}(\sqrt{\hat{g}_1})}{\sqrt{\hat{g}_1}}-\frac{\partial_{x_{n_1}}p_1}{p_1}=\frac{n\tau}{i}-\tau$$

and the same relation (times the proper proportionality constant) holds for g_2 so that again we get the desired relation between $\frac{d_1}{b_1}$ and $\frac{d_2}{b_2}$.

The case when $b(x)=0$ follows by analyzing the symbol in the same fashion.

Now assume k^2 is not a Dirichlet eigenvalue on Ω for either L_1 or L_2. By (5.4), we know (for $j=1,2$)

$$\Lambda_j\Big(a_j(x)G_j^e(x,y)+b_j(x)\gamma_+^{1,e}G_j^e(x,y)\Big)=c_j(x)G_j^e(x,y)+d_j(x)\gamma_+^{1,e}G_j^e(x,y)$$

for $y\in\mathbb{R}^n\backslash\Omega$. By Lemma 5.5, the behavior of the Dirichlet-to-Neumann operator on the above functions determines the operator. The above relations on the elements of the transfer matrix therefore imply $\Lambda_1=\rho\Lambda_2$. □

By Lemma 7.2, we can apply Theorem 5.3 to locate the obstacle. Therefore, Lemma 7.2 and Theorem 7.5 prove Theorems 2.2 and 2.3.

Acknowledgments. The author thanks Gregory Eskin for initially suggesting work on transmission problems, as well as for his constant guidance and support. He is also deeply indebted to Michael Hitrik and James Ralston for numerous discussions during the preparation of this paper.

REFERENCES

[1] T. S. Angell, R. E. Kleinman, and F. Hettlich, *The resistive and conductive problems for the exterior Helmholtz equation*, SIAM J. Appl. Math., 50 (1990), pp. 1607–1622.

[2] D. Balian, D. Bessis, and G. A. Mezincescu, *Form of kinetic energy in effective-mass Hamiltonians for heterostructures*, Phys. Rev. B, 51 (1995), pp. 17624–17629.

[3] P. Bovik and P. Olsson, *Effective boundary conditions for the scattering of two-dimensional SH waves from a curved thin elastic layer*, Proc. Roy. Soc. Lond. Ser. A, 439 (1992), pp. 257–269.

[4] D. Cakoni, D. Colton, and H. Haddar, *The linear sampling method for anisotropic media*, J. Comput. Appl. Math., 46 (2001), pp. 285–299.

[5] D. Colton, J. Coyle, and P. Monk, *Recent developments in inverse acoustic scattering theory*, SIAM Rev., 42 (2000), pp. 369–414.
[6] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley & Sons, New York, NY, 1983.
[7] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer, Berlin, 1998.
[8] D. Colton and R. Kress, *Using fundamental solutions in inverse scattering*, Inverse Problems, 22 (2006), pp. R49–R66.
[9] D. Colton, R. Kress, and P. Monk, *Inverse scattering from an orthotropic medium*, J. Comput. Appl. Math., 81 (1997), pp. 269–298.
[10] G. Eskin, *Boundary Value Problems for Elliptic Pseudodifferential Equations*, Transl. Math. Monogr. 52, AMS, Providence, RI, 1981.
[11] G. Eskin, *Introduction to Partial Differential Equations*, UCLA Lecture Notes, University of California, Los Angeles, 2000.
[12] G. Eskin, *Inverse boundary value problems and the Aharonov-Bohm effect*, Inverse Problems, 19 (2003), pp. 49–62.
[13] S. Flügge, *Practical Quantum Mechanics* I, Springer-Verlag, Berlin, 1971.
[14] T. Gerlach and R. Kress, *Uniqueness in inverse obstacle scattering with conductive boundary condition*, Inverse Problems, 12 (1996), pp. 619–625.
[15] P. Hähner, *On uniqueness of the shape of a penetrable, anisotropic obstacle*, J. Comput. Appl. Math., 116 (2000), pp. 167–180.
[16] F. Hettlich, *On the uniqueness of the inverse conductive scattering problem for the Helmholtz equation*, Inverse Problems, 10 (1994), pp. 129–144.
[17] L. Hörmander, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1963.
[18] L. Hörmander, *The Analysis of Linear Partial Differential Operators* I, Springer, Berlin, 1990.
[19] L. Hörmander, *The Analysis of Linear Partial Differential Operators* III, Springer, Berlin, 1985.
[20] V. Isakov, *Uniqueness in the inverse transmission problem*, Comm. Partial Differential Equations, 15 (1990) pp. 1565–1587.
[21] A. Kirsch and L. Päivärinta, *On recovering obstacles inside inhomogeneities*, Math. Methods Appl. Sci., 21 (1998), pp. 619–51.
[22] R. Kohn and M. Vogelius *Determining conductivity by boundary measurements*, Comm. Pure Appl. Math., 38 (1985), pp. 643–667.
[23] J. Lee and G. Uhlmann, *Determining anisotropic real-analytic conductivities by boundary measurements*, Comm. Pure Appl. Math., 42 (1985), pp. 1097–1112.
[24] P. A. Martin, *Acoustic scattering by inhomogeneous obstacles*, SIAM J. Appl. Math., 64 (2003), pp. 297–308.
[25] G. Nakamura, Z. Sun, and G. Uhlmann, *Global identifiability for an inverse problem for the Schrödinger equation in a magnetic field*, Math. Ann., 303 (1995), pp. 377–388.
[26] S. O'Dell, *Inverse scattering for the Laplace–Beltrami operator with complex electromagnetic potentials and embedded obstacles*, Inverse Problems, 22 (2006), pp. 1579–1603.
[27] M. Piana, *On uniqueness for anisotropic inhomogeneous inverse scattering problems*, Inverse Problems, 14 (1998), pp. 1565–1579.
[28] R. Potthast, *Electromagnetic scattering from an orthotropic medium*, J. Integral Equations Appl., 11 (1999), pp. 197–215.
[29] R. Potthast and I. Stratis, *The singular sources method for an inverse transmission problem*, Computing, 75 (2005), pp. 237–255.
[30] M. Shubin, *Pseudodifferential Operators and Spectral Theory*, Springer-Verlag, Berlin, 1987.
[31] M. Taylor, *Partial Differential Equations* II, Appl. Math. Sci. 116, Springer-Verlag, New York, 1996.
[32] I. V. Tokatly, A. G. Tsibizov, and A. A. Gorbatsevich, *Interface electronic states and boundary conditions for envelope functions*, Phys. Rev. B, 65 (2002), pp. 165328. Available online at http://prola.aps.org/pdf/PRB/v65/i16/e165328.
[33] N. Valdivia, *Uniqueness in inverse obstacle scattering with conductive boundary conditions*, Appl. Anal., 83 (2004), pp. 825–861.
[34] K. Wapenaar, E. Slob, and J. Fokkema, *Reciprocity and power balance for piecewise continuous media with imperfect interfaces*, J. Geophys. Res., 109 (2004), paper B10301.
[35] P. Werner, *Zur mathematischen Theorie akustischer Wellenfelder*, Arch. Ration Mech. Anal., 6 (1960), pp. 231–260.

SIAM J. MATH. ANAL.
Vol. 39, No. 5, pp. 1627–1667

EXPLICIT SOLUTIONS FOR A CLASS OF NONLINEAR PDEs THAT ARISE IN ALLOCATION PROBLEMS*

PAUL DUPUIS† AND JIM X. ZHANG‡

Abstract. To exploit large deviation approximations for allocation and occupancy problems one must solve a deterministic optimal control problem (or equivalently, a calculus of variations problem). As this paper demonstrates, and in sharp contrast to the great majority of large deviation problems for processes with state dependence, for allocation problems one can construct more or less explicit solutions. Two classes of allocation problems are studied. The first class considers objects of a single type with a parameterized family of placement probabilities. The second class considers only equally likely placement probabilities but allows for more than one type of object. In both cases, we identify the Hamilton–Jacobi–Bellman equation, whose solution characterizes the minimal cost, explicitly construct solutions, and identify the minimizing trajectories. The explicit construction is possible because of the very tractable properties of the relative entropy function with respect to optimization.

Key words. explicit solutions, allocation, occupancy problems, nonlinear PDE, large deviations, calculus of variations

AMS subject classifications. 35C99, 34H05, 65K10, 60F10, 60K30

DOI. 10.1137/060659739

1. Introduction.

Allocation and occupancy problems are concerned with the random placement of objects into containers. The objects (usually referred to as balls or tokens) can be of a single type or many, in which case they are often distinguished by "color." The containers are variously called urns or cells and have many interpretations such as physical partitions (photo-electric receptors in a grid) and temporal partitions (the days of the year).

There are also many rules for how a given ball may be assigned to a given cell. The simplest such rule, in the context of a single color, uses what are called Maxwell–Boltzmann (MB) statistics. Here, each cell is equally likely to receive each ball. Other rules consider the balls as being placed sequentially, and the likelihood that a given ball is placed in a given cell depends on the current contents of that cell (relative to the contents of all other cells). Examples in this category are Bose–Einstein (BE) statistics, for which a cell that already contains balls is more likely to receive the next ball, and Fermi–Dirac (FD) statistics, where the reverse holds. The precise definitions of BE and FD will be given below.

A key random variable associated with an allocation is the *empirical measure*. After all (or some) of the balls have been placed, one can form the (random) probability measure $(\eta_0, \eta_1, \ldots)$ on $\{0, 1, \ldots\}$, with η_0 equal to the fraction of cells that are empty, η_1 the fraction that contain 1, etc. For example, one could be particularly concerned that at least 90% of the cells are nonempty after the random allocation. In

*Received by the editors May 12, 2006; accepted for publication (in revised form) October 8, 2007; published electronically January 30, 2008.
http://www.siam.org/journals/sima/39-5/65973.html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (dupuis@dam.brown.edu). This author's research was supported in part by the National Science Foundation (NSF-DMS-0306070 and NSF-DMS-0404806) and the Army Research Office (DAAD19-02-1-0425 and W911NF-05-1-0286).

‡CRSC, North Carolina State University, Raleigh, NC 27695 (jimxzhang@ncsu.edu). This author's research was supported in part by the National Science Foundation (NSF-DMS-0306070).

this case one is interested in the distribution of the first component of the empirical measure, and in particular $P\{\eta_0 \le 0.1\}$.

While methods from combinatorial probability provide exact formulas for certain classes of allocation problems, they do not apply universally, nor are they always of great practical utility; see the discussion in [2] on this point. Hence one turns to approximations. The simplest approximation is a law of large numbers (LLN) limit, under which the number of cells and number of balls placed into the cells both tend to ∞ with some fixed ratio. If η is indexed by the number of cells n, then the LLN limit identifies the (deterministic) probability distribution that η^n tends to as $n \to \infty$. This identifies the "typical" behavior of the allocation scheme for large n. The limit can often be identified as the solution to a system of ordinary differential equations (ODEs) at time t (and for an appropriate initial condition), where t is limiting ratio of the number of balls to the number of cells, i.e., the mean number of balls per cell.

If in contrast one is concerned with probabilities of atypical behavior, then one considers large deviation asymptotics. For example, if it is usual that 50% of the cells are empty when n is large, then under some technical assumptions large deviation asymptotics assert that $-\frac{1}{n}\log P\{\eta_0 \le 0.1\}$ tends to some constant $c > 0$, thus identifying the exponential rate of decay of the probability. The parameter c is usually identified as the solution of a calculus of variations problem, and using the well-known relation between problems in calculus of variations and Hamilton–Jacobi equations, c can also be characterized as the value (at a particular point) of the solution to a nonlinear partial differential equation (PDE).

The explicit identification of c is in general a daunting task. Whilst there are a small number of cases for which analytic expressions are available, in most cases one must attempt numerical approximation, and so one is limited to only low dimensional problems (i.e., in our setting to the first few components of the empirical distribution). Even putting aside the restriction of numerical methods to low dimensions, one would prefer analytic expressions for c since they have many other uses. Beyond simply identifying the rate of decay, analytic expressions for c can be used

- to characterize the most likely way that a rare event will occur,
- to construct efficient Monte Carlo schemes (known as importance sampling schemes) for nonasymptotic approximations, and
- in statistical estimation and model inference for occupancy models.

The purpose of the present paper is to show that explicit solutions can be obtained for the PDEs that are associated with a wide variety of allocation problems and to introduce techniques that can be applied to even broader classes of problems. As remarked previously, explicit solutions are not common. Among the classes of nonlinear, first order PDEs with explicit solutions (in general dimension) are those associated with the linear quadratic regulator and those linked to the Hopf–Lax formula. Both these examples exploit some significant underlying simplification. In the first example it is the fact that the value function for the control problem is expected to be quadratic in the spatial variable, and in the second example it is the independence of the running cost from the state variable. The optimization problems related to allocation problems are qualitatively quite different from either of these, as can be seen from both the form of the value functions and the structure of the minimizing trajectories. There is significant state dependence and no a priori obvious form for the value function. In the setting of allocation problems, it seems that the attractive properties of the *relative entropy* function are largely responsible for the existence of explicit solutions. It is these properties which allow for convenient calculation and representation of the various derivatives in terms of Lagrange multipliers, the key ingredient in the proof.

In the next section we analyze the single color model. After introducing the general model and formally reviewing the large deviation context, we discuss a formal and heuristic derivation of the explicit solution. The associated Hamilton–Jacobi–Bellman (HJB) equation is then introduced, and a solution is proposed in the form of a finite dimensional minimization problem that can be easily and efficiently solved using Lagrange multiplier techniques. The value of the minimization problem is shown to be smooth for an appropriate class of terminal costs, its derivatives are characterized via multipliers, and the HJB equation is shown to hold. The section concludes with the identification of the minimizing trajectories. The third and final section repeats these steps for a model with different colors.

2. Allocation models with differing assignment probabilities.

2.1. Probabilistic background and the variational problem. In this section, we formulate a general single color occupancy problem. After describing the model, we outline the relevant large deviation properties on path space and the related variational problems.

In the occupancy problem considered here cells are distinguished according to the number of balls contained therein. The full collection of models will be indexed by a parameter a. This parameter takes values in the set $(0, \infty] \cup \{-1, -2, \ldots, \}$, and its interpretation is as follows. Suppose that a ball is about to be thrown, and that any two cells (labeled, say, A and B) are selected. A cell is said to be of *category* i if it contains i balls. Suppose that cell A is of category i, while B is of category j. Then the probability that the ball is thrown into cell A, conditioned on the state of all the cells, and that the ball is thrown into either cell A or B, is

$$\frac{a+i}{(a+i)+(a+j)}.$$

When $a = \infty$ we interpret this to mean that the two cells are equally likely. Also, when $a < 0$ we use this ratio to define the probabilities only when $0 \leq i \vee j \leq -a$ and $i < -a$ or $j < -a$, so the formula gives a well-defined probability. The probability that a ball is placed in a cell of category $-a$ is 0. Thus under this model, cells can only be of category $0, 1, \ldots, -a$, and we only throw balls into categories $0, 1, \ldots, -a-1$.

When $a \in (0, \infty)$, cells that already contain balls are more likely to receive the next ball. When $a < 0$ the opposite is true. The cases $a = 1$, $a = \infty$, $a \in -\mathbb{N}$ correspond to what were called Bose–Einstein statistics, Maxwell–Boltzmann statistics, and Fermi–Dirac statistics, respectively, in the introduction.

Suppose that before we throw a ball there are already tn balls in all the cells, and that the occupancy state is $(x_0, x_1, \ldots, x_{I+})$. Here $x_i, i = 0, 1, \ldots, I$, denotes the fraction of cells that contain i balls, and x_{I+} denotes the fraction containing more than I balls. Throughout this paper we use this convention so that the state space of the occupancy process is finite dimensional. (Explicit formulas analogous to the ones derived here also hold in the infinite dimensional case, though one must be more careful in defining the PDE.) When the occupancy state is $(x_0, x_1, \ldots, x_{I+})$, the "un-normalized" or "relative" probability of throwing into a category i cell with $i \leq I$ is simply $(a+i)x_i$. Let us temporarily abuse notation, and let $x_{I+1}, x_{I+2}, \ldots$ denote the exact fraction in each category i with $i > I$. Since there are tn balls in the cells before we throw, $\sum_{i=0}^{\infty} i x_i = t$. Thus the (normalized and true) probability that the ball is placed in a cell that contains exactly i balls, $i = 0, 1, \ldots, I$, is $\frac{a+i}{a+t} x_i$, and the probability that the ball is placed in a cell that has more than I balls is $1 - \sum_{j=0}^{I} \frac{a+j}{a+t} x_j$.

In order to define both the LLN and large deviation approximations, it is convenient to introduce an occupancy *process*. We introduce a time variable t that ranges from 0 to T. At a time t that is of the form l/n, with $0 \leq l \leq \lfloor nT \rfloor$ an integer, l balls have been thrown. Let $X^n(t) = \left\{ X_0^n(t), X_1^n(t), \ldots, X_I^n(t), X_{I+}^n(t) \right\}$ be the *occupancy state* at that time. As noted previously, $X_i^n(t)$ denotes the fraction of cells that contain i balls at time t, $i = 0, 1, \ldots, I$, and $X_{I+}^n(t)$ the fraction of cells that contain more than I balls. The definition of X^n is extended to all $t \in [0, T]$ not of the form l/n by piecewise linear interpolation. Note that $X^n(t)$ is indeed a probability vector in $\mathbb{R}^{I+2}$. If

$$\mathcal{S}_I \doteq \left\{ x \in \mathbb{R}^{I+2} : x_i \geq 0,\ 0 \leq i \leq I+1, \quad \text{and} \quad \sum_{i=0}^{I+1} x_i = 1 \right\},$$

then for any $t \in [0, T], X^n(t) \in \mathcal{S}_I$. Thus X^n takes values in $\mathcal{U} \doteq C([0, T], \mathcal{S}_I)$. We equip $\mathcal{U}$ with the usual supremum norm and on $\mathcal{S}_I$ we take the usual L_1 norm. The generator of X^n sampled at times of the form $t = l/n$ is

$$\begin{aligned}
\mathcal{L}^n f(x, t) &= nE\left[f(X^n(t + 1/n)) - f(X^n(t)) | X^n(t) = x \right] \\
&= n \sum_{k=0}^{I} \left(\frac{a+k}{a+t} \right) x_k \left[f(x + (e_{k+1} - e_k)/n) - f(x) \right],
\end{aligned}$$

where e_k are the standard basis unit vectors.

It is often the case that one is interested in the large deviation properties at the terminal time T (i.e., those of $X^n(T)$) and for a general initial condition of the form $X^n(t) = (x_0, \ldots, x_{I+})$. Here there is often a detour—one first identifies the large deviation properties of the *process*, and then solves for the large deviation properties of $X^n(T)$ via the so-called contraction principle. This theorem represents the sought after exponential rate of decay as the solution to a calculus of variations problem, and therein lies the link to a PDE.

For our purposes an informal description of the process level large deviation properties will suffice. We first define the *rate function* on path space. Given (x, t) and a continuous trajectory φ with $\varphi(t) = x$, the rate $\mathcal{I}(\varphi; x, t)$ identifies the decay rate for the probability that X^n is in a small neighborhood of φ:

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \limsup_{n \to \infty} -\frac{1}{n} \log P \left\{ \sup_{t \leq s \leq T} |X^n(s) - \varphi(s)| < \delta \,\middle|\, X^n(t) = x^n \right\} \\
&= \lim_{\delta \downarrow 0} \liminf_{n \to \infty} -\frac{1}{n} \log P \left\{ \sup_{t \leq s \leq T} |X^n(s) - \varphi(s)| < \delta \,\middle|\, X^n(t) = x^n \right\} \\
&= \mathcal{I}(\varphi; x, t).
\end{aligned}$$

Here x^n is any sequence of initial conditions that can occur with positive probability and which satisfy $x^n \to x$ as $n \to \infty$. The proof of such a result and the identification of the rate function are given in [7]. $\mathcal{I}(\varphi; x, t)$ can be represented as the integral, over $[t, T]$, of a nonnegative "cost" which measures the likelihood that the increments of X^n follow the increments of φ, with higher cost corresponding to lower likelihood (the LLN trajectory has zero cost). The integral form of $\mathcal{I}(\varphi; x, t)$ is a consequence of the Markov property.

The specific form of $\mathcal{I}(\varphi; x, t)$ is as follows. Define the linear map $M : \mathcal{S}_I \mapsto \mathbb{R}^{I+2}$

$$M_i[\theta] = \begin{cases} -\theta_0, & i = 0, \\ \theta_{i-1} - \theta_i, & 1 \le i \le I, \\ \theta_I, & i = I+1. \end{cases}$$

Let $\varphi \in \mathcal{U}$ be given with $\varphi(t) = x$. Suppose there is a Borel measurable function $\theta : [t, T] \longmapsto \mathcal{S}_I$ such that for any $s \in [t, T]$,

$$\varphi(s) = \varphi(t) + \int_t^s M[\theta](u) du. \tag{2.1}$$

We interpret $\theta_i(s)$ as the rate at which balls are thrown into cells that contain i balls at time s. This rate will be viewed as a perturbation of the LLN limit rate at which balls would be thrown, and the cost for this perturbation will measure the likelihood that sure a perturbation occurs (with large cost corresponding to unlikely perturbations).

The mapping $M[\theta]$ accounts for the fact that when a ball is placed in a category i cell X_i^n decreases by $1/n$ and X_{i+1}^n increases by $1/n$. The rates $\theta(s)$ are unique in the sense that if another $\tilde{\theta} : [t, T] \longmapsto \mathcal{S}_I$ satisfies (2.1), then $\tilde{\theta} = \theta$ a.e. on $[t, T]$. We call φ a *valid* occupancy state process if there exists $\theta : [t, T] \longmapsto \mathcal{S}_I$ satisfying (2.1). In this case θ is called the occupancy rate process associated with φ. For $x \in \mathbb{R}^{I+2}$ and $t \in [0, -a1_{\{a<0\}} + \infty 1_{\{a>0\}})$, define the vector $\rho(t, x) \in \mathbb{R}^{I+2}$ by

$$\rho_k(t, x) = \frac{a+k}{a+t} x_k \qquad \text{for } k = 0, 1, \ldots, I, \tag{2.2}$$

and

$$\rho_{I+}(t, x) = 1 - \sum_{k=0}^{I} \frac{a+k}{a+t} x_k.$$

For each a, $\rho_k(t, x)$ gives the LLN limiting probability that at time t the next ball will be placed in a category k cell, given that the statistics of model a are used and that $X^n(t) = x$. A direct calculation shows that if

$$x \in \mathcal{S}_I \quad \text{and} \quad \sum_{k=0}^{I+1} k x_k \le t, \tag{2.3}$$

then $\rho(t, x)$ is indeed a probability vector in $\mathbb{R}^{I+2}$, i.e., $\rho(t, x) \in \mathcal{S}_I$. It is easy to observe that if φ is valid, then $\varphi(s)$ satisfies (2.3) for all $s \in [0, T]$. This shows that $\rho(s, \varphi(s)) \in \mathcal{S}_I$. For future use we define

$$\tau(x, t) \doteq \left(t - \sum_{k=0}^{I} k x_k \right) \Big/ x_{I+} \tag{2.4}$$

if $x_{I+} > 0$ and $\tau(x, t) \doteq I + 1$ if $x_{I+} = 0$. Thus $\tau(x, t)$ can be interpreted as the mean number of balls per cell among those of category $I+$. With this notation,

$$\rho_{I+}(x, t) = (a + \tau(x, t)) x_{I+1} / (a + t), \tag{2.5}$$

and so $\rho_{I+}(x, t)$ in some sense takes a form very similar to that of $\rho_k(x, t)$ for $k = 0, 1, \ldots, I$.

Let $\delta > 0$ be small. Observe that the occupancy state will not change very much over $[t, t+\delta]$ while $n\delta$ balls are placed into cells. Let θ denote the empirical measure on the categories where these balls are placed. Then the new occupancy state is the sum of the old state plus $\delta M[\theta]$. Since the change in state is determined by an empirical distribution for (at least approximately) independent and identically distributed (iid) random variables, Sanov's theorem [1, Theorem 2.2.1] suggests that the cost appearing in the integral representation for $\mathcal{I}(\varphi; x, t)$ should be defined in terms of the famous *relative entropy* function. For two probability measures α and β on a Polish space $\mathcal{A}$, the relative entropy of α with respect to β is defined by

$$R(\alpha||\beta) \doteq \int_{\mathcal{A}} \left(\log \frac{d\alpha}{d\beta}\right) d\alpha$$

whenever α is absolutely continuous with respect to β (and with the convention that $0 \log 0 = 0$). In all other cases we set $R(\alpha||\beta) = \infty$. When two probability vectors ρ and $\nu \in \mathcal{S}_I$ appear in the relative entropy function, we interpret them as probability measures on the simplex $\{0, 1, \ldots, I, I+1\}$, and thus

$$R(\rho||\nu) \doteq \sum_{i=0}^{I+1} \rho_i \log \frac{\rho_i}{\nu_i}.$$

Important properties of relative entropy are that it is nonnegative, jointly convex, and lower semicontinuous in (α, β), and $R(\alpha||\beta) = 0$ if and only if $\alpha = \beta$ [1, Lemma 1.4.3].

As observed, when $\varphi(s)$ is *valid,* $\rho(s, \varphi(s)) \in \mathcal{S}_I$, which makes $R(\theta(s)||\rho(s, \varphi(s)))$ well defined. If in addition $\varphi(t) = x$, define

$$\mathcal{I}(\varphi; x, t) = \int_t^T R(\theta(s)||\rho(s, \varphi(s)))ds. \tag{2.6}$$

If φ is not valid or $\varphi(t) \neq x$, then define $\mathcal{I}(\varphi; x, t) = \infty$.

This defines the rate function for the models introduced at the beginning of this section. Now suppose that one wishes to approximate probabilities involving $X^n(T)$. Since the probability that X^n (as a process) is close to a given trajectory φ decays exponentially, decay rates of quantities such as $P\{X^n(T) \in A | X^n(t) = x^n\}$ can (under appropriate regularity conditions on A) be found as follows. Among all trajectories φ with $\varphi(t) = x$ and $\varphi(T) \in A$, identify the one with the *smallest* decay rate c. Then c is also the exponential decay rate of $P\{X^n(T) \in A | X^n(t) = x^n\}$. Hence the variational problem to be solved is

$$V(x, t) = \inf_{\varphi:\varphi(t)=x \textit{ and } \varphi(T)\in A} \mathcal{I}(\varphi; x, t). \tag{2.7}$$

If one is interested in expected values other than probabilities, then variational problems of the more general form

$$V(x, t) = \inf_{\varphi:\varphi(t)=x} [\mathcal{I}(\varphi; x, t) + F(\varphi(T))] \tag{2.8}$$

arise, and one is often particularly interested in the initial condition that corresponds to starting with all cells empty: $t = 0, x_0 = 1$ and $x_k = 0, k > 0$. We will refer to this as the *empty* initial condition.

Not all initial conditions are *feasible* in the sense that they can be reached with finite cost from the empty initial condition. Feasibility in this context depends on the underlying parameter a.

DEFINITION 2.1 (feasible domain). *Define $\mathcal{D}_a$, the feasible domain for the occupancy model with parameter a, as follows:*

- *When $a > 0$,*

$$\mathcal{D}_a \doteq \left\{ (x,t) \in \mathcal{S}_I \times [0,T) : x_{I+1} > 0 \text{ and } t > \sum_{i=0}^{I+1} i x_i \right\}$$
$$\bigcup \left\{ (x,t) \in \mathcal{S}_I \times [0,T) : x_{I+1} = 0 \text{ and } t = \sum_{i=0}^{I} i x_i \right\};$$

- *and when $a < 0$ and $I = -a - 1$,*

$$\mathcal{D}_a \doteq \left\{ (x,t) \in \mathcal{S}_I \times [0,T) : t = \sum_{i=0}^{I+1} i x_i \right\}.$$

In the first case the second set in the union reflects the fact that when $x_{I+1} = 0$ the number of balls thrown is exactly $\sum_{i=0}^{I} i x_i$, and similarly for the second case. When $a \in -\mathbb{N}$ it is only possible to throw balls into the categories $0, 1, \dots, -a-1$, and the only possible categories are $0, 1, \dots, -a$. Thus if there are n cells, there can at most be $-an$ balls thrown, and therefore $T \le -a$. When $T = -a$ all the cells have exactly $-a$ balls, which is not an interesting case to study. As a consequence, throughout this paper we assume $T < -a$. Also, because of the restriction on the possible categories we can (without loss) assume that $I = -a - 1$. Hence for $a < 0$ we assume without loss that

$$T < -a, \quad I = -a - 1. \tag{2.9}$$

2.2. LLN limits and formal derivation of the explicit solution. When constructing explicit solutions one needs some insight into the form of the solution. In this section we present a formal derivation of an explicit solution to (2.8) for the case $F(x) = 1_y(x) \cdot \infty$. Before doing so we calculate the LLN limits of the occupancy processes, a necessary ingredient in the solution.

Equations for the LLN limits can easily be derived directly, or alternatively by noting that they are the zero cost trajectories in the variational problem (2.8) with $F \equiv 0$. It will suffice to consider initial conditions of the form $x = e_k, k = 0, 1, \dots, I$, where $(e_k)_j$ is 1 if $j = k$ and 0 otherwise. Since the relative entropy vanishes only if $\theta(s) = \rho(s, \varphi(s))$, the LLN limits can be characterized by the system of ODEs

$$\dot{\varphi}(s) = M[\rho(s, \varphi(s))], \quad \varphi(t) = e_k. \tag{2.10}$$

Since the LLN limit is desired for all components of the occupancy process, we use the infinite system rather than the system truncated at $I+$. These are easy to solve because the equation for the jth component depends only on the $(j-1)$st component, and so one can solve first for the kth component and then bootstrap. To write the solution in explicit form, we need some notation. For all $a \in \mathbb{R}$, $a \neq 0$, and $i \in \mathbb{N}$, let

$$\binom{a}{i} \doteq \frac{\prod_{j=0}^{i-1}(a-j)}{i!}.$$

Note that if $a \in \mathbb{N}$ and $i > a$, then $\binom{a}{i} = 0$, and that if $a \notin \mathbb{N} \cup \{0\}$ and $i \in \mathbb{N}$, then $\binom{a}{i} \neq 0$. For $i \in \mathbb{N} \cup \{0\}$ and $a > 0, s \geq 0$ or $a \in -\mathbb{N}, 0 \leq s \leq -a$, define

$$\mathcal{Q}_i^a(s) \doteq \left(-\frac{s}{a}\right)^i \binom{-a}{i} \left(1 + \frac{s}{a}\right)^{-a-i}.$$

One can easily check that the solution to (2.10) is $\varphi_i(s) = 0$ if $i < k$, and

$$\varphi_{k+i}(s) = \mathcal{Q}_i^{a+k}\left(\frac{a+k}{a+t}(s-t)\right)$$

if $i \geq 0$. In the limit $a \to \infty$ (MB statistics) one obtains the Poisson distribution

$$\mathcal{Q}_i^{a+k}\left(\frac{a+k}{a+t}(s-t)\right) \to \mathcal{P}_i(s-t) = e^{-(s-t)}(s-t)^i/i!.$$

For the remainder of this section we assume $a \neq \infty$, with the understanding that analogous statements for $a = \infty$ can be obtained by passing to the limit.

We next present a formal and heuristic solution to the variational problem (2.7) that is based on probabilistic intuition. We do not attempt to directly solve for the minimizing φ, but rather heuristically derive an alternate rate function for $X^n(T)$. This alternative rate function is based on a different construction of the process. However, since rate functions are unique it should coincide with (2.7) when $A = \{y\}$. Recall that the variational problem is intended to approximate the normalized logarithm of a probability. If one decomposes a probability into products or conditional products, this will correspond to a decomposition of the quantity being minimized as a sum.

We wish to solve (2.7) when $A = \{y\}$. Suppose that x_i is interpreted as the size of the "pool" of cells that at time t are in category i. Hence there are approximately nx_i cells in pool i. The cells in a given pool are fixed for the rest of the construction. Let π_i^k denote the probability that a cell of category k at time t ends up being a cell of category $k+i$ at time T. Then satisfaction of the terminal constraint requires

$$y_i = \sum_{k=0}^{i} x_k \pi_{i-k}^k, \quad 0 \leq i \leq I, \qquad y_{I+1} = 1 - \sum_{k=0}^{I} y_k.$$

We use $y \doteq x \times \pi$ as shorthand for the last display. We require that the π^k be probabilities, and also require a constraint that corresponds to the fact that $n(T-t)$ balls will be placed in the prelimit problem:

$$x_k \sum_{j=0}^{\infty} \pi_j^k = x_k, \quad 0 \leq k \leq I+1, \quad \sum_{k=0}^{I+1} x_k \sum_{j=0}^{\infty} j\pi_j^k = T-t. \tag{2.11}$$

Let $\mathcal{F}(x,t;y,T)$ denote the set of $\pi = (\pi^0, \pi^1, \ldots, \pi^I, \pi^{I+1})$ which satisfy the last two displays. A terminal point y is *feasible* (for the given initial time and condition) if $\mathcal{F}(x,t;y,T)$ is not empty.

Owing to the fact that the un-normalized relative probabilities are *affine* in the number of balls currently in each cell, we can consider the allocation from a different perspective. We first study the random evolution of the number of balls contained in each distinct pool of cells (recall that the partition of cells into pools is determined by their status at time t). Because the un-normalized probabilities are affine, this can be done without knowing the details of how the balls are to be placed *within* the pool. This process is also Markovian, and its large deviation properties are easy to identify.

Once we know the total number of balls that will end up in each pool, we then consider the question of how they are distributed among cells within the pool. Here we make an approximation that is formal but reasonable. If the number of cells within a pool is large, then the statistical dependence between any two cells should be low. If the dependence in some sense disappears in the limit $n \to \infty$, then one might guess that the rate function for the empirical distribution within the pool is given by Sanov's theorem, where the distribution of the random number of balls within each cell is just the LLN distribution appropriate to that particular pool. By Sanov's theorem the rate function for the empirical distribution within the pool is a certain relative entropy. However, we must also impose the constraint on the (previously determined) number of balls that were placed into this particular pool, which adds a constraint to the rate function. Finally, the overall rate function is found by combining these two rates (allocation between pools and allocation within each pool). The function found in this manner will be proved to be the solution to the calculus of variations problem.

An argument based on Sanov's theorem shows that the variational problem for the allocation between the pools is

$$\inf \int_t^T R\left(u(s)\,\|w(s)\right) ds,$$

where

$$w_k(s) = \frac{(a+k)x_k + \int_t^s u_k(\tau)d\tau}{a+s}$$

is the probability that a ball is placed into pool k at time s. This is, in un-normalized form, equal to ax_k+[number of balls per cell in pool k]x_k, and the normalization is just $a+s$. The initial and terminal conditions are

$$w_k(t) = \frac{(a+k)x_k}{a+t}, \quad w_k(T) = \frac{(a+k)x_k + z_k(T-t)}{a+T},$$

where z_k is the mean number of additional balls per unit time put into pool k. The Euler–Lagrange equations for this problem are easily constructed and solved, and one obtains as the optimal trajectory

$$w_k(s) = \frac{1}{s+a}\left((a+k)x_k + (s-t)z_k\right)$$

(of course satisfaction of the Euler–Lagrange equations is not, in general, a sufficient condition for optimality, but since our discussion is simply to motivate the form of the solution, this point is of no consequence). The cost is

$$\int_t^T R\left([(s+a)w(s)]'\,\|w(s)\right) ds,$$

and for the optimal trajectory,

$$[(s+a)w_k(s)]' = z_k.$$

The integral can be explicitly evaluated and equals

$$\sum_{k=0}^{I+} z_k \cdot (T-t) \cdot \log\left(\frac{a+t}{a+k} \cdot \frac{z_k}{x_k}\right)$$
$$-\sum_{k=0}^{I+} [x_k \cdot (a+k) + z_k \cdot (T-t)] \cdot \log\left(\frac{a+k+\frac{z_k}{x_k}(T-t)}{a+k+\frac{a+k}{a+t}(T-t)}\right).$$

This identifies the first part of the overall rate function.

The second part is found by considering placement within each pool. The mean additional number of balls per cell in pool k is $(z_k/x_k)(T-t)$. According to the LLN, the number of additional balls in a typical cell from pool k has distribution $\mathcal{Q}^{a+k}\big(\frac{z_k}{x_k}(T-t)\big)$ if $k \le I$ and $\mathcal{Q}^{a+\tau(x,t)}\big(\frac{z_{I+}}{x_{I+}}(T-t)\big)$ if $k = I+$. Approximating the true empirical measure within a given pool by that of the empirical measure for iid random variables with the corresponding distribution, one formally obtains from Sanov's theorem the rate function $R\big(\pi^k \| \mathcal{Q}^{a+k}\big(\frac{z_k}{x_k}(T-t)\big)\big)$, together with the constraint $\sum_{i=0}^{\infty} i\pi_i^k = \frac{z_k}{x_k}(T-t)$ on the number of balls placed in pool k. Combining the different contributions from the various pools with the contribution due to the allocation between the pools and then applying the terminal constraint, one (again formally) obtains the rate function

$$\inf\left\{\sum_{k=0}^{I} x_k R\left(\pi^k \,\middle\|\, \mathcal{Q}^{a+k}\left(\frac{z_k}{x_k}(T-t)\right)\right)\right.$$
$$+ x_{I+} R\left(\pi^{I+} \,\middle\|\, \mathcal{Q}^{a+\tau(x,t)}\left(\frac{z_{I+}}{x_{I+}}(T-t)\right)\right) + \sum_{k=0}^{I+} z_k \cdot (T-t) \cdot \log\left(\frac{a+t}{a+k} \cdot \frac{z_k}{x_k}\right)$$
$$\left.-\sum_{k=0}^{I+} [x_k \cdot (a+k) + z_k \cdot (T-t)] \cdot \log\left(\frac{a+k+\frac{z_k}{x_k}(T-t)}{a+k+\frac{a+k}{a+t}(T-t)}\right)\right\},$$

where the infimum is over all π and z such that $\sum_{i=0}^{\infty} i\pi_i^k = \frac{z_k}{x_k}(T-t)$ and $x \times \pi = y$. However, a straightforward calculation using the specific form of $\mathcal{Q}^a$ and $\sum_{i=0}^{\infty} i\pi_i^k = \frac{z_k}{x_k}(T-t)$ gives

$$R\left(\pi^k \,\middle\|\, \mathcal{Q}^{a+k}\left(\frac{a+k}{a+t}(T-t)\right)\right) - R\left(\pi^k \,\middle\|\, \mathcal{Q}^{a+k}\left(\frac{z_k}{x_k}(T-t)\right)\right)$$
$$= \frac{z_k}{x_k}(T-t) \cdot \log\left(\frac{a+t}{a+k} \cdot \frac{z_k}{x_k}\right) - (a+k) \cdot \log\left(\frac{a+k+\frac{z_k}{x_k}(T-t)}{a+k+\frac{a+k}{a+t}(T-t)}\right)$$
$$- \frac{z_k}{x_k}(T-t) \cdot \log\left(\frac{a+k+\frac{z_k}{x_k}(T-t)}{a+k+\frac{a+k}{a+t}(T-t)}\right),$$

with an analogous result for $k = I+$. If follows that the rate function can be written in the simpler form

$$\inf_{\pi: x \times \pi = y}\left\{\sum_{k=0}^{I} x_k R\left(\pi^k \,\middle\|\, \mathcal{Q}^{a+k}\left(\frac{a+k}{a+t}(T-t)\right)\right)\right.$$
$$\left.+ x_{I+} R\left(\pi^{I+} \,\middle\|\, \mathcal{Q}^{a+\tau(x,t)}\left(\frac{a+\tau(x,t)}{a+t}(T-t)\right)\right)\right\},$$

with the infimum over z no longer necessary.

Let

$$\mathcal{J}(x,t;y) \doteq \inf_{\substack{\varphi\in C([t,T],\mathcal{S}_I)\\ \varphi(t)=x,\varphi(T)=y}} \mathcal{I}(x,t;\varphi). \tag{2.12}$$

The formal derivation just given suggests the following result, in which we also simplify further where the special cases of FD and MB statistics allow.

THEOREM 2.2 (explicit formula for the rate function). *Consider an initial condition $(x,t)\in\mathcal{D}_a$ and a feasible terminal condition y. If $a\in(0,\infty)$, define $\tau(x,t)$ by* (2.4)*. Then the quantity $\mathcal{J}(x,t;y)$ defined in* (2.12) *has the representation*

$$\begin{aligned}\mathcal{J}(x,t;y) = \min_{\pi\in\mathcal{F}(x,t;y,T)} \Bigg\{ &\sum_{k=0}^{I} x_k R\left(\pi^k \middle\| \mathcal{Q}^{a+k}\left(\frac{a+k}{a+t}(T-t)\right)\right)\\ &+ x_{I+1} R\left(\pi^{I+1} \middle\| \mathcal{Q}^{a+\tau(x,t)}\left(\frac{a+\tau(x,t)}{a+t}(T-t)\right)\right)\Bigg\}.\end{aligned}$$

If $a\in-\mathbb{N}$ with $I=-a-1$, then $\tau(x,t)=I+1$ and

$$\mathcal{J}(x,t;y) = \min_{\pi\in\mathcal{F}(x,t;y,T)} \left\{\sum_{k=0}^{I+1} x_k R\left(\pi^k \middle\| \mathcal{Q}^{a+k}\left(\frac{a+k}{a+t}(T-t)\right)\right)\right\}.$$

In the final case of $a=\infty$, we have

$$\mathcal{J}(x,t;y) = \min_{\pi\in\mathcal{F}(x,t;y,T)} \left\{\sum_{k=0}^{I+1} x_k R\left(\pi^k \,\|\mathcal{P}(T-t)\right)\right\}.$$

Remark 2.3. Although the minimization problems in Theorem 2.2 appear to be infinite dimensional, they can in fact be reduced to finite dimensional problems. This is because if π^k is the minimizer, then π^k_j takes a prescribed form for $j>I$. In fact, all π^k_j can be represented in terms of no more than $I+3$ Lagrange multipliers as in (2.20) below.

2.2.1. The Hamilton–Jacobi–Bellman equation. Given Theorem 2.2 one can solve the problem with a general terminal condition F. Conversely, if the problem with terminal cost can be solved for a sufficiently broad class of F, one can derive Theorem 2.2. This is how we will prove the theorem, and moreover, the proof will be based on the fact that finite dimensional representations analogous to those in Theorem 2.2, but with these terminal costs, are *classical sense* solutions to the associated PDE. The proof also has a number of side benefits, such as convenient representations for the various derivatives of the solution in terms of Lagrange multipliers.

The calculus of variations problem (2.8) has a natural control interpretation, where $\theta(s)$, $t\le s\le T$, is the control, $\dot\varphi(s)=M[\theta](s)$ are the dynamics, $R(\theta(s)||\rho(s,\varphi(s)))$ is the running cost, and $F(x)$ is the terminal cost. It is expected that if we define

$$V(x,t) \doteq \inf_{\varphi\in C([t,T],\mathcal{S}_I),\varphi(t)=x} \left\{\int_t^T R(\theta(s)||\rho(s,\varphi(s)))\,ds + F(\varphi(T))\right\}, \tag{2.13}$$

then $V(x,t)$ is a weak sense solution to the HJB equation

$$W_t + H(W_x, x, t) = 0$$

and terminal condition

$$W(x,T) = F(x).$$

Here the Hamiltonian $H(p,x,t)$ is defined by

$$H(p,x,t) \doteq \inf_{\theta\in\mathcal{S}_I} \left[\langle p, M[\theta]\rangle + R\left(\theta \,\|\rho\left(t,x\right)\right)\right]$$

and W_t and W_x denote the partial derivative with respect to t and gradient in x, respectively. Note that by the representation formula [1, Proposition 1.4.2], the infimum in the definition of $H(p,x,t)$ can be evaluated, yielding

$$\text{(2.14)}\qquad \left\{ \begin{array}{c} W_t = \log\left(\sum_{k=0}^{I} x_k \left(\frac{a+k}{a+t}\right)\exp\left(W_{x_k} - W_{x_{k+1}}\right) + x_{I+1}\left(\frac{a+\tau(x,t)}{a+t}\right)\right), \\ W(x,T) = F(x). \end{array}\right.$$

Note the use of the convenient expression (2.5) for $\rho_{I+}(x,t)$.

For a general smooth F, (2.14) need not have a smooth (C^1) solution. However, for affine terminal costs $F(x) = \langle \ell, x\rangle + b$ there is a C^1 solution (it is in fact the unique solution), and as remarked above, these solutions can be used to carry out a fairly complete analysis of the problem with more general terminal conditions. Indeed, for a general (proper) convex terminal cost $F(x)$, the Legendre transform gives a representation of the form

$$F(x) = \sup_{\beta\in\mathbb{R}^{I+2}} \left[\langle \beta, x\rangle - h(\beta)\right]$$

for some proper convex function h. Let $V^F(x,t)$ denote the solution (explicit or otherwise) to the calculus of variations problem (2.13) with terminal cost $F(\cdot)$. Then one can show

$$V^F(x,t) = \sup_{\beta\in\mathbb{R}^{I+2}} V^{\{\langle\beta,\cdot\rangle - h(\beta)\}}(x,t),$$

and an analogous formula for $U^F(x,t) \doteq \inf[\mathcal{J}(x,t;y) + F(y)]$. Given Proposition 2.4 below, $V^F = U^F$ then follows. Since $\infty \cdot 1_{\{y\}^c}$ is a proper convex function, the formula can be extended even further to very general F.

Observe that W is a solution of just the PDE alone (i.e., without the terminal condition) if and only if $W + c$ is a solution for any real number c. Since x is a probability vector, we can write

$$F(x) = \langle \ell, x\rangle + b = \sum_{i=0}^{I} (\ell_i - \ell_{I+1}) x_i + \ell_{I+1} + b,$$

and so by the previous sentence it suffices to prove the representation under the conditions $\ell_{I+1} = 0$ and $b = 0$.

2.3. Explicit solution for affine terminal costs.

PROPOSITION 2.4. *Consider $(x,t) \in \mathcal{D}_a$ and $F(y) = \langle \ell, y \rangle$, where $\ell \in \mathbb{R}^{I+2}$ and $\ell_{I+1} = 0$. Define*

$$V(x,t) \doteq \inf_{\varphi \in C([t,T],\mathcal{S}_I),\varphi(t)=x} \left\{ \int_t^T R\left(\theta(s) || \rho\left(s, \varphi(s)\right)\right) ds + F(\varphi(T)) \right\},$$

and

$$(2.15)\quad U(x,t) \doteq \min_{\pi \in \mathcal{F}(x,t;T)} \left\{ \sum_{k=0}^{I} x_k R\left(\pi^k \middle\| \mathcal{Q}^{a+k}\left(\frac{a+k}{a+t}(T-t)\right)\right) \right. \\ \left. + x_{I+1} R\left(\pi^{I+1} \middle\| \mathcal{Q}^{a+\tau(x,t)}\left(\frac{a+\tau(x,t)}{a+t}(T-t)\right)\right) + F(x \times \pi) \right\},$$

where $\pi \in \mathcal{F}(x,t;T)$ means that π satisfies the constraints in (2.11)*. Then $V(x,t) = U(x,t)$.*

The proof of this result is given in the next subsection. We close this subsection with remarks on the LLN limit distributions.

We will use the fact that if $a \in \mathbb{R}$ and $|z| < 1$, then the binomial expansion

$$(1+z)^{-a} = \sum_{i=0}^{\infty} \binom{-a}{i} z^i, \qquad \binom{a}{i} \doteq \frac{\prod_{j=0}^{i-1}(a-j)}{i!}$$

is valid, and if $-a \in \mathbb{N}$, then the sum contains only a finite number of nonzero terms and is valid for all $z \in \mathbb{R}$. Recall that for $i \in \mathbb{N} \cup \{0\}$ and $a > 0, s \geq 0$ or $a \in -\mathbb{N}, 0 \leq s \leq -a$, then

$$\mathcal{Q}_i^a(s) \doteq \left(-\frac{s}{a}\right)^i \binom{-a}{i} \left(1 + \frac{s}{a}\right)^{-a-i}.$$

If $a > 0, s \geq 0$, and $|s\theta/(a+s)| < 1$, then the binomial expansion gives

$$\begin{aligned}\sum_{i=0}^{\infty} \mathcal{Q}_i^a(s)\theta^i &= \sum_{i=0}^{\infty} \left(-\frac{s}{a}\right)^i \binom{-a}{i} \left(1+\frac{s}{a}\right)^{-a-i} \theta^i \\ &= \left(1+\frac{s}{a}\right)^{-a} \sum_{i=0}^{\infty} \left(-\frac{s}{s+a}\theta\right)^i \binom{-a}{i} \\ &= \left(1+\frac{s}{a}\right)^{-a} \left(1 - \frac{s\theta}{s+a}\right)^{-a} \\ &= \left(1 + \frac{s}{a} - \frac{s\theta}{a}\right)^{-a} \\ &= \left(1 + \frac{s}{a}(1-\theta)\right)^{-a}.\end{aligned}$$

We thus have the following expressions, where the second one may be justified by a very similar calculation (when $|s\theta/(a+s)| < 1$):

$$(2.16)\quad \sum_{i=0}^{\infty} \mathcal{Q}_i^a(s)\theta^i = \left(1 + \frac{s}{a}(1-\theta)\right)^{-a}, \quad \sum_{i=0}^{\infty} i\mathcal{Q}_i^a(s)\theta^i = s\theta\left(1 + \frac{s}{a}\left(1-\theta\right)\right)^{-a-1}.$$

Note also that when $-a \in \mathbb{N}$ and $0 \le s \le -a$ the number of nonzero summands is finite and the formulas again hold. If $-a \in \mathbb{N}$ and $0 \le s \le -a$ or $a > 0$ and $s \ge 0$, then $\mathcal{Q}_i^a(s) \ge 0$ for $i \in \mathbb{N} \cup \{0\}$. Letting $\theta \uparrow 1$ in the first expression shows that under these conditions $\mathcal{Q}_i^a(s)$ defines a probability measure on $\mathbb{N} \cup \{0\}$. When $a = 0$ we use the limiting values

$$\mathcal{Q}_0^0(s) = 1,\ \mathcal{Q}_i^0(s) = 0$$

for all $i \in \mathbb{N}$ and $s \ge 0$. For later use note that similar calculations show that if $-a \in \mathbb{N}$ and $0 \le s \le -a$ or $a > 0$, $s \ge 0$, and $s/(a+s) < 1$, then

$$\sum_{i=0}^{\infty} i^2 \mathcal{Q}_i^a(s) - \left[\sum_{i=0}^{\infty} i \mathcal{Q}_i^a(s) \right]^2 = s + \frac{s^2}{a}. \tag{2.17}$$

2.3.1. Analysis of the finite dimensional minimization problem. We now focus on proving Proposition 2.4. We will do so by proving that $U(x,t)$ is a classical sense solution to the HJB equation (2.14). A modification of the standard verification argument [4] can then be used to show that $V(x,t) = U(x,t)$. The classical verification argument consists of two parts. One first considers any valid occupancy process and control (φ, θ) for the initial condition (x,t). If U is a smooth solution to the PDE (2.14) in neighborhood of $\{(\varphi(s), s) : t \le s \le T\}$ and if $U(\varphi(T), T) = \langle \varphi(T), \ell \rangle$, then the chain rule implies that the cost along this trajectory is at least $U(x,t)$. The reverse inequality is proved by defining an optimal feedback control through the HJB equation, using this control to construct a trajectory, and then verifying (once again via the chain rule) that the cost for this control is $U(x,t)$. The characterization of $V(x,t)$ as an infimum over all valid occupancy processes and controls that start at (x,t) then gives $V(x,t) = U(x,t)$. However, we have to clarify here what is meant by a "classical sense" solution to (2.14). The difficulty is that $U(x,t)$ is only well defined on the set $\mathcal{D}_a$, which does not have an interior.

Given any point $(x,t) \in \mathcal{D}_a$, we will prove that one can extend $U(x,t)$ smoothly to a neighborhood of (x,t) in $\mathbb{R}^{I+2} \times \mathbb{R}$. To be more precise, for any such (x,t) we will show there exists a neighborhood $\mathcal{U} \subset \mathbb{R}^{I+2} \times \mathbb{R}$ of (x,t) and a function $\bar{U} \in C^\infty(\mathcal{U}, \mathbb{R})$, such that $\bar{U}(y,s) = U(y,s)$ for $(y,s) \in \mathcal{U} \cap \mathcal{D}_a$, and that $\bar{U}$ satisfies (2.14) in $\mathcal{U} \cap \mathcal{D}_a$. One can then use $\bar{U}$ in place of U in the verification argument, since any feasible trajectory will never leave $\mathcal{D}_a$.

To analyze $U(x,t)$ we formulate an appropriate Lagrangian. Let

$$\begin{aligned} f(x,t;\pi) \doteq \sum_{k=0}^{I} x_k R\left(\pi^k \,\middle\|\, \mathcal{Q}^{a+k}\left(\frac{a+k}{a+t}(T-t) \right) \right) \\ + x_{I+1} R\left(\pi^{I+1} \,\middle\|\, \mathcal{Q}^{a+\tau(x,t)}\left(\frac{a+\tau(x,t)}{a+t}(T-t) \right) \right) + \langle \ell, x \times \pi \rangle \end{aligned} \tag{2.18}$$

and for a set of Lagrange multipliers $\Lambda \doteq (\lambda, \mu) = (\lambda_0, \lambda_1, \ldots, \lambda_I, \lambda_{I+1}, \mu)$, let

$$\begin{aligned} &L(x,t;\Lambda;\pi) \\ &\quad \doteq f(x,t;\pi) + \sum_{k=0}^{I+1} \lambda_k x_k \left(1 - \sum_{j=0}^{\infty} \pi_j^k \right) + \mu \left(T - t - \sum_{k=0}^{I+1} x_k \sum_{j=0}^{\infty} j \pi_j^k \right). \end{aligned} \tag{2.19}$$

It follows from the definition of $U(x,t)$ that $U(x,t) = \inf_\pi \sup_\Lambda L(x,t;\Lambda;\pi)$.

Note that by the joint convexity of relative entropy, $L(x,t;\Lambda;\pi)$ is convex in π. Thus (2.15) is a standard convex programming problem with linear constraints, except that the minimization is over a variable π which is infinite dimensional. Hence the standard Lagrange multiplier method does not apply directly. If we temporarily ignore this issue, then to guess the form of the minimizer one would of course set $D_\pi L(x,t;\Lambda;\pi)=0$ to get $\pi=\pi(x,t;\Lambda)$, where D_π stands for the gradient in π and

$$\begin{aligned}
&\pi_j^k\,(x,t;\Lambda) = \mathcal{Q}_j^{a+k}\left(\frac{a+k}{a+t}(T-t)\right)e^{\lambda_k-1+j\mu-\ell_{k+j}}\\
&\qquad k=0,1,\ldots,I \text{ and } j\geq 0,\\
&\pi_j^{I+1}\,(x,t;\Lambda) = \mathcal{Q}_j^{a+\tau(x,t)}\left(\frac{a+\tau(x,t)}{a+t}(T-t)\right)e^{\lambda_{I+1}-1+j\mu}.
\end{aligned} \tag{2.20}$$

Here, for notational simplicity, we extend ℓ in Proposition 2.4 by letting $\ell_i=0$ when $i>I$. Note in particular that $\{\pi^k\}$ will depend on x only when $k=I+1$. Observe also that setting $D_\Lambda L(x,t;\Lambda;\pi)=0$ gives the constraints (2.11). For any $(x,t)\in\mathbb{R}^{I+2}\times\mathbb{R}$ and $\Lambda\in\mathbb{R}^{I+3}$, let $\pi(x,t;\Lambda)$ be determined by (2.20) and define $G:\mathbb{R}^{I+2}\times\mathbb{R}\times\mathbb{R}^{I+3}\mapsto\mathbb{R}^{I+3}$ by

$$G_k(x,t;\Lambda) = \left(1-\sum_{j=0}^{\infty}\pi_j^k\,(x,t;\Lambda)\right),\qquad k=0,1,\ldots,I+1,$$

$$G_{I+2}(x,t;\Lambda) = \left(T-t-\sum_{k=0}^{I+1}x_k\sum_{j=0}^{\infty}j\pi_j^k\,(x,t;\Lambda)\right).$$

In the next theorem we show that the $\pi(x,t;\lambda)$ defined in (2.20) indeed give the minimizer of (2.15).

Theorem 2.5. *For any $(x,t)\in\mathcal{D}_a$ define $U(x,t)$ by* (2.15). *Then there exists $\Lambda\in\mathbb{R}^{I+3}$ so that $G(x,t;\Lambda)=0$, and $\pi(x,t;\Lambda)$ is a minimizer of* (2.15). *Thus $U(x,t)=L\big(x,t;\Lambda;\pi(x,t;\Lambda)\big)$. In addition, the Λ that satisfies $G(x,t;\Lambda)=0$ is unique. Hence if $G(x,t;\Lambda)=0$ for some $\Lambda\in\mathbb{R}^{I+3}$, then $\pi(x,t;\Lambda)$ is a minimizer of* (2.15).

The proof is divided into three lemmas. For a point $(x,t)\in\mathcal{D}_a$, quantities of the following sort will appear frequently in the proofs of the lemmas:

$$\begin{aligned}
\bar\pi_j^k &\doteq \mathcal{Q}_j^{a+k}\left(\frac{a+k}{a+t}(T-t)\right)\\
\bar\pi_j^{I+1} &\doteq \mathcal{Q}_j^{a+\tau(x,t)}\left(\frac{a+\tau(x,t)}{a+t}(T-t)\right)\\
&\text{for } k=0,1,\ldots,I;\quad j=0,1,\ldots.
\end{aligned} \tag{2.21}$$

In particular, it will often be the case that (2.16) must be invoked, with a there replaced by $a+k$ [or $a+\tau(x,t)$] and s there replaced by the corresponding argument in the expression above. We note that the conditions required for (2.16) will always hold so long as $t\in[0,T]$. This is straightforward to check in the case of $a>0$. For the case $-a\in\mathbb{N}$, it uses that $-a=I+1$, $T\leq -a$, and that always $\tau(x,t)=I+1$. Thus, for example, for any $k\in\{0,\ldots,-a\}$ and $t\in[0,T]$, $(a+k)(T-t)/(a+t)\geq 0$ and $(T-t)/(-a-t)\leq 1$ shows that $(a+k)(T-t)/(a+t)\leq -a-k$, as required for (2.16).

LEMMA 2.6 (general properties). *For any* $(x,t) \in \mathcal{D}_a$ *define* $U(x,t)$ *by* (2.15). *Then* $\mathcal{F}(x,t;T)$ *is nonempty, minimizing measures* π^* *exist, and if* k *is such that* $x_k > 0$ *and* $j \in \{0,1,\ldots\}$, *then*

$$\bar{\pi}_j^k > 0 \quad \textit{implies} \quad \pi_j^{*k} > 0. \tag{2.22}$$

Proof. According to (2.16) the quantities in (2.21) are probabilities that satisfy (2.11). This shows that $\mathcal{F}(x,t;T)$ is nonempty. Next note that with this notation, we can rewrite (2.18) as

$$f(x,t;\pi) = R\left(x \otimes \pi \,\|x \otimes \bar{\pi}\,\right) + \langle \ell, x \times \pi \rangle, \tag{2.23}$$

where $(x \otimes \pi)_{i,j} = x_i \pi_j^i$. Since the relative entropy has compact level sets in the first argument [1, Lemma 1.4.3(c)], the existence of a minimizer of (2.15) follows. In addition, because of the strict convexity in that argument we know that the minimizer is unique up to those $\left\{\pi_\cdot^k\right\}$ with $x_k > 0$.

For a general initial condition (x,t) let $\mathcal{K} \doteq \{k : x_k > 0\}$. Then the choice of $\left\{\pi_\cdot^k : k \notin \mathcal{K}\right\}$ will not affect either the constraint (2.11) or the objective function (2.23). Hence we can consider the equivalent minimization problem over $\mathcal{M}_{(x,t)} = \left\{\pi_j^k : k \in \mathcal{K}, j = 0,1,\ldots\right\}$. As discussed in the previous paragraph, a minimizer in $\mathcal{M}_{(x,t)}$ exists and is unique. Let this minimizer be denoted π^*.

Lastly we must show (2.22). Let $\pi^\epsilon \doteq (1-\epsilon)\pi^* + \epsilon\bar{\pi}$, where $\bar{\pi}$ is defined in (2.21), and let $f(\epsilon) = f(x,t;\pi^\epsilon)$. By computing the derivative of $f(\epsilon)$ explicitly, it is readily observed that if (2.22) does not hold, then $f'(\epsilon) \to -\infty$ as $\epsilon \to 0$. Thus (2.22) must be true since otherwise π^* is not the minimizer. $\square$

LEMMA 2.7 (characterization of the minimizer). *For any* $(x,t) \in \mathcal{D}_a$ *define* $U(x,t)$ *by* (2.15). *Then there exists* $\Lambda \in \mathbb{R}^{I+3}$ *so that* $G(x,t;\Lambda) = 0$, *and* $\pi(x,t;\Lambda)$ *is a minimizer of* (2.15).

Proof. We want to argue that the minimizers must take the form of (2.20). However, there is a difficulty since $\mathcal{M}_{(x,t)}$ can be infinite dimensional. To deal with this we use a truncation argument adapted from one in [2]. For any $N \in \mathbb{N}$ let

$$T^{(N)} \doteq \sum_{k\in\mathcal{K}} x_k \sum_{j=0}^{N} j\pi_j^{*k}, \quad \alpha_k^{(N)} \doteq \sum_{j=0}^{N} \pi_j^{*k}, \quad k \in \mathcal{K},$$

and also let

$$f^{(N)}(x,t;\pi) \doteq f(x,t;\hat{\pi}),$$

where

$$\begin{aligned} \hat{\pi}_j^k &= \pi_j^k \quad &&\text{for } k \in \mathcal{K}, j \le N,\\ \hat{\pi}_j^k &= \pi_j^{*k} \quad &&\text{for } k \in \mathcal{K}, j > N. \end{aligned}$$

Since π^* is the minimizer of (2.15), we automatically obtain

$$U(x,t) = \min_{\pi} f^{(N)}(x,t;\pi), \tag{2.24}$$

where the minimum is subject to the constraints

$$\sum_{k\in\mathcal{K}} x_k \sum_{j=0}^{N} j\pi_j^{k} = T^{(N)}, \quad \sum_{j=0}^{N} \pi_j^{k} = \alpha_k^{(N)}, \quad k \in \mathcal{K}.$$

We can now apply the standard Lagrange multiplier method to (2.24). The first step is to formulate the Lagrangian for this finite dimensional problem:

$$L^N(x,t;\Lambda;\pi)$$
$$\doteq f^N(x,t;\pi) + \sum_{k\in\mathcal{K}} \lambda_k^{(N)} x_k \left(\alpha_k^{(N)} - \sum_{j=0}^{N} \pi_j^k\right) + \mu^{(N)}\left(T^{(N)} - \sum_{k=0}^{I+1} x_k \sum_{j=0}^{N} j\pi_j^k\right).$$

We have that $\left\{\pi_j^{*k} : k \in \mathcal{K}, j \le N\right\}$ satisfies the constraints in (2.24), and by (2.22) we know that $\pi_j^{*k} > 0$ if $\bar\pi_j^k > 0$. Hence by [6, Corollary 28.2.2] and [6, Theorem 28.3] applied to (2.24), there must exist a set of Lagrange multipliers $\lambda_k^{(N)}, \mu^{(N)}$ so that the minimizer of (2.24) π_j^{*k} has the form

$$\pi_j^{*k} = \bar\pi_j^k e^{\lambda_k^{(N)} - 1 + j\mu^{(N)} - \ell_{k+j}} \tag{2.25}$$

for $k \in \mathcal{K}$ and $0 \le j \le N$. If $k + j > I$, then since $\ell_{k+j} = 0$,

$$\frac{\pi_{j+1}^{*k}}{\pi_j^{*k}} = C \cdot e^{\mu^{(N)}},$$

where C does not depend on N. Thus $\mu^{(N)}$ is independent of N, and hence $\lambda^{(N)}$ is also independent of N. Since the choice of N is arbitrary, we then know that for all $k \in \mathcal{K}$ and $j = 0, 1, \dots, \pi_j^{*k}$ indeed has the form in (2.20) for a suitable choice of λ_k and μ. For $k \notin \mathcal{K}$, we can simply define π_j^{*k} as in (2.25) and then solve for λ_k from the normalization constraint $\sum_{j=0}^{\infty} \pi_j^{*k} = 1$. When defined in this way, $\Lambda^* = (\lambda_0, \dots, \lambda_{I+1}, \mu)$ automatically satisfies $G(x,t;\Lambda^*) = 0$.

For $k \in \mathcal{K}$ the corresponding λ_k are a Kuhn–Tucker vector as in [6, Corollary 28.2.2], and hence each $\lambda_k < \infty$. However, for $k \notin \mathcal{K}$ the finiteness of λ_k is not automatic.

To show the finiteness, we first insert the explicit form of $\pi_j^{I+1}(x,t;\Lambda)$ from (2.20) into $G_{I+1}(x,t,\Lambda) = 0$ to obtain

$$\sum_{j=0}^{\infty} \mathcal{Q}_j^{a+\tau(x,t)}\left(\frac{a+\tau(x,t)}{a+t}(T-t)\right) e^{\lambda_{I+1} - 1 + j\mu} = 1.$$

Using (2.16) to evaluate the sum gives

$$\lambda_{I+1} = (a+\tau(x,t)) \log\left(\frac{a+T-e^{\mu}(T-t)}{a+t}\right) + 1.$$

For notational simplicity define

$$\eta(t,\mu) \doteq \log\left(\frac{a+T-e^{\mu}(T-t)}{a+t}\right), \qquad \lambda(x,t;\mu) \doteq (a+\tau(x,t))\eta(t,\mu) + 1. \tag{2.26}$$

Then $\lambda_{I+1} = \lambda(x,t;\mu)$. Choose $C < \infty$ such that $|\ell_k| \le C$ for $0 \le k \le I$. Then

$$\sum_{j=0}^{\infty} \bar\pi_j^k e^{\lambda_k - 1 + j\mu - C} \le \sum_{j=0}^{\infty} \pi_j^k(x,t;\Lambda) \le \sum_{j=0}^{\infty} \bar\pi_j^k e^{\lambda_k - 1 + j\mu + C}.$$

A calculation of the same sort that gave the display above (2.26) gives

$$(a+k)\eta(t,\mu)+1-C \leq \lambda_k \leq (a+k)\eta(t,\mu)+1+C, \qquad k=0,\ldots,I.$$

Hence $\lambda_k < \infty$ so long as $\lambda_i < \infty$ for some $i = 0,1,\ldots,I,I+1$, which is true by [6, Corollary 28.2.2]. This completes the proof that for any $(x,t) \in \mathcal{D}_a$ there exists $\Lambda \in \mathbb{R}^{I+3}$ so that $G(x,t;\Lambda) = 0$ and $\pi(x,t;\Lambda)$ is a minimizer of (2.15). ◻

The next lemma will focus on the claim that for $(x,t) \in \mathcal{D}_a$, there is only one Λ that satisfies $G(x,t,\Lambda) = 0$, which together with the previous lemma completes the proof of Theorem 2.5.

LEMMA 2.8 (uniqueness of characterization). *For $(x,t) \in \mathcal{D}_a$, there is only one $\Lambda \in \mathbb{R}^{I+3}$ such that $G(x,t,\Lambda) = 0$.*

Proof. Recalling the definition of $\bar{\pi}$ in (2.21), notice that (2.20) is simply

$$\pi_j^k(x,t;\Lambda) = \bar{\pi}_j^k e^{\lambda_k - 1 + j\mu - \ell_{k+j}} \quad \text{for } k = 0,1,\ldots,I+1, \quad j = 0,1,\ldots.$$

As noted previously, for any $(x,t) \in \mathcal{D}_a$ we can assume that each $\{\bar{\pi}_\cdot^k\}$ is a valid probability vector. Thus $\pi_j^k(x,t;\Lambda) \geq 0$ and for each k at least one of $\{\pi_j^k(x,t;\Lambda), j = 0,1,\ldots\}$ is strictly positive. If

$$\alpha_k \doteq \sum_{j=0}^{\infty} \pi_j^k(x,t;\Lambda), \qquad T_k \doteq \sum_{j=0}^{\infty} j\pi_j^k(x,t;\Lambda),$$

then $\alpha_k > 0$ for each $k = 0,1,\ldots,I,I+1$, and any $\Lambda \in \mathbb{R}^{I+3}$. Using the particular dependency of $\pi_j^k(x,t;\Lambda)$ on Λ, one can compute

$$\begin{cases} \frac{\partial \pi_j^k(x,t;\Lambda)}{\partial \lambda_k} &= \pi_j^k(x,t;\Lambda), \\ \frac{\partial \pi_j^k(x,t;\Lambda)}{\partial \mu} &= j\pi_j^k(x,t;\Lambda), \\ \frac{\partial \pi_j^k(x,t;\Lambda)}{\partial \lambda_l} &= 0, \quad l \neq k. \end{cases}$$

It is straightforward to construct a dominating function of the form $\bar{\pi}_j^k \cdot C \cdot D^j$ for suitable constants C and D, and hence by the Lebesgue dominated convergence theorem one can compute $D_\Lambda G(x,t;\Lambda)$ explicitly as

$$\begin{pmatrix} -\alpha_0 & \cdots & 0 & -T_0 \\ \vdots & \cdots & \vdots & \vdots \\ 0 & \cdots & -\alpha_{I+1} & -T_{I+1} \\ -x_0 T_0 & \cdots & -x_{I+1}T_{I+1} & -\sum_{k=0}^{I+1} x_k \sum_{j=0}^{\infty} j^2 \pi_j^k(x,t;\Lambda) \end{pmatrix}.$$

Using elementary row operations to make the matrix upper triangular, we see that $\{-\alpha_k : k = 0,1,\ldots,I+1\}$ and $\sum_{k=0}^{I+1} \frac{x_k}{\alpha_k} T_k^2 - \sum_{k=0}^{I+1} x_k \sum_{j=0}^{\infty} j^2 \pi_j^k(x,t;\Lambda)$ are the eigenvalues of $D_\Lambda G(x,t;\Lambda)$. We have already observed that $\alpha_k > 0$ for all $k = 0,1,\ldots,I+1$. Also, for every $k = 0,1,\ldots,I+1$ the Cauchy–Schwarz inequality implies

$$\left(\sum_{j=0}^{\infty} j^2 \pi_j^k(x,t;\Lambda)\right)\left(\sum_{j=0}^{\infty} \pi_j^k(x,t;\Lambda)\right) \geq \left(\sum_{j=0}^{\infty} j\pi_j^k(x,t;\Lambda)\right)^2.$$

It is easy to verify that the necessary condition for equality [$\pi_j^k = j^2\pi_j^k$ for all j] does not hold. Hence the inequality is strict, and therefore

$$\left(\sum_{j=0}^{\infty} j^2 \pi_j^k(x,t;\Lambda)\right) > \frac{T_k^2}{\alpha_k}.$$

Thus $D_\Lambda G(x,t;\Lambda)$ is negative definite for all $\Lambda \in \mathbb{R}^{I+3}$.

Now we can prove the uniqueness of Λ. Suppose there are two different $\Lambda_1, \Lambda_2 \in \mathbb{R}^{I+3}$ such that $G(x,t;\Lambda_i) = 0,\ i = 1,2$. Define $\Lambda(\epsilon) \doteq \epsilon\Lambda_1 + (1-\epsilon)\Lambda_2$ and

$$h(\epsilon) \doteq \langle G(x,t;\Lambda(\epsilon)), \Lambda_1 - \Lambda_2\rangle.$$

Then $h'(\epsilon) = (\Lambda_1 - \Lambda_2)^T \cdot D_\Lambda G \cdot (\Lambda_1 - \Lambda_2)$. Since $D_\Lambda G(x,t;\Lambda)$ is always negative definite, $h'(\epsilon) < 0$ for all $0 < \epsilon < 1$. However, $h(0) = h(1) = 0$. This contradiction shows that $G(x,t;\Lambda) = 0$ has a unique solution in Λ. $\square$

The next theorem considers differentiability properties of $U(x,t)$. As mentioned previously, we first extend the definition of $U(x,t)$ to a neighborhood of (x,t) in $\mathbb{R}^{I+2} \times \mathbb{R}$, label this extension $\bar{U}(x,t)$, and then show $\bar{U}(x,t)$ is differentiable in the normal Euclidean sense. For our needs (a verification argument) the function $\bar{U}(x,t)$ can be used in lieu of $U(x,t)$.

THEOREM 2.9. *Fix $(x,t) \in \mathcal{D}_a$ and define $U(x,t)$ by* (2.15). *Then there is an open neighborhood $\mathcal{U}$ of (x,t) and an extension $\bar{U}$ of U from $\mathcal{D}_a \cap \mathcal{U}$ to $\mathcal{U}$ which is differentiable on $\mathcal{U}$.*

Proof. By Theorem 2.5, for any $(x,t) \in \mathcal{D}_a$ there exists Λ so that $G(x,t;\Lambda) = 0$ and $U(x,t) = L\big(x,t;\Lambda,\pi(x,t;\Lambda)\big)$. A natural approach to proving smoothness would be to apply the implicit function theorem. Recall that $\tau(x,t)$ is defined as the mean number of balls per cell in category $I+$ in (2.4). A difficulty with a direct application of the implicit function theorem is that this nonsmooth function appears in the constraints involving G_{I+1} and G_{I+2}. To avoid this difficulty we consider an equivalent but less obvious formulation of the constraint.

As discussed above (2.26), if the Lagrange multiplier λ_{I+1} is set to $\lambda(x,t;\mu)$, then the constraint $G_{I+1}(x,t;\Lambda) = 0$ will hold automatically. We will work with the reduced set of multipliers $\tilde{\Lambda} \doteq \{\lambda_0, \ldots, \lambda_I, \mu\}$ and the definition

$$\Lambda\left(x,t;\tilde{\Lambda}\right) \doteq \{\lambda_0, \lambda_1, \ldots, \lambda_I, \lambda(x,t;\mu), \mu\}. \tag{2.27}$$

Setting

$$H(x,t;\tilde{\Lambda}) \doteq L\left(x,t;\Lambda\left(x,t;\tilde{\Lambda}\right);\pi\left(x,t;\Lambda\left(x,t;\tilde{\Lambda}\right)\right)\right) \tag{2.28}$$

gives $U(x,t) = H(x,t;\tilde{\Lambda})$.

To apply the implicit function theorem we must show that there are smooth constraints that characterize $\tilde{\Lambda}$. For $i = 0,\ldots,I$ we use $\tilde{G}_i(x,t;\tilde{\Lambda}) = 0$, where $\tilde{G}_i(x,t;\tilde{\Lambda}) = G_i(x,t;\Lambda)$. These constraints are equivalent since π_j^k does not depend on λ_{I+1} for $k \le I$. Since $G_{I+1}(x,t;\Lambda) = 0$ holds automatically, we need only define $\tilde{G}_{I+1}$ so that $\tilde{G}_{I+1}(x,t;\tilde{\Lambda}) = 0$ is equivalent to $G_{I+2}(x,t;\Lambda) = 0$. We have

$$G_{I+2}(x,t;\Lambda) = T - t - \sum_{k=0}^{I} x_k \sum_{j=0}^{\infty} j\pi_j^k(x,t;\Lambda) - x_{I+1}\sum_{j=0}^{\infty} j\pi_j^{I+1}(x,t;\Lambda),$$

and thus set

$$\tilde{G}_{I+1}(x,t;\tilde{\Lambda}) \doteq G_{I+2}(x,t;\Lambda(x,t;\tilde{\Lambda})). \tag{2.29}$$

Since $\pi_j^k(\cdot)$ does not depend on λ_{I+1} when $k \le I$ we abuse notation and write the terms of the form $\pi_j^k(x,t;\Lambda(x,t;\tilde{\Lambda}))$ as $\pi_j^k(x,t;\tilde{\Lambda})$. Thus

$$\tilde{G}_{I+1}(x,t;\tilde{\Lambda}) = T - t - \sum_{k=0}^{I} x_k \sum_{j=0}^{\infty} j\pi_j^k(x,t;\tilde{\Lambda}) - x_{I+1}\sum_{j=0}^{\infty} j\pi_j^{I+1}(x,t;\Lambda(x,t;\tilde{\Lambda})).$$

Since for $(x,t) \in \mathcal{D}_a$ the value Λ such that $G(x,t;\Lambda) = 0$ exists and is unique, the value $\tilde{\Lambda}$ such that $\tilde{G}(x,t;\tilde{\Lambda}) = 0$ exists and is also unique.

We have that $D_{\tilde{\Lambda}}\tilde{G}(x,t;\tilde{\Lambda})$ equals

$$\begin{pmatrix} -\alpha_0 & \cdots & 0 & -T_0 \\ \vdots & \cdots & \vdots & \vdots \\ 0 & \cdots & -\alpha_I & -T_I \\ -x_0T_0 & \cdots & -x_IT_I & D_\mu \tilde{G}_{I+1}(x,t;\tilde{\Lambda}) \end{pmatrix},$$

where it is only the last entry that must be identified.

We pause to introduce a convention which will be used in the remainder of the paper. Whenever a differential operator of the form D_x precedes a composed function, the derivative is computed via the chain rule for precisely those arguments where a composed dependence on x is made explicit in the notation. Thus in computing $D_\mu \tilde{G}_{I+1}(x,t;\tilde{\Lambda})$ we use (2.29) and calculations in the last section to get

$$\begin{aligned} &D_\mu \tilde{G}_{I+1}(x,t;\tilde{\Lambda}) \\ &= D_\mu G_{I+2}(x,t;\lambda_0,\ldots,\lambda_I,\lambda(x,t;\mu),\mu) \\ &= D_\mu G_{I+2}(x,t;\lambda_0,\ldots,\lambda_I,\lambda_{I+1},\mu) + D_{\lambda_{I+1}} G_{I+2}(x,t;\lambda_0,\ldots,\lambda_I,\lambda_{I+1},\mu)\cdot D_\mu\lambda(x,t;\mu) \\ &= -\sum_{k=0}^{I+1} x_k \sum_{j=0}^{\infty} j^2 \pi_j^k(x,t;\Lambda(x,t;\tilde{\Lambda})) - x_{I+1}T_{I+1}\cdot D_\mu \lambda_{I+1}(x,t;\mu). \end{aligned}$$

It will be useful to express π^{I+1} in the $\mathcal{Q}^a(s)$ notation. We have

$$\pi_j^{I+1} = Q_j^{a+\tau(x,t)}\left(\frac{a+\tau(x,t)}{a+t}(T-t)\right)e^{\lambda_{I+1}-1}e^{j\mu}.$$

Recall that λ_{I+1} is chosen to make this a probability measure. By (2.16),

$$e^{\lambda_{I+1}-1} = \left(1 + \frac{T-t}{a+t}(1-e^\mu)\right)^{a+\tau(x,t)}.$$

Hence using a little algebra we can write

$$\begin{aligned} &\pi_j^{I+1} \\ &= \left(-\frac{T-t}{a+t}\right)^j \binom{-a-\tau(x,t)}{j} \left(1+\frac{T-t}{a+t}\right)^{-a-\tau(x,t)-j} e^{j\mu}\left(1+\frac{T-t}{a+t}(1-e^\mu)\right)^{a+\tau(x,t)} \\ &= \left(-\frac{e^\mu(T-t)}{a+T-e^\mu(T-t)}\right)^j \binom{-a-\tau(x,t)}{j} \left(\frac{a+T}{a+T-e^\mu(T-t)}\right)^{-a-\tau(x,t)-j} \\ &= Q_j^{a+\tau(x,t)}\left(\frac{e^\mu(T-t)(a+\tau(x,t))}{a+T-e^\mu(T-t)}\right). \end{aligned}$$

Again using (2.16),

$$T_{I+1} = \sum_{j=0}^{\infty} j\pi_j^{I+1}(x,t;\Lambda(x,t;\tilde{\Lambda})) = \frac{e^{\mu}(a+\tau(x,t))(T-t)}{a+T-e^{\mu}(T-t)}. \tag{2.30}$$

Recalling the definition

$$\lambda(x,t;\mu) = (a+\tau(x,t))\log\left(\frac{a+T-e^{\mu}(T-t)}{a+t}\right),$$

a direct calculation shows $D_\mu\lambda(x,t;\mu) = -T_{I+1}$, and hence

$$D_\mu\tilde{G}_{I+1}(x,t;\tilde{\Lambda}) = -\sum_{k=0}^{I+1} x_k \sum_{j=0}^{\infty} j^2\pi_j^k(x,t;\Lambda(x,t;\tilde{\Lambda})) + x_{I+1}T_{I+1}^2. \tag{2.31}$$

Since $\alpha_{I+1} = 1$, the eigenvalues of $D_{\tilde{\Lambda}}\tilde{G}(x,t;\tilde{\Lambda})$ are $-\alpha_0, -\alpha_1, \ldots, -\alpha_I$ and $\sum_{k=0}^{I+1}\frac{x_k}{\alpha_k}T_k^2 - \sum_{k=0}^{I+1} x_k \sum_{j=1}^{\infty} j^2\pi_j^k(x,t;\Lambda(x,t;\tilde{\Lambda}))$. By the same argument as was used for $D_\Lambda G(x,t;\Lambda)$ these are all negative, and hence $D_{\tilde{\Lambda}}\tilde{G}(x,t;\tilde{\Lambda})$ is invertible.

We next claim that $D_{\tilde{\Lambda}}\tilde{G}$ is smooth in (x,t). One can check that the only potentially difficult component is $D_\mu\tilde{G}_{I+1}(x,t;\tilde{\Lambda})$, and of this the only nontrivial part is

$$x_{I+1}\left(\sum_{j=0}^{\infty} j^2\pi_j^{I+1}(x,t;\Lambda(x,t;\tilde{\Lambda})) - T_{I+1}^2\right).$$

Using (2.17) and some algebra shows this term equals

$$-x_{I+1}(a+\tau(x,t))\cdot\frac{(t-T)e^{\mu}(a+T)}{[e^{\mu}(t-T)+a+T]^2}.$$

Although $\tau(x,t)$ is not smooth, $x_{I+1}(a+\tau(x,t))$ is always smooth, and thus $D_\mu\tilde{G}$ is smooth in (x,t). Note that the denominator does not vanish since $\eta(t,\mu) > -\infty$.

Therefore $\tilde{G}(\cdot;\cdot)$ is smooth in a neighborhood of $(x,t;\tilde{\Lambda})$. By the implicit function theorem, for any $(x,t)\in\mathcal{D}_a$ there exists a neighborhood $\mathcal{U}\subset\mathbb{R}^{I+1}\times\mathbb{R}$ of (x,t), a neighborhood $\mathcal{V}\subset\mathbb{R}^{I+2}$ of $\tilde{\Lambda}$, and a C^∞ function $g:\mathcal{U}\mapsto\mathcal{V}$, so that $\tilde{\Lambda} = g(x,t)$ and for every $(y,s)\in\mathcal{U}$, $\tilde{G}(y,s;g(y,s)) = 0$. Define

$$\bar{U}(y,s) = H(y,s;g(y,s)).$$

Since $g(y,s)$ is smooth in $\mathcal{U}$, $\bar{U}\in C^\infty(\mathcal{U},\mathbb{R})$, and by Theorem 2.5 $\bar{U}(y,s) = U(y,s)$ for $(y,s)\in\mathcal{U}\cap\mathcal{D}_a$. ∎

The next theorem expresses the derivatives in terms of the Lagrange multipliers.

THEOREM 2.10. *Fix $(x,t)\in D_a$, and let $\tilde{\Lambda}^*$ be the associated Lagrange multiplier. We have*

$$\begin{cases} D_{x_k}\bar{U}(x,t) - D_{x_{k+1}}\bar{U}(x,t) = \lambda_k^* - \lambda_{k+1}^* + \eta^*, \qquad k = 0,1,\ldots,I-1, \\ D_{x_I}\bar{U}(x,t) - D_{x_{I+1}}\bar{U}(x,t) = \lambda_I^* - 1 - (a+I)\eta^* \end{cases}$$

and $D_tU(x,t) = \eta^* - \mu^*$.

Proof. Consider any point $(x,t)\in\mathcal{D}_a$ and let $\tilde{\Lambda}^*$ be the associated Lagrange multiplier. By Theorem 2.9 there exists $\mathcal{U}\subset\mathbb{R}^{I+1}\times\mathbb{R}$ a neighborhood of (x,t),

$\mathcal{V} \subset \mathbb{R}^{I+2}$ a neighborhood of $\tilde{\Lambda}^*$, and a C^∞ function $\tilde{\Lambda} : \mathcal{U} \mapsto \mathcal{V}$ such that $\bar{U}(y,s) \doteq H(y,s;\tilde{\Lambda}(y,s))$ satisfies $\bar{U}(y,s) = U(y,s)$ for any $(y,s) \in \mathcal{U} \cap \mathcal{D}_a$.

Keeping in mind the convention regarding differential operators,

$$\begin{aligned} D_{x_k}\bar{U}(x,t) &= D_{x_k}H(x,t;\tilde{\Lambda}(x,t)) \\ &= D_{x_k}H(x,t;\tilde{\Lambda}^*) + D_{\tilde{\Lambda}}H(x,t;\tilde{\Lambda}^*)D_{x_k}\tilde{\Lambda}(x,t). \end{aligned}$$

Thus in the first line $H(x,t;\tilde{\Lambda}(x,t))$ is considered as the composed function of (x,t) (which by definition is $\bar{U}(x,t)$), and we take derivatives with respect to two arguments and evaluate at (x,t). In the second line, $D_{x_k}H(x,t;\tilde{\Lambda}^*)$ means $H(x,t;\tilde{\Lambda}^*)$ is now a function of the independent variables $(x,t,\tilde{\Lambda}^*)$ and we take derivatives with respect to x_k and then evaluate it at $(x,t;\tilde{\Lambda}^*)$. In all calculations, vectors are interpreted as row vectors.

With the notation established, we can proceed. Note that by definition (2.28),

$$\begin{aligned} &D_{\tilde{\Lambda}}H(x,t;\tilde{\Lambda}) \\ &= D_\Lambda L\left(x,t;\Lambda;\pi\right) D_{\tilde{\Lambda}}\Lambda\left(x,t;\tilde{\Lambda}\right) + D_\pi L\left(x,t;\Lambda;\pi\right) D_{\tilde{\Lambda}}\pi\left(x,t;\Lambda\left(x,t;\tilde{\Lambda}\right)\right). \end{aligned}$$

Since $\Lambda^* = \Lambda(x,t;\tilde{\Lambda}^*)$ and π^* are chosen so that $D_\pi L\left(x,t;\Lambda^*;\pi^*\right) = D_\Lambda L\left(x,t;\Lambda^*;\pi^*\right) = 0$, we have $D_{\tilde{\Lambda}}H(x,t;\tilde{\Lambda}^*) = 0$. This gives

$$D_{x_k}\bar{U}(x,t) = D_{x_k}H(x,t;\tilde{\Lambda}^*), \qquad k = 0,1,\ldots,I+1. \tag{2.32}$$

By the same argument, we have $D_t\bar{U}(x,t) = D_tH(x,t;\tilde{\Lambda}^*)$.

Next, we insert the explicit form of $\pi(x,t;\Lambda)$ from (2.20) into (2.19) to get

$$\begin{aligned} &L(x,t;\Lambda,\pi(x,t;\Lambda)) \\ &= \sum_{k=0}^{I+1} x_k \sum_{j=0}^{\infty} \left[\lambda_k - 1 + j\mu\right] \pi_j^k(x,t;\Lambda) + \sum_{k=0}^{I+1} \lambda_k x_k \left(1 - \sum_{j=0}^{\infty} \pi_j^k(x,t;\Lambda)\right) \\ &\quad + \mu\left(T - t - \sum_{k=0}^{I+1} x_k \sum_{j=0}^{\infty} j\pi_j^k(x,t;\Lambda)\right) \\ &= -\sum_{k=0}^{I} x_k \sum_{j=0}^{\infty} \pi_j^k(x,t;\Lambda) - x_{I+1}\sum_{j=0}^{\infty} \pi_j^{I+1}(x,t;\Lambda) + \sum_{k=0}^{I+1} \lambda_k x_k + \mu\left(T-t\right). \end{aligned}$$

In the definition of $H(x,t;\tilde{\Lambda})$, λ_{I+1} is replaced by $\lambda(x,t;\mu)$ so that automatically $\sum_{j=0}^{\infty} \pi_j^{I+1}(x,t;\Lambda) = 1$. Using $x_{I+1}\lambda(x,t;\mu) - x_{I+1} = x_{I+1}(a+\tau(x,t))\eta(t,\mu)$ from (2.26) and the definition of $\tau(x,t)$,

$$\begin{aligned} H(x,t;\tilde{\Lambda}) &= \sum_{k=0}^{I} x_k\lambda_k + \mu(T-t) - \sum_{k=0}^{I} x_k \sum_{j=0}^{\infty} \pi_j^k(x,t;\tilde{\Lambda}) \\ &\quad + x_{I+1}(a+\tau(x,t))\eta(t,\mu) \\ &= \sum_{k=0}^{I} x_k\lambda_k + \mu(T-t) - \sum_{k=0}^{I} x_k \sum_{j=0}^{\infty} \pi_j^k(x,t;\tilde{\Lambda}) \\ &\quad + a\left(1 - \sum_{k=0}^{I} x_k\right)\eta(t,\mu) + \left(t - \sum_{k=0}^{I} kx_k\right)\eta(t,\mu). \end{aligned}$$

Recall from the explicit expression (2.20) that for $k = 0, 1, \dots, I$, $\pi_j^k(x, t; \Lambda)$ does not depend on x or on λ_{I+1}. Hence the x dependence can be omitted in $\pi_j^k(x, t; \tilde{\Lambda})$ in the last display, and we do so from now on.

By (2.32),

$$\bar{U}_{x_k} = \lambda_k^* - 1 - (a + k)\eta^*, \qquad k = 0, 1, \dots, I,$$

and $\bar{U}_{x_{I+1}} = 0$, where $\eta^* = \eta(t, \mu^*)$. This implies the first claim of the theorem. Similarly

$$\begin{aligned}
&D_t\bar{U}(x,t)\\
&= D_t H(x, t; \tilde{\Lambda}^*)\\
&= \eta^* - \mu^* - D_t\left\{\sum_{k=0}^{I} x_k \sum_{j=0}^{\infty} \pi_j^k(t; \tilde{\Lambda}^*)\right\} + \left\{a\left(1 - \sum_{k=0}^{I} x_k\right) + \left(t - \sum_{k=0}^{I} k x_k\right)\right\} D_t\eta(t, \mu^*)\\
&= \eta^* - \mu^* - D_t\left\{\sum_{k=0}^{I} x_k \sum_{j=0}^{\infty} \pi_j^k(t; \tilde{\Lambda}^*)\right\} + x_{I+1}(a + \tau(x,t))D_t\eta(t, \mu^*).
\end{aligned}$$

We now will verify that

$$-D_t\left\{\sum_{k=0}^{I} x_k \sum_{j=0}^{\infty} \pi_j^k(t; \tilde{\Lambda}^*)\right\} + x_{I+1}(a + \tau(x,t))D_t\eta(t; \mu^*) = 0, \tag{2.33}$$

which implies $D_t U(x,t) = \eta^* - \mu^*$. By the explicit formula for $\pi_j^k(t; \tilde{\Lambda})$ in (2.20) and the definition of $\mathcal{Q}_j^a(s)$, for all $k = 0, 1, \dots, I; j = 0, 1, \dots$

$$D_t\pi_j^k(t; \tilde{\Lambda}) = j\pi_j^k(t; \tilde{\Lambda})\frac{a + T}{(t - T)(a + t)} + \frac{(a + k + j)}{a + t}\pi_j^k(t; \tilde{\Lambda}).$$

A suitable dominating function can be found, and thus by the Lebesgue dominated convergence theorem,

$$\begin{aligned}
&D_t\left\{\sum_{k=0}^{I} x_k \sum_{j=0}^{\infty} \pi_j^k(t; \tilde{\Lambda})\right\}\\
&= \sum_{k=0}^{I} x_k \sum_{j=0}^{\infty}\left[j\pi_j^k(t; \tilde{\Lambda})\frac{a + T}{(t - T)(a + t)} + \frac{(a + k + j)}{a + t}\pi_j^k(t; \tilde{\Lambda})\right].
\end{aligned}$$

Using that $\pi(x, t; \Lambda^*)$ satisfies the constraint (2.11) and some elementary algebra,

$$\begin{aligned}
&D_t\left\{\sum_{k=0}^{I} x_k \sum_{j=0}^{\infty} \pi_j^k(t; \tilde{\Lambda}^*)\right\}\\
&= -\frac{x_{I+1}\sum_{j=0}^{\infty} j\pi_j^{I+1}(x, t; \Lambda^*)}{t - T} - \frac{x_{I+1}(a + \tau(x,t))}{a + t}.
\end{aligned}$$

Equation (2.30) and the definition of η in (2.26) give

$$T_{I+1} = \frac{T - t}{a + t}(a + \tau(x,t))e^{\mu^* - \eta^*}, \tag{2.34}$$

and hence

$$D_t \left\{ \sum_{k=0}^{I} x_k \sum_{j=0}^{\infty} \pi_j^k(t;\tilde{\Lambda}^*) \right\} = -\frac{a+\tau(x,t)}{a+t} x_{I+1} \left(1 - e^{\mu^*-\eta^*}\right). \tag{2.35}$$

The definition of $\eta(t,\mu)$ in (2.26) and $(a+t)e^{\eta} = a+T-e^{\mu}(T-t)$ gives

$$D_t\eta(t,\mu) = \frac{e^{\mu}}{a+T-(T-t)e^{\mu}} - \frac{1}{a+t} = \frac{e^{\mu-\eta}-1}{a+t}.$$

Finally, combining this with (2.35) gives (2.33). ☐

Our final theorem shows that $\bar{U}$ satisfies the HJB equation (2.14) in the classical sense on $\mathcal{U}$. When combined with a standard verification argument as in [4], this will imply $V(x,t) = \bar{U}(x,t) = U(x,t)$ on $\mathcal{D}_a$.

THEOREM 2.11. *$\bar{U}$ satisfies* (2.14) *on $\mathcal{D}_a$.*

Proof. Having derived various expressions for the derivatives of $\bar{U}$ in terms of the Lagrange multipliers in Theorem 2.10, to show that $\bar{U}(x,t)$ satisfies the PDE (2.14) it remains to show

$$e^{-\mu^*+\eta^*} = \sum_{k=0}^{I-1} x_k \frac{a+k}{a+t} e^{\lambda_k^*-\lambda_{k+1}^*+\eta^*} + x_I \frac{a+I}{a+t} e^{\lambda_I^*-1-(a+I)\eta^*} + x_{I+1}\frac{a+\tau(x,t)}{a+t}. \tag{2.36}$$

Recall from (2.20) that for $k=0,1,\ldots,I$ and $j\geq 0$,

$$\pi_j^{*k} = \mathcal{Q}_j^{a+k}\left(\frac{a+k}{a+t}(T-t)\right) e^{\lambda_k^*-1+j\mu^*-\ell_{k+j}}.$$

Using the definition of $\mathcal{Q}_j^{a+k}$, for $k=0,1,\ldots,I-1$,

$$\pi_j^{*k+1} = \frac{(j+1)(a+t)}{(a+k)(T-t)} \pi_{j+1}^{*k} e^{\lambda_{k+1}^*-\lambda_k^*} e^{-\mu^*}.$$

Now sum both sides from $j=0$ to ∞ and use the fact that $\sum_{j=0}^{\infty}\pi_j^{*k}=1$ to get

$$e^{\lambda_k^*-\lambda_{k+1}^*} = e^{-\mu^*}\cdot \frac{(a+t)\sum_{j=1}^{\infty} j\pi_j^{*k}}{(T-t)(a+k)} \quad \text{for } k=0,1,\ldots,I-1. \tag{2.37}$$

Inserting (2.37) into (2.36), a little algebra shows that satisfaction of the PDE is equivalent to

$$\begin{aligned}(T-t) &= \sum_{k=0}^{I-1} x_k \sum_{j=0}^{\infty} j\pi_j^{*k} + x_I \frac{a+I}{a+t}(T-t)e^{\lambda_I^*-1-(a+I+1)\eta^*+\mu^*} \\ &\quad + x_{I+1}\frac{a+\tau(x,t)}{a+t}(T-t)e^{-\eta^*+\mu^*}.\end{aligned} \tag{2.38}$$

Since $\pi_j^{*I} = \mathcal{Q}_j^{a+I}\left(\frac{a+I}{a+t}(T-t)\right)e^{\lambda_I^*-1+j\mu^*}$ for $j\geq 1$, by (2.16),

$$\begin{aligned}\sum_{j=1}^{\infty} j\pi_j^{*I} &= \sum_{j=1}^{\infty} j\mathcal{Q}_j^{a+I}\left(\frac{a+I}{a+t}(T-t)\right)e^{\lambda_I^*-1+j\mu^*} \\ &= \frac{T-t}{a+t}(a+I)e^{\lambda_I^*-1-\eta^*(a+I+1)+\mu^*},\end{aligned} \tag{2.39}$$

where the last equality uses the definition of η in (2.26).

Inserting (2.34) and (2.39) into $\sum_{k=0}^{I+1} x_k \sum_{j=0}^{\infty} j\pi_j^k = T - t$, we then have

$$(T-t) = \sum_{k=0}^{I-1} x_k \sum_{j=0}^{\infty} j\pi_j^{*k} + x_I \frac{a+I}{a+t}(T-t)e^{\lambda_I^* - 1 - \eta^*(a+I+1)+\mu^*} + x_{I+1}\frac{a+\tau(x,t)}{a+t}(T-t)e^{-\eta^*+\mu^*}.$$

We have thus verified (2.38), which completes the proof that $\bar{U}$ satisfies (2.14). ∎

2.4. Minimizing trajectories. The minimizing trajectories associated with the calculus of variations problem have important qualitative and computational uses. Perhaps the most important is that they identify the most likely way a rare event will occur [5].

Identification of the minimizing trajectories in the MB case was done in [2] using the Euler–Lagrange equations, a system of nonlinear ODEs. The solutions to these equation are called extremals, and in general being an extremal is neither a necessary nor sufficient condition for minimality. In [2], a direct but detailed argument using Lagrange multiplier techniques was used to show the extremals were indeed minimizers. Here we take a different tack. We start in [2] with a two-parameter family of solutions to the Euler–Lagrange equations that identify the extremals. The two parameters are themselves characterized as the solution to a pair of nonlinear algebraic equations. These parameters and the form of the extremals suggest values for the Lagrange multipliers in the explicit representation (2.15), and indeed it is shown that characterizing equations $G(x,t;\Lambda) = 0$ for the unique Lagrange multipliers are satisfied. Having identified the minimal cost, all that remains is to show that the cost along the extremal is the same as this minimal value. This is done by explicitly evaluating an integral.

The main goal of this section is to argue that the extremals are minimizers and exhibit the relation between the two parameters used to identify the extremals and the Lagrange multipliers used in the formula for the minimal cost. Not all details will be given, and to simplify the presentation only the initial condition with $t = 0$ and all cells empty is considered. The statement of the case of general initial conditions is exactly analogous to Theorem 2.8 in [2], and any details that are omitted are similar to ones appearing in [2]. In addition, we present only the case $a \in (0,\infty)$. In this case it is simpler to work with an infinite dimensional version of the extremals. This is analogous to what is called the *exponential* case in [2]. The arguments for the cases $a = \infty$ and $a < 0$ are analogous to that of $a \in (0,\infty)$. As noted above, the case $a = \infty$ has already been considered in [2]. In the case $a < 0$, the extremals satisfy $\varphi_I = 0$. Because of this the arguments are somewhat simpler than in the case of $a > 0$, and it is analogous to the *polynomial* case of [2].

In the case of the empty initial condition the extremal can be identified as follows. Let $y \in \mathcal{S}_I$ be given. Then a family of solutions to the Euler–Lagrange equations for the problem of minimizing the cost subject to this terminal condition are

$$\varphi_0(t) = C\mathcal{Q}_0^a(\rho t) + \sum_{k=0}^{I} \left(y_k - C\mathcal{Q}_k^a(\rho T)\right)\left(1 - \frac{t}{T}\right)^k,$$

$$\varphi_i(t) = \frac{t^i}{i!}(-1)^i \varphi_0^{(i)}(t), \quad 1 \le i \le I,$$

$$\varphi_{I+}(t) = 1 - \sum_{i=0}^{I} \varphi_i(t).$$

This is exactly analogous to the form found in [2] for the special case of $a = \infty$, with the Poisson distribution $\mathcal{P}(t)$ in that case replaced by the family of probability distributions $\mathcal{Q}^a(t)$. It is useful to extend the definition of $\varphi_i(t)$ to $\varphi_i(t) = C\mathcal{Q}_i^a(\rho t)$ for $i > I$, while maintaining the distinction between $\varphi_{I+}(t)$ and $\varphi_I(t)$.

The parameters $\rho > 0$ and $C \geq 0$ are chosen so that the measure corresponding to $\varphi_i(T)$ is a probability measure, and moreover, one for which the number of balls per cell at time T equals T. Specifically, ρ is chosen so that

$$\frac{\rho T - \sum_{i=0}^{I} i\mathcal{Q}_i^a(\rho T)}{1 - \sum_{i=0}^{I} \mathcal{Q}_i^a(\rho T)} = \frac{T - \sum_{i=0}^{I} iy_i}{1 - \sum_{i=0}^{I} y_i}$$

holds, and then

$$C \doteq \frac{1 - \sum_{i=0}^{I} y_i}{1 - \sum_{i=0}^{I} \mathcal{Q}_i^a(\rho T)} = \frac{T - \sum_{i=0}^{I} iy_i}{\rho T - \sum_{i=0}^{I} i\mathcal{Q}_i^a(\rho T)}.$$

Solutions to these equations exist for $\rho \in (0, \infty)$ and $C \in [0, \infty)$ and are unique.

To show that this is indeed a minimizing trajectory we relate the constants ρ and C to the Lagrange multipliers appearing in the finite dimensional representation (2.15). Recall that the minimizer to this problem takes the form

$$\pi_j^0(1, 0; \Lambda) = \mathcal{Q}_j^a(T)e^{\lambda_0 - 1 + j\mu - \ell_j},$$

with $\ell_j = 0$ if $j > I$. Using the form of the minimizing trajectory, at time $t = T$,

$$\begin{aligned} &\varphi_j(T) = y_j, && 0 \leq j \leq I, \\ &\varphi_j(T) = C\mathcal{Q}_j^a(\rho T), && I < j. \end{aligned}$$

Thus for all $j > I$ we will want

$$\mathcal{Q}_j^a(T)e^{\lambda_0 - 1 + j\mu} = C\mathcal{Q}_j^a(\rho T).$$

Since

$$\frac{\mathcal{Q}_j^a(\rho T)}{\mathcal{Q}_j^a(T)} = \frac{\left(\frac{-\rho T}{a+\rho T}\right)^j \left(\frac{a+\rho T}{a}\right)^{-a}}{\left(\frac{-T}{a+T}\right)^j \left(\frac{a+T}{a}\right)^{-a}} = \rho^j \left(\frac{a+T}{a+\rho T}\right)^j \left(\frac{a+T}{a+\rho T}\right)^a,$$

this suggests

$$\begin{aligned} \mu &= \log \rho + \log\left(\frac{a+T}{a+\rho T}\right), \\ \lambda_0 - 1 &= a \log\left(\frac{a+T}{a+\rho T}\right) + \log C, \end{aligned}$$

and

$$\ell_k = -\log y_k + \log \mathcal{Q}_k^a(\rho T) + \log C$$

when $k \leq I$. Hence the minimizing trajectory for a problem with a finite terminal cost ℓ will terminate at a point y with each $y_k > 0$, an assumption we make for the rest of this section. The argument when one or more $y_k = 0$ can be handled by a limiting argument. We remark in passing that similar considerations allow one to explicitly identify the Lagrange multipliers for all initial conditions (x,t) that lie on the extremal in terms of C, ρ, and the values y_k.

To show that $\pi^0(1,0;\Lambda)$ is the minimizing probability measure in (2.15) the constraints (2.11) must be demonstrated. One constraint is that $\pi^0_j(1,0;\Lambda)$ be a probability measure. Since the definitions of the Lagrange multipliers enforce $\pi^0_j(1,0;\Lambda) = \varphi_j(T)$, this follows from $\sum_{i=0}^I y_i = 1 - C + C\sum_{i=0}^I \mathcal{Q}^a_i(\rho T)$ and

$$\sum_{j=0}^{\infty} \varphi_j(T) = \sum_{j=0}^{I} y_j + \sum_{j=I+1}^{\infty} C\mathcal{Q}^a_j(\rho T) = 1 - C + C = 1.$$

The only other constraint to check is the conservation condition:

$$\begin{aligned}
\sum_{j=0}^{\infty} j\pi^0_j(1,0;\Lambda) &= \sum_{j=0}^{\infty} j\mathcal{Q}^a_j(T)e^{\lambda_0 - 1 + j\mu - \ell_j} \\
&= \sum_{j=0}^{\infty} jC\mathcal{Q}^a_j(\rho T) + \sum_{j=0}^{I} jy_j - \sum_{j=0}^{I} jC\mathcal{Q}^a_j(\rho T) \\
&= C\rho T - \sum_{j=0}^{I} jC\mathcal{Q}^a_j(\rho T) + \sum_{j=0}^{I} jy_j \\
&= T - \sum_{i=0}^{I} iy_i + \sum_{j=0}^{I} jy_j \\
&= T,
\end{aligned}$$

where the equations characterizing C and ρ are used for the fourth equality.

We have identified the optimal measure for the terminal cost ℓ. To complete the argument that φ is a minimizer we need only show that the cost along this trajectory equals

$$\begin{aligned}
&R\left(\pi^0(1,0;\Lambda) \,\|\, \mathcal{Q}^a(T)\right) \\
&= \sum_{j=0}^{\infty} \pi^0_j(1,0;\Lambda) \log\left(\frac{\pi^0_j(1,0;\Lambda)}{\mathcal{Q}^a_j(T)}\right) \\
&= \sum_{j=0}^{\infty} \mathcal{Q}^a_j(T)e^{\lambda_0 - 1 + j\mu - \ell_j} \log\left(e^{\lambda_0 - 1 + j\mu - \ell_j}\right) \\
&= \sum_{j=0}^{I} y_j \log\left(\frac{y_j}{\mathcal{Q}^a_j(T)}\right) + \sum_{j=I+1}^{\infty} C\mathcal{Q}^a_j(\rho T) \log\left(\frac{C\mathcal{Q}^a_j(\rho T)}{\mathcal{Q}^a_j(T)}\right) \\
&= \sum_{j=0}^{\infty} \varphi_j(T) \log\left(\frac{\varphi_j(T)}{\mathcal{Q}^a_j(T)}\right) \\
&= \sum_{j=0}^{\infty} \varphi_j(T) \log\left(\varphi_j(T)\right) - \sum_{j=0}^{\infty} \varphi_j(T) \log\left(\mathcal{Q}^a_j(T)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=0}^{\infty} \varphi_j(T) \log\left(\varphi_j(T)\right) \\
&\quad - \sum_{j=0}^{\infty} \varphi_j(T) \left[\log\left(\frac{T}{a}\right)^j + \log\left(\frac{\prod_{k=0}^{j-1}(a+k)}{j!}\right) + \log\left(\frac{a+T}{a}\right)^{-a-j} \right] \\
&= \sum_{j=0}^{\infty} \varphi_j(T) \log\left(\varphi_j(T)\right) - T \log\left(\frac{T}{a}\right) + (a+T)\log\left(\frac{a+T}{a}\right) \\
&\quad - \sum_{j=0}^{\infty} \varphi_j(T) \log\left(\frac{\prod_{k=0}^{j-1}(a+k)}{j!}\right).
\end{aligned}
$$

The notion of a completely monotone function is useful here. Although it is clear from the construction of the Lagrange multipliers that $\varphi_j(T)$ is a probability measure, the same cannot be said yet for $\varphi_j(t)$. A monotone function γ is completely monotone on $[0,T]$ if it is infinitely differentiable on $[0,T]$ and if for all $t \in [0,T]$ and $i \geq 0$,

$$(-1)^i \gamma^{(i)}(t) \geq 0.$$

The same argument as [2, p. 2794] shows that $\varphi_0(t)$ is completely monotone on $[0,T]$, and hence for all $t \in [0,T]$ and $i \geq 0$, $\varphi_i(t) \geq 0$. From the explicit formula for $\varphi_0(t)$ we actually have $(-1)^i \varphi_0^{(i)}(t) > 0$ for $t \in (0,T)$. It is also easy to show that for all $t \in [0,T]$ and $i \geq 0$, $\varphi_i(t)$ can be interpreted as the ith term in the Taylor series expansion of $\varphi_0(0)$ about t, and so for each t, $\{\varphi_i(t), i = 0, 1, \ldots\}$ is a probability measure on $\{0, 1, \ldots\}$.

To evaluate the cost

$$\int_0^T R\left(\theta \,\|\rho(t,\varphi(t))\right) dt,$$

it is convenient to work, as in [2], with the cumulative occupancy functions

$$\psi_j(t) = \sum_{i=0}^{j} \varphi_i(t).$$

The dynamics then take the form $\psi_j^{(1)}(t) = -\theta_j(t)$, and so with the convention $\psi_{-1}(t) = 0$ the cost can be expressed as

$$
\begin{aligned}
\int_0^T \Bigg[\sum_{i=0}^{I} -\psi_i^{(1)}(t) \log\left(\frac{-\psi_i^{(1)}(t)}{\frac{a+i}{a+t}\left(\psi_i(t) - \psi_{i-1}(t)\right)}\right) + \left(1 + \psi_0^{(1)}(t) + \cdots + \psi_I^{(1)}(t)\right) \\
\times \log\left(\frac{1 + \psi_0^{(1)}(t) + \cdots + \psi_I^{(1)}(t)}{\frac{a}{a+t}\left(1 - \psi_I(t)\right) + \frac{1}{a+t}\sum_{k=I+1}^{\infty} k\left(\psi_k(t) - \psi_{k-1}(t)\right)}\right) \Bigg] dt.
\end{aligned}
$$

We have

$$
\begin{aligned}
-\psi_i^{(1)}(t) &= -\sum_{k=0}^{i} \varphi_i^{(1)}(t) \\
&= -\sum_{k=0}^{i} \frac{(-t)^k}{k!} \psi_0^{(k+1)}(t) + \sum_{k=1}^{i} \frac{(-t)^{k-1}}{(k-1)!} \psi_0^{(k)}(t) \\
&= -\frac{(-t)^i}{i!} \psi_0^{(i+1)}(t),
\end{aligned}
$$

and so

$$\frac{-\psi_i^{(1)}(t)}{\varphi_i(t)} = \frac{-\frac{(-t)^i}{i!}\psi_0^{(i+1)}(t)}{\frac{(-t)^i}{i!}\psi_0^{(i)}(t)} = \frac{-\psi_0^{(i+1)}(t)}{\psi_0^{(i)}(t)}.$$

Since for $i > I$,

$$\begin{aligned}
\frac{-\psi_i^{(1)}(t)}{\frac{a+i}{a+t}\left(\psi_i(t) - \psi_{i-1}(t)\right)} &= \frac{a+t}{a+i}\frac{-\varphi_0^{(i+1)}(t)}{\varphi_0^{(i)}(t)} \\
&= \frac{a+t}{a+i}\frac{\mathcal{Q}_{i+1}^a(\rho t)(i+1)}{\mathcal{Q}_i^a(\rho t)t} \\
&= \frac{a+t}{a+i}\frac{(-\rho t)(-a-i)}{(a+\rho t)t} \\
&= \frac{(a+t)\rho}{a+\rho t}
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1+\psi_0^{(1)}(t)+\cdots+\psi_I^{(1)}(t)}{\frac{a}{a+t}\left(1-\psi_I(t)\right)+\frac{1}{a+t}\sum_{k=I+1}^{\infty}k\left(\psi_k(t)-\psi_{k-1}(t)\right)} \\
&= \frac{\sum_{i=I+1}^{\infty}-\frac{(-t)^i}{i!}\varphi_0^{(i+1)}(t)}{\frac{a}{a+t}\left(\sum_{i=I+1}^{\infty}\frac{(-t)^i}{i!}\varphi_0^{(i)}(t)\right)+\frac{1}{a+t}\sum_{i=I+1}^{\infty}i\frac{(-t)^i}{i!}\varphi_0^{(i)}(t)} \\
&= \frac{\sum_{i=I+1}^{\infty}-\frac{(-t)^i}{i!}\varphi_0^{(i+1)}(t)}{\frac{a+i}{a+t}\left(\sum_{i=I+1}^{\infty}\frac{(-t)^i}{i!}\varphi_0^{(i)}(t)\right)},
\end{aligned}$$

we can write the integral as

$$\int_0^T\sum_{i=0}^{\infty}-\psi_i^{(1)}(t)\log\left(\frac{-\psi_i^{(1)}(t)}{\frac{a+i}{a+t}\left(\psi_i(t)-\psi_{i-1}(t)\right)}\right)dt.$$

In several places below we will need the existence of a dominating function to justify the interchange of summation and integration. A suitable function can be found using the same calculations as those used to establish (2.16). Also, this dominating function will work only on closed subintervals of $(0, T)$, and so a careful argument will first evaluate the integral on $[\varepsilon, T-\varepsilon]$ and then use monotone convergence to take the limit $\varepsilon \downarrow 0$.

We evaluate the integral by the following calculation, each line of which is explained below:

$$
\begin{aligned}
&\int_0^T \sum_{i=0}^{\infty} -\psi_i^{(1)}(t) \log\left(\frac{a+t}{a+i}\frac{-\psi_0^{(i+1)}(t)}{\psi_0^{(i)}(t)}\right) dt \\
&= \int_0^T \sum_{i=0}^{\infty} \left(-\psi_i^{(1)}(t)\log\left|\psi_0^{(i+1)}(t)\right| + \psi_i^{(1)}(t)\log\left|\psi_0^{(i)}(t)\right| - \psi_i^{(1)}(t)\log(a+i)\right) dt \\
&\quad + \int_0^T \left(\sum_{i=0}^{\infty} -\psi_i^{(1)}(t)\right)\log(a+t)\, dt \\
&= \int_0^T \sum_{i=0}^{\infty}\left[\left(\psi_i^{(1)}(t) - \psi_{i-1}^{(1)}(t)\right)\log\left|\psi_0^{(i)}(t)\right| - \psi_i^{(1)}(t)\log(a+i)\right] dt \\
&\quad + \int_0^T \log(a+t)\, dt \\
&= \sum_{i=0}^{\infty}\left(\left[\varphi_i(t)\log\left|\psi_0^{(i)}(t)\right|\right]_0^T + \int_0^T\left(\sum_{i=0}^{\infty} -\psi_i^{(1)}(t)\right) dt + \sum_{i=0}^{\infty}\left[\psi_i(t)\right]_0^T \log(a+i)\right) \\
&\quad + (a+T)\log(a+T) - (a+T) - a\log a + a \\
&= \sum_{i=0}^{\infty}\left(\left[\varphi_i(T)\log\left|\varphi_0^{(i)}(T)\right| - \varphi_i(0)\log\left|\varphi_0^{(i)}(0)\right|\right]_0^T + \log(a+i)\left(\sum_{k=0}^{i}\varphi_i(T) - \sum_{k=0}^{i}\varphi_i(0)\right)\right) \\
&\quad + T + (a+T)\log(a+T) - (a+T) - a\log a + a \\
&= \sum_{i=0}^{\infty}\left(\varphi_i(T)\log\left(\varphi_i(T)i!/T^i\right) - \log(a+i)\left(\sum_{k=i+1}^{\infty}\varphi_k(T)\right)\right) \\
&\quad + (a+T)\log(a+T) - a\log a \\
&= \sum_{i=0}^{\infty}\left(\varphi_i(T)\log\varphi_i(T) \;\; + \sum_{i=0}^{\infty}\varphi_i(T)\log(i!) - \varphi_i(T)\log\left(\prod_{k=0}^{i-1}(a+k)\right)\right) \\
&\quad - T\log T + (a+T)\log(a+T) - a\log a \\
&= \sum_{i=0}^{\infty}\varphi_i(T)\log\varphi_i(T) + \sum_{i=0}^{\infty}\varphi_i(T)\log\left(\frac{i!}{\prod_{k=0}^{i-1}(a+k)}\right) \\
&\quad - T\log T + (a+T)\log(a+T) - a\log a.
\end{aligned}
$$

The first line separates the $a+t$ term in the logarithm. The second equality uses the convention $\psi_{-1}(t) = 0$ and that $\sum_{i=0}^{\infty} -\psi_i^{(1)}(t) = 1$ (all balls go into some cell). The third line uses integration by parts and the fourth uses the definition of the cumulative occupancies. The fifth line uses the definition of φ_i in terms of derivatives of φ_0, and the sixth equality uses summation by parts. Since

$$
-T\log\left(\frac{T}{a}\right) + (a+T)\log\left(\frac{a+T}{a}\right) = -T\log T + (a+T)\log(a+T) - a\log a,
$$

the cost along this trajectory equals the minimum, and the argument is complete.

3. Allocation models with balls of different color. In this section we extend the techniques to the case of allocation models where the balls are of more than one type. To keep the notation simple, we actually consider just two colors, the extension to the more general case being straightforward. Another simplification is that we consider only MB statistics. The interested reader can combine the methods from this section and the last to treat more general statistical models.

3.1. Coloration process. We construct an allocation model with colored balls as follows. Balls are thrown into one of n cells sequentially. The throwing process is modeled by a collection of iid random variables, each uniformly distributed on the set $\{1,\dots,n\}$, with each value of the set corresponding to a cell. There is also a coloration process. At each discrete time a ball is assigned color $Y_l^n \in \{1,2\}$ and then placed into the cell determined by the throwing process. These two processes are independent.

The occupancy process in this case is defined as follows. The natural state space is

$$S_{I,J} \doteq \left\{ x \in \mathbb{R}^{I+2} \times \mathbb{R}^{J+2} : x_{i,j} \geq 0, 0 \leq i \leq I+1, 0 \leq j \leq J+1, \sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j} = 1 \right\}.$$

If $i \in \{0,\dots,I\}$ and $j \in \{0,\dots,J\}$, then $X_{i,j}^n(l/n)$ is the fraction of cells containing exactly i color-1 and j color-2 balls when l balls have been thrown. In an analogous fashion $X_{I+,j}^n(l/n), X_{i,J+}^n(l/n)$ and $X_{I+,J+}^n(l/n)$ are defined. By definition $X_{0,0}^n(0) \doteq 1$, and $X_{i,j}^n(0) \doteq 0$ for all other values of (i,j).

To describe the large deviation asymptotics of these allocation processes we must specify those of the coloration processes. Cumulative coloration processes $\{r^n, n \in \mathbb{N}\}$ are defined for $t = l/n$ by

$$r_1^n\left(\frac{l}{n}\right) \doteq \frac{1}{n}\sum_{k=1}^{l} 1_{\{Y_k^n = 1\}}, \quad r_2^n\left(\frac{l}{n}\right) \doteq \frac{1}{n}\sum_{k=1}^{l} 1_{\{Y_k^n = 2\}}.$$

We will assume that these processes satisfy a large deviations principle with a rate function of the form $J(\phi) = \int_0^T c(\dot{\phi}(s))ds$. Thus $c(\dot{\phi}(s))$ is a measure of the local (in time) log likelihood that a fraction $\dot{\phi}_i(s)$ of the balls are color i. A mild technical assumption that is needed to prove a large deviations result for the occupancy process is that $c(a) = 0$ for some point a with $a_i > 0, i = 1,2$. Since c is a rate function, there is at least one probability vector a at which $c(a) = 0$. The assumption that this occurs at a point where both components are positive is very mild and means simply that the LLN limit cannot concentrate exclusively on one color.

Examples of coloration processes which satisfy these properties are *deterministic, iid*, and *Markovian*. In the iid case colors are selected by an iid sequence of random variables. In the Markov case the color is chosen by a finite state ergodic Markov chain. The so-called deterministic case seeks to achieve a deterministic fraction a_k of color k, with $a_k \in (0,1)$. This can be done as follows. If N_{l-1}^k balls of color k have been thrown in the first $l-1$ throws (with $N_{l-1}^1 + N_{l-1}^2 = l-1$), and if $N_{l-1}^1/n \leq a_1 l/n$, then we color the lth ball 1, and otherwise color it 2.

The specific form for c in all these cases is spelled out in [3]. For reasons to be explained below, the focus in this paper will be on the iid and deterministic cases, where $c(\rho)$ equals $R(\rho\,\|a)$ and $\infty \cdot 1_{\{a\}^c}(\rho)$, respectively. Under a suitable restriction needed to ensure convexity that is also described below, the same methods can be applied to the Markovian case as well.

3.2. Variational problem and PDE. For this problem the feasible domain is

$$\mathcal{D} \doteq \left\{ (x,t) \in S_{I,J} \times [0,T) : \sum_{i=0}^{I} x_{i,J+1} + \sum_{j=0}^{J} x_{I+1,j} + x_{I+1,J+1} > 0 \text{ and} \right.$$
$$\left. t > \sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j} \text{ or } \sum_{i=0}^{I} x_{i,J+1} + \sum_{j=0}^{J} x_{I+1,j} + x_{I+1,J+1} = 0 \text{ and } t = \sum_{i=0}^{I}\sum_{j=0}^{J} x_{i,j} \right\}.$$

We next describe the large deviations variational problem. Recall that $S_{I,J}$ are the set of all probability measures on $\{0, I+1\} \times \{0, J+1\}$, which can also be interpreted as real $(I+2) \times (J+2)$ matrices. For $\alpha \in S_{I,J}$ define the linear maps

$$M^1_{i,j}[\alpha] = \alpha_{i-1,j} 1_{\{i \geq 1\}} - \alpha_{i,j} 1_{\{i \leq I\}}, M^2_{i,j}[\alpha] = \alpha_{i,j-1} 1_{\{j \geq 1\}} - \alpha_{i,j} 1_{\{j \leq J\}}.$$

The rate function for the coloration over an interval $[t,T]$ is assumed to be of the form $\int_t^T c(\rho) ds$, where $\rho(s) = (\rho_1(s), \rho_2(s))$ are the colored fractions at time s. The local (in time) and total coloration fractions satisfy $q_k = \int_t^T \rho_k(s) ds/[T-t]$, and for a trajectory of the form $(\rho_1(s), \rho_2(s)) = (q_1, q_2)$, the cost is of course $[T-t]c(q)$. The rate function for the occupancy process on path space is then

$$\mathcal{I}(x,t;\varphi) = \inf_{\theta,\rho} \int_t^T \left[\rho_1 R(\theta^1 \,\|\varphi\,) + \rho_2 R(\theta^2 \,\|\varphi\,) + c(\rho) \right] ds,$$

where the infimum is over all θ, ρ such that

$$\varphi(u) - \varphi(t) = \int_t^u \left(\rho_1 M^1[\theta^1] + \rho_2 M^2[\theta^2] \right) ds.$$

For a terminal cost F we consider

$$V(x,t) = \inf_{\varphi \in C([t,T], S_{I,J}), \varphi(t) = x} \left\{ \mathcal{I}(x,t;\varphi) + F(\varphi(T)) \right\}.$$

Then V should be a weak sense solution to

$$W_t + H(W_x, x, t) = 0$$

and the terminal condition, where

$$H(p,x,t) = \inf_{\rho,\theta^1,\theta^2} \left[\left\langle p, \rho_1 M^1[\theta^1] + \rho_2 M^2[\theta^2] \right\rangle + \rho_1 R(\theta^1 \,\|x\,) + \rho_2 R(\theta^2 \,\|x\,) + c(\rho) \right].$$

If $b(\gamma)$ is the Legendre transform $b(\gamma) = \sup_\rho [\langle \gamma, \rho \rangle - c(\rho)]$, then we can also write

$$\begin{aligned} H(p,x,t) &= -\sup_\rho \left[-\sum_{m=1,2} \rho_m \left(\inf_{\theta^m} [\langle p, M^m[\theta^m] \rangle + R(\theta^m \,\|x\,)] \right) - c(\rho) \right] \\ &= -b\left(-\inf_{\theta^1} \left[\left\langle p, M^1[\theta^1] \right\rangle + R(\theta^1 \,\|x\,) \right], -\inf_{\theta^2} \left[\left\langle p, M^2[\theta^2] \right\rangle + R(\theta^2 \,\|x\,) \right] \right). \end{aligned}$$

The variational formula for exponential integrals in terms of relative entropy [1, Proposition 1.4.2] asserts that

$$\begin{aligned}
\inf_{\theta^1}\left[\left\langle p, M^1\left[\theta^1\right]\right\rangle + R(\theta^1\,\|x\,)\right] &= \inf_{\theta^1}\left[\sum_{i,j,i\geq 1} p_{i,j}\theta^1_{i-1,j} - \sum_{i,j,i\leq I} p_{i,j}\theta^1_{i,j} + R(\theta^1\,\|x\,)\right]\\
&= \inf_{\theta^1}\left[\sum_{i,j,i\leq I}(p_{i+1,j}-p_{i,j})\theta^1_{i,j} + R(\theta^1\,\|x\,)\right]\\
&= -\log\left(\sum_{i,j,i\leq I} e^{-(p_{i+1,j}-p_{i,j})}x_{i,j} + \sum_{i,j,i=I+1} x_{i,j}\right).
\end{aligned}$$

Using the analogous formula for $m=2$, one obtains

$$\begin{aligned}
H(p,x,t) = -b\Bigg[&\log\left(\sum_{i,j,i\leq I} e^{-(p_{i+1,j}-p_{i,j})}x_{i,j} + \sum_{i,j,i=I+1} x_{i,j}\right),\\
&\log\left(\sum_{i,j,j\leq J} e^{-(p_{i,j+1}-p_{i,j})}x_{i,j} + \sum_{i,j,j=J+1} x_{i,j}\right)\Bigg].
\end{aligned} \tag{3.1}$$

3.3. Explicit solution. Let $\pi_{i,j}(r_1,r_2)$ denote the probability of throwing r_m additional balls of color m, $m=1,2$, into cells of category (i,j), and let $q=(q_1,q_2)$ be the fraction of balls of colors $(1,2)$. For $x\in S_{I,J}$, we say that $(\pi,q)\in\mathcal{F}(x,t;y,T)$ if for all i,j,

$$x_{i,j}\sum_{r_1,r_2=0}^{\infty}\pi_{i,j}(r_1,r_2) = x_{i,j}, \quad \sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j}\sum_{r_1,r_2=0}^{\infty} r_m\pi_{i,j}(r_1,r_2) = q_m(T-t)$$

for $m=1,2$ and

$$\begin{aligned}
y_{k,l} &= \sum_{i=0}^{k}\sum_{j=0}^{l} x_{i,j}\pi_{i,j}(k-i,l-j),\\
y_{I+1,l} &= \sum_{r=0}^{\infty}\sum_{i=0}^{I+1}\sum_{j=0}^{l} x_{i,j}\pi_{i,j}(I+1-i+r,l-j),\\
y_{k,J+1} &= \sum_{r=0}^{\infty}\sum_{i=0}^{k}\sum_{j=0}^{J+1} x_{i,j}\pi_{i,j}(k-i,J+1-j+r),\\
y_{l+1,J+1} &= \sum_{s=0}^{\infty}\sum_{r=0}^{\infty}\sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j}\pi_{i,j}(I+1-i+s,J+1-j+r).
\end{aligned}$$

We also denote y by $x\times\pi$. If the coloration turns out to be (q_1,q_2), then there are $q_m(T-t)n$ balls of color m thrown, and the LLN limit for the empirical fraction of cells of category (i,j) is $\mathcal{P}_i(q_1(T-t))\mathcal{P}_j(q_2(T-t))$.

The same sort of argument as in section 2.2 then suggests that the explicit form for the solution to the variational problem should be

$$\min_{(\pi,q)\in\mathcal{F}(x,t;y,T)}\left\{\sum_{i=0}^{I+1}\sum_{j=0}^{J+1}x_{i,j}R\left(\pi_{i,j}\left\|\mathcal{P}(q_1(T-t))\times\mathcal{P}(q_2(T-t))\right.\right)+(T-t)c(q)\right\}.$$

However, an interesting feature of the case with color is that the quantity being minimized in this formula is not always convex. In a previous paper [3], a useful assumption that guaranteed the convexity of the large deviation rate on path space was that $c(\rho)+h(\rho)$ be convex, where $h(\rho)$ is the entropy function $h(\rho)\doteq-\rho_1\log\rho_1-\rho_2\log\rho_2$. We will show that this same condition, not surprisingly, gives convexity here as well. Let a be the point with $a_i>0, i=1,2$, for which $c(a)=0$. Then we can write, under the constraint that relates $\pi_{i,j}$ and q_m,

$$\begin{aligned}
&\sum_{i=0}^{I+1}\sum_{j=0}^{J+1}x_{i,j}R\left(\pi_{i,j}\left\|\mathcal{P}(q_1(T-t))\times\mathcal{P}(q_2(T-t))\right.\right)+(T-t)c(q)\\
&=\sum_{i=0}^{I+1}\sum_{j=0}^{J+1}x_{i,j}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\pi_{i,j}(k,l)\log\left(\frac{\pi_{i,j}(k,l)}{\mathcal{P}_k(a_1(T-t))\mathcal{P}_l(a_2(T-t))}\right)\\
&\quad+\sum_{i=0}^{I+1}\sum_{j=0}^{J+1}x_{i,j}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\pi_{i,j}(k,l)\left[(T-t)c(q)-\log\left(\left[\frac{q_1}{a_1}\right]^k\left[\frac{q_2}{a_2}\right]^l\right)\right]\\
&=\sum_{i=0}^{I+1}\sum_{j=0}^{J+1}x_{i,j}R\left(\pi_{i,j}\left\|\mathcal{P}(a_1(T-t))\times\mathcal{P}(a_2(T-t))\right.\right)+(T-t)\left[c(q)-R(q\,\|a)\right].
\end{aligned}$$

The mapping $q\to[c(q)-R(q\,\|a)]$ is convex if and only if $c(q)+h(q)$ is convex, and so convexity of $c(q)+h(q)$ is sufficient for the minimization problem to be convex in $(\pi_{i,j},q_m)$. Note that in the deterministic case this condition holds with strict convexity, and that in the iid case $c(q)+h(q)$ is convex but never strictly convex (it is in fact always linear in q). Hence this is in a certain sense a borderline case, and one for which there may be nonuniqueness of minimizers. In the case of Markov coloring the condition may or may not hold; see [3] for further details.

This alternative rewriting of the objective function also has a practical benefit, in that the quantity to be minimized is now the sum of a convex function of π and a convex function of q, with no "cross terms." As a consequence, the formulas for $\pi_{i,j}$ and q_m also separate, and hence can be solved for explicitly in terms of the multipliers.

Having already restricted our attention to the case where $c(q)+h(q)$ is convex, we now make a final restriction. To parallel the very explicit computations of the single color model, we need a specific form for c, and in particular a form that allows us to solve for the minimizers in terms of multipliers. This can be done when the rate function for the coloration has a representation in terms of relative entropy, which is the case for all the models introduced previously. The particular form we choose is $c(q)=bR(q\,\|a)$, where $b\in(1,\infty)$. The limit $b\uparrow\infty$ gives the deterministic coloration with parameters a_1 and a_2, and the limit $b\downarrow 1$ gives the iid coloration with parameters a_1 and a_2.

Define

$$\mathcal{J}(x,t;y)\doteq\inf_{\substack{\varphi\in C([t,T],\mathcal{S}_{I,J})\\ \varphi(t)=x,\varphi(T)=y}}\mathcal{I}(x,t;\varphi).$$

THEOREM 3.1. *Consider the allocation problem with either the deterministic or iid coloration process with parameters $a_1 > 0$ and $a_2 > 0$, an initial condition $(x, t) \in \mathcal{D}$, and a feasible terminal condition y. Then the quantity $\mathcal{J}(x, t; y)$ has the representation*

$$\min_{(\pi,q)\in\mathcal{F}(x,t;y,T)} \left\{ \sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j} R\left(\pi_{i,j} \,\|\mathcal{P}(a_1(T-t))\times\mathcal{P}(a_2(T-t))\right) + (T-t)(b-1)R(q\,\|a) \right\}.$$

Proof. We will prove the representation for $b \in (1, \infty)$. Taking limits and using monotonicity in b will then establish the corresponding result for $b = 1$ and $b = \infty$. The same line of argument as in the single color case is followed. Hence we consider linear terminal conditions $F(y) = \langle y, \ell\rangle$ with $\ell = \ell_{i,j}, i = 0, \ldots, I+1, j = 0, \ldots, J+1$, and define

$$\begin{aligned} U(x,t) = \inf \Bigg\{ & \sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j} R\left(\pi_{i,j} \,\|\mathcal{P}(a_1(T-t))\times\mathcal{P}(a_2(T-t))\right) \\ & + (T-t)(b-1)R(q\,\|a) + \langle \ell, x\times\pi\rangle \Bigg\}. \end{aligned}$$

The infimum is over $\mathcal{F}(x, t, T)$, which is defined to be the set of all collections $(\pi_{i,j}, q_m)$ such that

$$\sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j} \sum_{r_1=0}^{\infty}\sum_{r_2=0}^{\infty} r_m \pi_{i,j}(r_1, r_2) = q_m(T-t), \qquad m = 1, 2,$$

and

$$q_m \geq 0, \quad m = 1, 2, \quad q_1 + q_2 = 1.$$

To study this problem, define

$$\begin{aligned} & f(x,t;\pi,q) \\ & \doteq \sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j} R\left(\pi_{i,j} \,\|\mathcal{P}(a_1(T-t))\times\mathcal{P}(a_2(T-t))\right) + (T-t)(b-1)R(q\,\|a) + \langle \ell, x\times\pi\rangle, \end{aligned}$$

introduce Lagrange multipliers $\Lambda = (\lambda_{i,j}, i = 0, \ldots, I+1, j = 0, \ldots, J+1; \mu_m, m = 1, 2; \theta)$, and define

$$\begin{aligned} L(x,t;\Lambda,\pi,q) \doteq{}& f(x,t;\pi,q) + \sum_{i=0}^{I+1}\sum_{j=0}^{J+1} \lambda_{i,j} x_{i,j} \left(1 - \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \pi_{i,j}(k,l)\right) \\ & + \sum_{m=1,2} \mu_m \left(q_m(T-t) - \sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j} \sum_{r_1=0}^{\infty}\sum_{r_2=0}^{\infty} r_m \pi_{i,j}(r_1,r_2)\right) \\ & + \theta(1 - q_1 - q_2). \end{aligned}$$

Analogously to the single color case,

$$\pi_{i,j}(k,l;x,t;\Lambda) = \mathcal{P}_k(a_1(T-t))\mathcal{P}_l(a_2(T-t))e^{\lambda_{i,j}-1+k\mu_1+l\mu_2-\ell_{i+k,j+l}}.$$

The equation for q is

$$(T-t)(b-1)\left[\log\left(\frac{q_m}{a_m}\right)+1\right]+\mu_m(T-t)-\theta=0,$$

so that

$$q_m(x,t;\Lambda)=a_m e^{-\frac{\mu_m}{b-1}}e^{\frac{\theta}{(T-t)(b-1)}}e^{-1}.$$

For $i=0,\dots,I,I+1, j=0,\dots,J,J+1$ and $m=1,2$ let

$$\begin{aligned}
G_{i,j}(x,t;\Lambda) &= \left(1-\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\pi_{i,j}(k,l;x,t;\Lambda)\right),\\
G_m(x,t;\Lambda) &= \left(q_m(x,t;\Lambda)(T-t)-\sum_{i=0}^{I+1}\sum_{j=0}^{J+1}x_{i,j}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}r_m\pi_{i,j}(k,l;x,t;\Lambda)\right),\\
G_3(x,t;\Lambda) &= (1-q_1(x,t;\Lambda)-q_2(x,t;\Lambda)). \qquad \square
\end{aligned}$$

When discussing uniqueness of the multipliers we must work with a matrix indexed by the subscripts of these functions, and the particular ordering of the i,j as subscripts is unimportant.

We next present three lemmas that are analogues of ones proved in the case of a single color. Since the proofs of the first two are also direct analogues they are omitted.

LEMMA 3.2 (general properties). *For any $(x,t)\in\mathcal{D}$, $\mathcal{F}(x,t;T)$ is nonempty, minimizing measures π^* exist, and if $x_{i,j}>0$, then $\pi^*_{i,j}(k,l)>0$ for all k and l.*

LEMMA 3.3 (characterization of the minimizer). *For any $(x,t)\in\mathcal{D}$ there exists $\Lambda\in\mathbb{R}^{(I+2)\times(J+2)+3}$ so that $G(x,t;\Lambda)=0$, and $\pi_{i,j}(k,l;x,t;\Lambda)$, $q_m(x,t;\Lambda)$ is a minimizer in the definition of $U(x,t)$.*

LEMMA 3.4 (uniqueness of characterization). *For $(x,t)\in\mathcal{D}$, there is only one $\Lambda\in\mathbb{R}^{(I+2)\times(J+2)+3}$ such that $G(x,t,\Lambda)=0$.*

Proof. We have

$$\begin{cases}
\frac{\partial\pi_{i,j}(k,l;x,t;\Lambda)}{\partial\lambda_{i,j}} &= \pi_{i,j}(k,l;x,t;\Lambda),\\
\frac{\partial\pi_{i,j}(k,l;x,t;\Lambda)}{\partial\mu_1} &= k\pi_{i,j}(k,l;x,t;\Lambda),\\
\frac{\partial\pi_{i,j}(k,l;x,t;\Lambda)}{\partial\mu_2} &= l\pi_{i,j}(k,l;x,t;\Lambda),\\
\frac{\partial q_m(x,t;\Lambda)}{\partial\mu_m} &= -\frac{1}{b-1}q_m(x,t;\Lambda),\\
\frac{\partial q_m(x,t;\Lambda)}{\partial\theta} &= \frac{1}{(T-t)(b-1)}q_m(x,t;\Lambda),
\end{cases}$$

and all other partial derivatives are zero. As in the single color case it is enough to show the negative definiteness of $D_\Lambda G$. Using a suitable dominating function to justify the interchange of differentiation and summation and the definitions

$$\begin{aligned}
\alpha_{i,j} &= \sum_{k,l=0}^{\infty}\pi_{i,j}(k,l;x,t;\Lambda),\\
T_m^{i,j} &= \sum_{r_1,r_2=0}^{\infty}r_m\pi_{i,j}(r_1,r_2;x,t;\Lambda),\\
C_{1,1} &= \sum_{i=0}^{I+1}\sum_{j=0}^{J+1}x_{i,j}\sum_{r_1,r_2=0}^{\infty}r_1^2\pi_{i,j}(r_1,r_2;x,t;\Lambda),
\end{aligned}$$

$$C_{2,2} = \sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j} \sum_{r_1,r_2=0}^{\infty} r_2^2 \pi_{i,j}(r_1,r_2;x,t;\Lambda),$$
$$C_{1,2} = \sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j} \sum_{r_1,r_2=0}^{\infty} r_1 r_2 \pi_{i,j}(r_1,r_2;x,t;\Lambda),$$

$D_\Lambda G$ equals [with $q_m = q_m(x,t;\Lambda)$]

$$\begin{pmatrix} -\alpha_{0,0} & \cdots & 0 & -T_1^{0,0} & -T_2^{0,0} & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & -\alpha_{I+1,J+1} & -T_1^{I+1,j+1} & -T_2^{I+1,j+1} & 0 \\ -x_{0,0}T_1^{0,0} & \cdots & -x_{I+1,J+1}T_1^{I+1,J+1} & -\frac{T-t}{b-1}q_1 - C_{1,1} & -C_{1,2} & -\frac{1}{b-1}q_1 \\ -x_{0,0}T_2^{0,0} & \cdots & -x_{I+1,J+1}T_2^{I+1,J+1} & -C_{1,2} & -\frac{T-t}{b-1}q_2 - C_{2,2} & -\frac{1}{b-1}q_2 \\ 0 & \cdots & 0 & \frac{1}{b-1}q_1 & \frac{1}{b-1}q_2 & -\frac{q_1+q_2}{(T-t)(b-1)} \end{pmatrix}.$$

Since diagonalizing all, save the lower right 3×3 submatrix, produces strictly negative values on the diagonal, we need only check the negative definiteness of

$$\begin{pmatrix} \sum_{i,j}\frac{x_{i,j}}{\alpha_{i,j}}(T_1^{i,j})^2 - \frac{T-t}{b-1}q_1 - C_{1,1} & \sum_{i,j}\frac{x_{i,j}}{\alpha_{i,j}}T_1^{i,j}T_2^{i,j} - C_{1,2} & -\frac{1}{b-1}q_1 \\ \sum_{i,j}\frac{x_{i,j}}{\alpha_{i,j}}T_1^{i,j}T_2^{i,j} - C_{1,2} & \sum_{i,j}\frac{x_{i,j}}{\alpha_{i,j}}(T_2^{i,j})^2 - \frac{T-t}{b-1}q_2 - C_{2,2} & -\frac{1}{b-1}q_2 \\ \frac{1}{b-1}q_1 & \frac{1}{b-1}q_2 & -\frac{q_1+q_2}{(T-t)(b-1)} \end{pmatrix}.$$

Since

$$\begin{pmatrix} -\frac{T-t}{b-1}q_1 & 0 & -\frac{1}{b-1}q_1 \\ 0 & -\frac{T-t}{b-1}q_2 & -\frac{1}{b-1}q_2 \\ \frac{1}{b-1}q_1 & \frac{1}{b-1}q_2 & -\frac{q_1+q_2}{(T-t)(b-1)} \end{pmatrix}$$

is obviously negative definite, we need only check the 2×2 matrix

$$\begin{pmatrix} \sum_{i,j}\frac{x_{i,j}}{\alpha_{i,j}}(T_1^{i,j})^2 - C_{1,1} & \sum_{i,j}\frac{x_{i,j}}{\alpha_{i,j}}T_1^{i,j}T_2^{i,j} - C_{1,2} \\ \sum_{i,j}\frac{x_{i,j}}{\alpha_{i,j}}T_1^{i,j}T_2^{i,j} - C_{1,2} & \sum_{i,j}\frac{x_{i,j}}{\alpha_{i,j}}(T_2^{i,j})^2 - C_{2,2} \end{pmatrix}.$$

However, letting $\pi_{i,j}(r_1,r_2) = \pi_{i,j}(r_1,r_2;x,t;\Lambda)$ and pre- and postmultiplying by the nonzero vector (z_1,z_2) produces

$$\sum_{i,j}\frac{x_{i,j}}{\alpha_{i,j}}\left(\sum_{r_1,r_2}(z_1r_1+z_2r_2)\pi_{i,j}(r_1,r_2)\right.$$
$$\left. - \sum_{r_1,r_2}(z_1r_1+z_2r_2)^2\pi_{i,j}(r_1,r_2)\cdot\sum_{r_1,r_2}\pi_{i,j}(r_1,r_2)\right) \le 0.$$

Thus the entire matrix is negative definite. □

Since we have restricted our attention to the case of MB there is no analogue of the nonsmooth function $\tau(x,t)$, and hence the existence of a smooth extension of U to a neighborhood of $\mathcal{D}$ follows directly from the implicit function theorem. The next result expresses the derivatives in terms of the multipliers.

THEOREM 3.5. *Fix $(x,t) \in D$, and let Λ^* be the associated Lagrange multiplier. Then*

$$D_{x_{i,j}} U(x,t) = (\lambda^*_{i,j} - 1)$$

and

$$D_t U(x,t) = (b-1) - \frac{\theta^*}{(T-t)}.$$

Proof. With

$$H(x,t;\Lambda) \doteq L(x,t;\Lambda,\pi_{i,j}(\cdot,x,t;\Lambda), q_m(x,t;\Lambda))$$

we can write

$$U(x,t) = H(x,t;\Lambda(x,t)),$$

where $\Lambda(x,t)$ is the unique solution to the constraint equations. Then

$$D_{x_{i,j}} U(x,t) = D_{x_{i,j}} H(x,t;\Lambda^*) + D_\Lambda H(x,t;\Lambda^*) D_{x_{i,j}} \Lambda(x,t).$$

As in the single color case $D_\Lambda H(x,t;\Lambda^*) = 0$, and so $D_{x_{i,j}} U(x,t) = D_{x_{i,j}} H(x,t;\Lambda^*)$, and an analogous argument also gives $D_t U(x,t) = D_t H(x,t;\Lambda^*)$. Note that both $\pi_{i,j}(k,l;x,t;\Lambda)$ and $q_m(x,t;\Lambda)$ are actually independent of x, and hence can be written as $\pi_{i,j}(k,l;t;\Lambda)$ and $q_m(t;\Lambda)$. Keeping in mind the t dependence but temporarily suppressing both t and Λ^* in the notation, we will use

$$(T-t)(b-1)R(q\,\|a) = -(T-t)\left[q_1\mu_1^* + q_2\mu_2^*\right] + [\theta^* - (T-t)(b-1)](q_1+q_2).$$

The quantity to be differentiated is thus

$$\begin{aligned}
&\sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j} \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \left[\lambda^*_{i,j} - 1 + k\mu_1^* + l\mu_2^*\right] \pi_{i,j}(k,l) - (T-t)\left[q_1\mu_1^* + q_2\mu_2^*\right] \\
&\quad + [\theta^* - (T-t)(b-1)](q_1+q_2) \\
&\quad + \sum_{i=0}^{I+1}\sum_{j=0}^{J+1} \lambda^*_{i,j} x_{i,j} \left(1 - \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \pi_{i,j}(k,l)\right) \\
&\quad + \sum_{m=1,2} \mu_m^* \left(q_m(T-t) - \sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j} \sum_{r_1=0}^{\infty}\sum_{r_2=0}^{\infty} r_m \pi_{i,j}(r_1,r_2)\right) \\
&\quad + \theta^*(1-q_1-q_2) \\
&= -\sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j} \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \pi_{i,j}(k,l) - (T-t)(b-1)(q_1+q_2) + \sum_{i=0}^{I+1}\sum_{j=0}^{J+1} \lambda^*_{i,j} x_{i,j} + \theta^*.
\end{aligned}$$

Since

$$\begin{aligned}
\frac{\partial \pi_{i,j}(k,l;t;\Lambda)}{\partial t} &= e^{\lambda_{i,j} - 1 + k\mu_1 + l\mu_2 - \ell_{i+k,j+l}} \frac{\partial \mathcal{P}_k(a_1(T-t))\mathcal{P}_l(a_2(T-t))}{\partial t} \\
&= \frac{e^{\lambda_{i,j} - 1 + k\mu_1 + l\mu_2 - \ell_{i+k,j+l}}}{k!l!} \frac{\partial e^{-(T-t)} a_1^k (T-t)^k a_2^l (T-t)^l}{\partial t} \\
&= \left[1 - \frac{k}{T-t} - \frac{l}{T-t}\right] \pi_{i,j}(k,l;t;\Lambda)
\end{aligned}$$

and

$$\frac{\partial q_m(t;\Lambda)}{\partial t} = \left[\frac{\theta}{(T-t)^2(b-1)}\right] q_m(t;\Lambda),$$

it follows that

$$D_{x_{i,j}} U(x,t) = (\lambda^*_{i,j} - 1)$$

and

$$\begin{aligned} D_t U(x,t) &= -\sum_{i=0}^{I+1}\sum_{j=0}^{J+1} x_{i,j} \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\left[1 - \frac{k}{T-t} - \frac{l}{T-t}\right]\pi_{i,j}(k,l) \\ &\quad + (b-1)\,(q_1+q_2) - \frac{\theta^*}{(T-t)}\,(q_1+q_2) \\ &= -1 + q_1 + q_2 + (b-1) - \frac{\theta^*}{(T-t)} \\ &= (b-1) - \frac{\theta^*}{(T-t)}. \qquad \square \end{aligned}$$

For the particular rate function of interest here,

$$\begin{aligned} b(\gamma) &= \sup\left[\langle \gamma, q\rangle - bR(q\,\|a\,)\right] \\ &= -b\inf\left[-\frac{1}{b}\langle \gamma, q\rangle + R(q\,\|a\,)\right] \\ &= b\log\left(e^{\frac{\gamma_1}{b}}a_1 + e^{\frac{\gamma_2}{b}}a_2\right). \end{aligned}$$

Hence by (3.1), the PDE to be satisfied by U takes the form

$$\begin{aligned} &W_t - b\log\left(\left(\sum_{i,j,i\le I} e^{-(W_{x_{i+1,j}} - W_{x_{i,j}})}x_{i,j} + \sum_{i,j,i=I+1} x_{i,j}\right)^{\frac{1}{b}} a_1\right. \\ &\qquad \left. + \left(\sum_{i,j,j\le J} e^{-(W_{x_{i,j+1}} - W_{x_{i,j}})}x_{i,j} + \sum_{i,j,j=J+1} x_{i,j}\right)^{\frac{1}{b}} a_2\right) = 0. \end{aligned} \tag{3.2}$$

THEOREM 3.6. *U satisfies* (3.2) *on $\mathcal{D}$.*

Proof. By Theorem 3.5, this will be true if

(3.3)

$$\begin{aligned} &(b-1) - \frac{\theta^*}{(T-t)} - b\log\left(\left(\sum_{i,j,i\le I} e^{-(\lambda^*_{i+1,j} - \lambda^*_{i,j})}x_{i,j} + \sum_{i,j,i=I+1} x_{i,j}\right)^{\frac{1}{b}} a_1\right. \\ &\qquad\qquad \left. + \left(\sum_{i,j,j\le J} e^{-(\lambda^*_{i,j+1} - \lambda^*_{i,j})}x_{i,j} + \sum_{i,j,j=J+1} x_{i,j}\right)^{\frac{1}{b}} a_2\right) = 0. \end{aligned}$$

For $i \le I$,

$$\frac{\pi_{i+1,j}(k,l)}{\pi_{i,j}(k+1,l)} = \frac{e^{\lambda_{i+1,j}-\lambda_{i,j}} e^{-\mu_1}(k+1)}{a_1(T-t)}.$$

Summing on k and l gives

$$e^{\lambda_{i,j}-\lambda_{i+1,j}} = \frac{e^{-\mu_1}}{a_1(T-t)} \sum_{r_1,r_2} r_1 \pi_{i,j}(r_1,r_2),$$

and the analogous formula

$$e^{\lambda_{i,j}-\lambda_{i,j+1}} = \frac{e^{-\mu_2}}{a_2(T-t)} \sum_{r_1,r_2} r_2 \pi_{i,j}(r_1,r_2)$$

applies for $j \le J$. We also have

$$1 = \frac{e^{-\mu_2}}{a_2(T-t)} \sum_{r_1,r_2} r_2 \pi_{I+1,j}(r_1,r_2) = \frac{e^{-\mu_2}}{a_2(T-t)} \sum_{r_1,r_2} r_2 \pi_{i,J+1}(r_1,r_2)$$

if $j \le J+1$ or $i \le I+1$. Hence

$$\sum_{i,j,i\le I} e^{-(\lambda^*_{i+1,j}-\lambda^*_{i,j})} x_{i,j} + \sum_{i,j,i=I+1} x_{i,j} = e^{-\mu_1^*} \frac{q_1}{a_1},$$

$$\sum_{i,j,j\le J} e^{-(\lambda^*_{i,j+1}-\lambda^*_{i,j})} x_{i,j} + \sum_{i,j,j=J+1} x_{i,j} = e^{-\mu_2^*} \frac{q_2}{a_2}.$$

Now $q_1 + q_2 = 1$ implies

$$\begin{aligned} 0 &= \log\left(q_1(t;\Lambda) + q_2(t;\Lambda)\right) \\ &= \log\left(a_1 e^{-\frac{\mu_1}{b-1}} e^{\frac{\theta}{(T-t)(b-1)}-1} + a_2 e^{-\frac{\mu_2}{b-1}} e^{\frac{\theta}{(T-t)(b-1)}-1}\right) \\ &= \log\left(a_1 e^{-\frac{\mu_1}{b-1}} + a_2 e^{-\frac{\mu_2}{b-1}}\right) + \frac{\theta}{(T-t)(b-1)} - 1. \end{aligned}$$

The left-hand side of (3.3) becomes

$$\begin{aligned} &(b-1) - \frac{\theta^*}{(T-t)} - b\log\left(\left(e^{-\mu_1^*} q_1/a_1\right)^{\frac{1}{b}} a_1 + \left(e^{-\mu_2^*} q_2/a_2\right)^{\frac{1}{b}} a_2\right) \\ &= (b-1) - \frac{\theta^*}{(T-t)} - b\log\left(\left(e^{-\mu_1^*} e^{-\frac{\mu_1^*}{b-1}} e^{\frac{\theta^*}{(T-t)(b-1)}-1}\right)^{\frac{1}{b}} a_1\right. \\ &\qquad\qquad \left. + \left(e^{-\mu_2^*} e^{-\frac{\mu_2^*}{b-1}} e^{\frac{\theta^*}{(T-t)(b-1)}-1}\right)^{\frac{1}{b}} a_2\right) \\ &= -(b-1) + \frac{\theta^*}{(T-t)} - b\left[\log\left(\left(e^{-\frac{b\mu_1^*}{b-1}}\right)^{\frac{1}{b}} a_1 + \left(e^{-\frac{b\mu_2^*}{b-1}}\right)^{\frac{1}{b}} a_2\right) + \frac{\theta^*}{b(T-t)(b-1)} - \frac{1}{b}\right] \\ &= -b + \frac{b\theta^*}{(T-t)(b-1)} - \frac{b\theta^*}{(T-t)(b-1)} + b \\ &= 0, \end{aligned}$$

and the theorem is proved. $\square$

3.4. Minimizing trajectories. We end this section by stating without proof the form of the minimizing trajectories. As in the case of a single color we consider only the empty initial condition. In contrast with that case, here the minimizing q must be determined first via Lagrange multipliers. Once q is given, we define

$$\varphi_{0,0}(t_1,t_2) \doteq C\mathcal{P}_0(\rho q_1 t_1)\mathcal{P}_0(\rho q_2 t_2) + \sum_{k=0}^{I}\sum_{l=0}^{J}(y_{k,l} - C\mathcal{P}_k(\rho q_1 T)\mathcal{P}_l(\rho q_2 T))\left(1-\frac{t_1}{T}\right)^k\left(1-\frac{t_2}{T}\right)^l$$

and

$$\varphi_{i,j}(t_1,t_2) \doteq \frac{(-t_1)^i}{i!}\frac{(-t_2)^j}{j!}\varphi_{0,0}^{(i,j)}(t_1,t_2).$$

In terms of these functions we set

$$\varphi_{i,j}(t) \doteq \varphi_{i,j}(t,t),$$

and for $i \le I$ and $j \le J$,

$$\varphi_{I+,j}(t) = \sum_{i=I+1}^{\infty}\varphi_{i,j}(t),\ \varphi_{i,J+}(t) = \sum_{j=J+1}^{\infty}\varphi_{i,j}(t),\ \varphi_{I+J+}(t) = \sum_{i=I+1}^{\infty}\sum_{j=J+1}^{\infty}\varphi_{i,j}(t).$$

With q determined via Lagrange multipliers, the parameters $\rho > 0$ and $C \ge 0$ are chosen so that

$$\frac{\rho T - \sum_{i=0}^{I}\sum_{j=0}^{J}(i+j)\mathcal{P}_i(\rho q_1 T)\mathcal{P}_j(\rho q_2 T)}{1-\sum_{i=0}^{I}\sum_{j=0}^{J}\mathcal{P}_i(\rho q_1 T)\mathcal{P}_j(\rho q_2 T)} = \frac{T-\sum_{i=0}^{I}\sum_{j=0}^{J}(i+j)y_{i,j}}{1-\sum_{i=0}^{I}\sum_{j=0}^{J}y_{i,j}}$$

and

$$C \doteq \frac{1-\sum_{i=0}^{I}\sum_{j=0}^{J}y_{i,j}}{1-\sum_{i=0}^{I}\sum_{j=0}^{J}\mathcal{P}_i(\rho q_1 T)\mathcal{P}_j(\rho q_2 T)} = \frac{T-\sum_{i=0}^{I}\sum_{j=0}^{J}(i+j)y_{i,j}}{\rho T - \sum_{i=0}^{I}\sum_{j=0}^{J}(i+j)\mathcal{P}_i(\rho q_1 T)\mathcal{P}_j(\rho q_2 T)}.$$

The proof that these trajectories achieve the minimal cost parallels that of the single color case, with an appropriate modification of the notion of completely monotone that is suitable for functions of two independent variables.

REFERENCES

[1] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, John Wiley & Sons, New York, 1997.
[2] P. Dupuis, C. Nuzman, and P. Whiting, *Large deviation asymptotics for occupancy problems*, Ann. Probab., 32 (2004), pp. 2765–2818.
[3] P. Dupuis, C. Nuzman, and P. Whiting, *Large deviation principle for occupancy problems with colored balls*, J. Appl. Probab., 44 (2007), pp. 115–141.
[4] W. H. Fleming and R. Rishel, *Deterministic and Stochastic Optimal Control*, Springer, Berlin, 1975.
[5] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York, 1984.
[6] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[7] J. Zhang and P. Dupuis, *Large deviation principle for general occupancy models*, in Combin. Probab. Comput., to appear.

SIAM J. MATH. ANAL.
Vol. 39, No. 5, pp. 1668–1685

LONG-TIME EXISTENCE OF SMOOTH SOLUTIONS FOR THE RAPIDLY ROTATING SHALLOW-WATER AND EULER EQUATIONS*

BIN CHENG† AND EITAN TADMOR‡

Abstract. We study the stabilizing effect of rotational forcing in the nonlinear setting of two-dimensional shallow-water and more general models of compressible Euler equations. In [*Phys. D*, 188 (2004), pp. 262–276] Liu and Tadmor have shown that the pressureless version of these equations admit a global smooth solution for a large set of subcritical initial configurations. In the present work we prove that when rotational force dominates the pressure, it *prolongs* the lifespan of smooth solutions for $t \lesssim \ln(\delta^{-1})$; here $\delta \ll 1$ is the ratio of the pressure gradient measured by the inverse squared Froude number, relative to the dominant rotational forces measured by the inverse Rossby number. Our study reveals a "nearby" periodic-in-time approximate solution in the small δ regime, upon which hinges the long-time existence of the exact smooth solution. These results are in agreement with the close-to-periodic dynamics observed in the "near-inertial oscillation" (NIO) regime which follows oceanic storms. Indeed, our results indicate the existence of a smooth, "approximate periodic" solution for a time period of *days*, which is the relevant time period found in NIO obesrvations.

Key words. shallow-water equations, rapid rotation, pressureless equations, critical threshold, two-dimensinoal Euler equations, long-time existence

AMS subject classifications. 76U05, 76E07, 76N10

DOI. 10.1137/070693643

1. Introduction and statement of main results.

We are concerned here with two-dimensional systems of nonlinear Eulerian equations driven by pressure and rotational forces. It is well known that in the absence of rotation, these equations experience a finite-time breakdown: for generic smooth initial conditions, the corresponding solutions will lose C^1-smoothness due to shock formation. The presence of rotational forces, however, has a stabilizing effect. In particular, the pressureless version of these equations admit global smooth solutions for a large set of so-called subcritical initial configurations [17]. It is therefore a natural extension to investigate the balance between the regularizing effects of rotation versus the tendency of pressure to enforce finite-time breakdown (we mention in passing the recent work [21] on a similar regularizing balance of different competing forces in the one-dimensional Euler–Poisson equations). In this paper we prove the long-time existence of rapidly rotating flows characterized by "nearby" periodic flows. Thus, rotation *prolongs* the lifespan of smooth solutions over increasingly long time periods, which grow longer as the rotation forces become more dominant over pressure.

Our model problem is the rotational shallow water (RSW) system of equations. This system models largescale geophysical motions in a thin layer of fluid under the

*Received by the editors June 4, 2007; accepted for publication (in revised form) October 9, 2007; published electronically January 30, 2008. This research was supported in part by NSF grants 04-07704 and 07-07949 and by ONR grant N00014-91-J-1076.
http://www.siam.org/journals/sima/39-5/69364.html

†Department of Mathematics, Center of Scientific Computation and Mathematical Modeling (CSCAMM), University of Maryland, College Park, MD 20742. Current address: Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 (bincheng@umich.edu).

‡Department of Mathematics, Center of Scientific Computation and Mathematical Modeling (CSCAMM), University of Maryland, College Park, MD 20742 (tadmor@cscamm.umd.edu).

influence of the Coriolis rotational forcing (see, e.g., [18, section 3.3], [10, section 2.1]),

$$\partial_t h + \nabla \cdot (h\mathbf{u}) = 0, \tag{1.1a}$$

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + g\nabla h - f\mathbf{u}^\perp = 0. \tag{1.1b}$$

It governs the unknown velocity field $\mathbf{u} := \big(u^{(1)}(t,x,y), u^{(2)}(t,x,y)\big)$ and height $h := h(t,x,y)$, where g and f stand for the gravitational constant and the Coriolis frequency, respectively. Recall that (1.1a) observes the conservation of mass and (1.1b) describe balance of momentum by the pressure gradient, $g\nabla h$, and rotational forcing, $f\mathbf{u}^\perp := f\big(u^{(2)}, -u^{(1)}\big)$.

For convenience, we rewrite system (1.1) in terms of rescaled, nondimensional variables. To this end, we introduce the characteristic scales, H for total height h, D for height fluctuation $h-H$, U for velocity u, L for spatial length, and correspondingly, L/U for time, and we make the change of variables

$$\mathbf{u} = \mathbf{u}'\left(\frac{t'L}{U}, x'L, y'L\right)U, \qquad h = H + h'\left(\frac{t'L}{U}, x'L, y'L\right)D.$$

Discarding all the primes, we arrive at a nondimensional system,

$$\partial_t h + \mathbf{u} \cdot \nabla h + \left(\frac{H}{D} + h\right)\nabla \cdot \mathbf{u} = 0,$$

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \frac{gD}{U^2}\nabla h - \frac{fL}{U}\mathbf{u}^\perp = 0.$$

We are concerned here with the regime where the pressure gradient and compressibility are of the same order, $\frac{gD}{U^2} \approx \frac{H}{D}$. Thus we arrive at the (symmetrizable) RSW system,

$$\partial_t h + \mathbf{u} \cdot \nabla h + \left(\frac{1}{\sigma} + h\right)\nabla \cdot \mathbf{u} = 0, \tag{1.2a}$$

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \frac{1}{\sigma}\nabla h - \frac{1}{\tau}J\mathbf{u} = 0. \tag{1.2b}$$

Here σ and τ, given by

$$\sigma := \frac{U}{\sqrt{gH}}, \quad \tau := \frac{U}{fL}, \tag{1.2c}$$

are, respectively, the Froude number measuring the inverse pressure forcing and the Rossby number measuring the inverse rotational forcing. We use J to denote the 2×2 rotation matrix $J := \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$.

To trace their long-time behavior, we approximate (1.2a), (1.2b) with the successive iterations,

$$\partial_t h_j + \mathbf{u}_{j-1} \cdot \nabla h_j + \left(\frac{1}{\sigma} + h_j\right)\nabla \cdot \mathbf{u}_{j-1} = 0, \qquad j = 2, 3, \ldots, \tag{1.3a}$$

$$\partial_t \mathbf{u}_j + \mathbf{u}_j \cdot \nabla \mathbf{u}_j + \frac{1}{\sigma}\nabla h_j - \frac{1}{\tau}J\mathbf{u}_j = 0, \qquad j = 1, 2, \ldots, \tag{1.3b}$$

subject to initial conditions, $h_j(0,\cdot) = h_0(\cdot)$ and $\mathbf{u}_j(0,\cdot) = \mathbf{u}_0(\cdot)$. Observe that, given j, (1.3a,b) are only weakly coupled through the dependence of $\mathbf{u}_j$ on h_j, so that we need only specify the initial height h_1. Moreover, for $\sigma \gg \tau$, the momentum equations (1.2b) are "approximately decoupled" from the mass equation (1.2a) since rotational forcing is substantially dominant over pressure forcing. Therefore, a first approximation of constant height function will enforce this decoupling, serving as the starting point of the above iterative scheme,

$$h_1 \equiv \text{constant}. \tag{1.3c}$$

This, in turn, leads to the first approximate velocity field, $\mathbf{u}_1$, satisfying the pressureless equations,

$$\partial_t \mathbf{u}_1 + \mathbf{u}_1 \cdot \nabla \mathbf{u}_1 - \frac{1}{\tau} J \mathbf{u}_1 = 0, \quad \mathbf{u}_1(0, \cdot) = \mathbf{u}_0(\cdot). \tag{1.4}$$

Liu and Tadmor [17] have shown that there is a "large set" of so-called subcritical initial configurations $\mathbf{u}_0$, for which the pressureless equations (1.4) admit global smooth solutions. Moreover, the pressureless velocity $\mathbf{u}_1(t, \cdot)$ is in fact $2\pi\tau$-periodic in time. The regularity of $\mathbf{u}_1$ is discussed in section 2.

Having the pressureless solution, $(h_1 \equiv \text{constant}, \mathbf{u}_1)$ as a first approximation for the RSW solution $(h, \mathbf{u})$, in section 3 we introduce an improved approximation of the RSW equations, $(h_2, \mathbf{u}_2)$, which solves an "adapted" version of the second iteration $(j = 2)$ of (1.3). This improved approximation satisfies a specific linearization of the RSW equations around the pressureless velocity $\mathbf{u}_1$, with only a one-way coupling between the momentum and the mass equations. Building on the regularity and periodicity of the pressureless velocity $\mathbf{u}_1$, we show that the solution of this linearized system subject to subcritical initial data $(h_0, \mathbf{u}_0)$, is globally smooth; in fact, both $h_2(t, \cdot)$ and $\mathbf{u}_2(t, \cdot)$ retain $2\pi\tau$-periodicity in time.

Next, we turn to estimating the deviation between the solution of the linearized RSW system, $(h_2, \mathbf{u}_2)$, and the solution of the full RSW system, $(h, \mathbf{u})$. To this end, we introduce a new nondimensional parameter,

$$\delta := \frac{\tau}{\sigma^2} = \frac{gH}{fLU},$$

measuring the *relative* strength of rotation versus the pressure forcing, and we assume that rotation is the dominant forcing in the sense that $\delta \ll 1$. Using the standard energy method, we show in Theorem 4.1 and the follow-up Remark 4.1 that, starting with H^m subcritical initial data, the RSW solution $\big(h(t, \cdot), \mathbf{u}(t, \cdot)\big)$ remains sufficiently close to $\big(h_2(t, \cdot), \mathbf{u}_2(t, \cdot)\big)$ in the sense that

$$\|h(t, \cdot) - h_2(t, \cdot)\|_{H^{m-3}} + \|\mathbf{u}(t, \cdot) - \mathbf{u}_2(t, \cdot)\|_{H^{m-3}} \lesssim \frac{e^{C_0 t}\delta}{(1 - e^{C_0 t}\delta)^2},$$

where constant $C_0 = \widehat{C}_0(m, |\nabla \mathbf{u}_0|_\infty, |h_0|_\infty) \cdot \|\mathbf{u}_0, h_0\|_m$. In particular, we conclude that for a large set of subcritical initial data, the RSW equations (1.2) admit smooth, "approximate periodic" solutions for long time, $t \le t_\delta := \ln(\delta^{-1})$, in the rotationally dominant regime $\delta \ll 1$.

We comment that our formal notion of "approximate periodicity" emphasizes the existence of a periodic approximation $(h_2, \mathbf{u}_2)$ nearby the actual flow $(h, \mathbf{u})$, with an up-to $O(\delta) \ll 1$-error for sufficiently long time. Therefore, strong rotation stabilizes the flow by imposing on it approximate periodicity, which in turn postpones finite-time breakdown of classical solutions for a long time. A convincing example is provided by the so-called "near-inertial oscillation" (NIO) regime, which is observed during the days that follow oceanic storms; see, e.g., [22]. These NIOs are triggered when storms pass by (large U's) and only a thin layer of the oceans is reactive (small aspect ratio H/L), corresponding to $\delta = \frac{gH}{fLU} \ll 1$. Specifically, with Rossby number $\tau \sim \mathcal{O}(0.1)$ and Froude number $\sigma \sim \mathcal{O}(1)$ we find $\delta \sim 0.1$, which yields the existence of a smooth, "approximate periodic" solution for $t \sim 2$ days. We note that the clockwise rotation of cyclonic storms on the Northern Hemisphere produce negative vorticity, which is a preferred scenario of the subcritical condition (2.1b). Our results are consistent with the observations regarding the stability and approximate periodicity of the NIO regime.

Next, we generalize our result to Euler systems describing the isentropic gasdynamics, in section 4.2, and ideal gasdynamics, in section 4.3. We regard these two systems as successive generalizations of the RSW system using the following formalism:

$$\partial_t \rho + \nabla \cdot (\rho \mathbf{u}) = 0, \tag{1.5a}$$

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \rho^{-1} \nabla \widetilde{p}(\rho, S) = fJ\mathbf{u}, \tag{1.5b}$$

$$\partial_t S + \mathbf{u} \cdot \nabla S = 0. \tag{1.5c}$$

Here, the physical variables ρ, S are, respectively, the density and entropy. We use $\widetilde{p}(\rho, S)$ for the gas-specific pressure law relating pressure to density and entropy. For the ideal gasdynamics, the pressure law is given as $\widetilde{p} := A\rho^\gamma e^S$, where A, γ are two gas-specific physical constants. The isentropic gas equations correspond to constant S, for which the entropy equation (1.5c) becomes redundant. Setting $A = g, \gamma = 2$ yields the RSW equations with ρ playing the same role as h.

The general Euler system (1.5) can be symmetrized by introducing a "normalized" pressure function,

$$p := \frac{\sqrt{\gamma}}{\gamma - 1} \widetilde{p}^{\frac{\gamma-1}{2\gamma}}(\rho, S),$$

and by replacing the density equation (1.5a) with a pressure equation,

$$\partial_t p + \mathbf{u} \cdot \nabla p + \frac{\gamma - 1}{2} p \nabla \cdot \mathbf{u} = 0. \tag{1.5d}$$

We then nondimensionalize the above system (1.5b), (1.5c), (1.5d) into

$$\begin{aligned} \partial_t p + \mathbf{u} \cdot \nabla p + \frac{\gamma - 1}{2} \left(\frac{1}{\sigma} + p \right) \nabla \cdot \mathbf{u} = 0, \\ \partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \frac{\gamma - 1}{2} \left(\frac{1}{\sigma} + p \right) e^{\sigma S} \nabla p = \frac{1}{\tau} J\mathbf{u}, \\ \partial_t S + \mathbf{u} \cdot \nabla S = 0. \end{aligned}$$

The same methodology introduced for the RSW equations still applies to the more general Euler system, independent of the pressure law. In particular, our first approximation, the pressureless system, remains the same as in (1.4) since it ignores any effect of pressure. We then obtain the second approximation $(p_2, \mathbf{u}_2, S_2)$ (or $(p_2, \mathbf{u}_2)$ in the isentropic case) from a specific linearization around the pressureless velocity $\mathbf{u}_1$. Thanks to the fact that h, p, and S share a similar role as passive scalars transported by $\mathbf{u}$, the same regularity and periodicity argument can be employed for $(p_2, \mathbf{u}_2, S_2)$

in these general cases as for $(h_2, \mathbf{u}_2)$ in the RSW case. The energy estimate, however, needs careful modification for the ideal gas equations due to additional nonlinearity. Finally, we conclude in Theorem 4.2 and 4.3 that, in the rotationally dominant regime $\delta \ll 1$, the exact solution stays "close" to the globally smooth, $2\pi\tau$-periodic approximate solution $(p_2, \mathbf{u}_2, S_2)$ for long time in the sense that, starting with H^m subcritical data, the following estimate holds true for time $t \lesssim \ln(\delta^{-1})$:

$$\|p(t,\cdot) - p_2(t,\cdot)\|_{m-3} + \|\mathbf{u}(t,\cdot) - \mathbf{u}_2(t,\cdot)\|_{m-3} + \|S(t,\cdot) - S_2(t,\cdot)\|_{m-3} < \frac{e^{C_0 t}\delta}{1 - e^{C_0 t}\delta}.$$

Our results confirm the stabilization effect of rotation in the nonlinear setting, when it interacts with the slow components of the system, which otherwise tend to destabilize of the dynamics. The study of such interaction is essential to the understanding of rotating dynamics, primarily to geophysical flows. We can mention only a few works from the vast literature available on this topic, and we refer the reader to the recent book of Chemin et al. [6], and the references therein, for a state-of-the-art of the mathematical theory for rapidly rotating flows. Embid and Majda [7, 8] studied the singular limit of RSW equations under the two regimes $\tau^{-1} \sim \sigma^{-1} \to \infty$ and $\tau^{-1} \sim \mathcal{O}(1)$, $\sigma^{-1} \to \infty$. Extensions to more general skew-symmetric perturbations can be found in the work of Gallagher [9]. The series of works of Babin, Mahalov, and Nicolaenko (consult [1, 2, 3, 4, 5] and references therein) establish long-term stability effects of the rapidly rotating three-dimensional Euler, Navier–Stokes, and primitive equations. Finally, we mention the work of Zeitlin, Reznik, and Ben Jelloul [23, 24] which categorize several relevant scaling regimes and, correspondingly, derive formal asymptotics in the nonlinear setting.

We comment here that the approach pursued in the above literature relies on identifying the limiting system as $\tau \to 0$, which filters out fast scales. The full system is then approximated to a first order by this slowly evolving limiting system. A rigorous mathematical foundation along these lines was developed by Schochet [19], which can be traced back to the earlier works of Klainerman and Majda [13] and Kreiss [14] (see also [20]). The key point was the separation of (linear) fast oscillations from the slow scales. The novelty of our approach, inspired by the critical threshold phenomena [16], is to adopt the rapidly oscillating and fully *nonlinear* pressureless system as a first approximation and then consider the full system as a perturbation of this fast scale. This enables us to preserve both slow and fast dynamics, and especially, the rotation-induced time periodicity.

2. First approximation—the pressureless system. We consider the pressureless system

$$\partial_t \mathbf{u}_1 + \mathbf{u}_1 \cdot \nabla \mathbf{u}_1 - \frac{1}{\tau} J \mathbf{u}_1 = 0 \tag{2.1a}$$

subject to initial condition $\mathbf{u}_1(0,\cdot) = \mathbf{u}_0(\cdot)$. We begin by recalling the main theorem in [17] regarding the global regularity of the pressureless equations (2.1a).

THEOREM 2.1. *Consider the pressureless equations* (2.1a) *subject to C^1-initial data $\mathbf{u}_1(0,\cdot) = u_0(\cdot)$. Then, the solution $\mathbf{u}_1(t,\cdot)$ stays C^1 for all time if and only if the initial data satisfy the critical threshold condition,*

$$\tau\omega_0(x) + \frac{\tau^2}{2}\eta_0^2(x) < 1 \quad \textit{for all } x \in \mathbb{R}^2. \tag{2.1b}$$

Here, $\omega_0(x) = -\nabla \times \mathbf{u}_0(x) = \partial_y u_0 - \partial_x v_0$ is the initial vorticity and $\eta_0(x) := \lambda_1 - \lambda_2$ is the (possibly complex-valued) spectral gap associated with the eigenvalues of gradient *matrix* $\nabla\mathbf{u}_0(x)$. *Moreover, these globally smooth solutions,* $\mathbf{u}_1(t,\cdot)$, *are* $2\pi\tau$*-periodic in time.*

In [17], Liu and Tadmor gave two different proofs of (2.1b). One was based on the spectral dynamics of $\lambda_j(\nabla\mathbf{u})$; the other was based on the flow map associated with (2.1a), and here we note yet another version of the latter, based on the Riccati-type equation satisfied by the gradient matrix $M =: \nabla\mathbf{u}_1$,

$$M' + M^2 = \tau^{-1}JM.$$

Here $\{\cdot\}' := \partial_t + \mathbf{u}_1\cdot\nabla$ denotes differentiation along the particle trajectories

$$\Gamma_0 := \{(x,t) \mid \dot{x}(t) = \mathbf{u}_1(x(t),t),\ x(t_0) = x_0\}. \tag{2.2}$$

Starting with $M_0 = M(t_0,x_0)$, the solution of this equation along the corresponding trajectory Γ_0 is given by

$$M = e^{tJ/\tau}\left(I + \tau^{-1}J\left(I - e^{tJ/\tau}\right)M_0\right)^{-1}M_0,$$

and a straightforward calculation based on the Cayley–Hamilton theorem (for computing the inverse of a matrix) shows that

$$\max_{t,x}|\nabla\mathbf{u}_1| = \max_{t,x}|M| = \max_{t,x}\left|\frac{\text{polynomial}(\tau, e^{tJ/\tau}, \nabla\mathbf{u}_0)}{(1-\tau\omega_0 - \frac{\tau^2}{2}\eta_0^2)_+}\right|. \tag{2.3}$$

Thus the critical threshold (2.1b) follows. The periodicity of $\mathbf{u}_1$ is proved upon integrating $\mathbf{u}_1{}' = \frac{1}{\tau}Ju$ and $x' = \mathbf{u}_1$ along particle trajectories Γ_0. It turns out both $x(t)$ and $\mathbf{u}_1(t,x(t))$ are $2\pi\tau$-periodic, which clearly implies that $\mathbf{u}_1(t,\cdot)$ shares the same periodicity. It follows that there exists a critical Rossby number, $\tau_c := \tau_c(\nabla\mathbf{u}_0)$, such that the pressureless solution, $\mathbf{u}_1(t,\cdot)$, remains smooth for global time whenever $\tau \in (0,\tau_c)$. This emphasizes the stabilization effect of the rotational forcing for a "large" class of subcritical initial configurations, [17, section 1.2]. Observe that the critical threshold τ_c *need not* be small, and in fact, $\tau_c = \infty$ for rotational initial data such that $\eta_0^2 < 0$, $\omega_0 < \sqrt{-\eta_0^2}$. We shall always limit ourselves, however, to a *finite* value of the critical threshold, τ_c.

In the next corollary we show that, in fact, the pressureless solution retains higher-order smoothness of the subcritical initial data. To this end, we introduce the following notations.

Notations. Here and below, $\|\cdot\|_m$ denotes the usual H^m-Sobolev norm over the two-dimensional torus $\mathbb{T}^2$ and $|\cdot|_\infty$ denotes the L^∞ norm. We abbreviate $a \lesssim_m b$ for $a \le cb$ whenever the constant c depends only on the dimension m. We let $\widehat{C}_0$ denote m-dependent constants that have possible nonlinear dependence on the initial data $|h_0|_\infty$ and $|\nabla\mathbf{u}_0|_\infty$. The constant, $C_0 := \widehat{C}_0\cdot\|(h_0,\mathbf{u}_0)\|_m$, will be used for estimates involving Sobolev regularity, emphasizing that C_0 depends *linearly* on the H^m-size of initial data, h_0 and $\mathbf{u}_0)\|$, and possibly nonlinearly on their L^∞-size.

COROLLARY 2.2. *Fix an integer* $m > 2$ *and consider the pressureless system* (2.1a) *subject to subcritical initial data,* $u_0 \in H^m$. *Then, there exists a critical value* $\tau_c := \tau_c(\nabla\mathbf{u}_0) < \infty$ *such that for* $\tau \in (0,\tau_c]$ *we have, uniformly in time,*

$$|\nabla\mathbf{u}_1(t,\cdot)|_\infty \le \widehat{C}_0, \tag{2.4a}$$

$$\|\mathbf{u}_1(t,\cdot)\|_m \le C_0. \tag{2.4b}$$

Proof. We recall the expression for $|\nabla \mathbf{u}_1|_\infty$ in (2.3). By a continuity argument, there exists a value $\tau_c > 0$ such that $1 - \tau\omega_0 - \frac{\tau^2}{2}\eta_0^2 > 0$ for all $\tau \in (0, \tau_c]$, which in turn implies (2.4a) with a constant $\widehat{C}_0$ that depends on $|\nabla \mathbf{u}_0|_\infty$ and on τ_c.

Having control on the L^∞ norm of $\nabla \mathbf{u}_1$, we employ the standard energy method to obtain the inequality,

$$\frac{d}{dt}\|\mathbf{u}_1(t,\cdot)\|_m \lesssim_m |\nabla \mathbf{u}_1(t,\cdot)|_{L^\infty}\|\mathbf{u}_1(t,\cdot)\|_m.$$

Since $\mathbf{u}_1(t,\cdot)$ is $2\pi\tau$-periodic, it suffices to consider its energy growth over $0 \leq t < 2\pi\tau < 2\pi\tau_c$. Combining with estimate (2.4a) and solving the above Gronwall inequality, we prove the H^m estimate (2.4b). ∎

3. Second approximation—the linearized system. Once we establish the global properties of the pressureless velocity $\mathbf{u}_1$, it can be used as the starting point for a second iteration of (1.3). We begin with the approximate height, h_2, governed by (1.3a),

$$\partial_t h_2 + \mathbf{u}_1 \cdot \nabla h_2 + \left(\frac{1}{\sigma} + h_2\right)\nabla \cdot \mathbf{u}_1 = 0, \qquad h_2(0,\cdot) = h_0(\cdot). \tag{3.1}$$

Recall that $\mathbf{u}_1$ is the solution of the pressureless system (2.1a) subject to *subcritical* initial data $\mathbf{u}_0$, so that $\mathbf{u}_1(t,\cdot)$ is smooth, $2\pi\tau$-periodic in time. The following key lemma shows that the periodicity of $\mathbf{u}_1$ imposes the same periodicity on passive scalars transported by such $\mathbf{u}_1$'s.

LEMMA 3.1. *Let scalar function w be governed by*

$$\partial_t w + \nabla \cdot (\mathbf{u}_1 w) = 0, \tag{3.2}$$

where $\mathbf{u}_1(t,\cdot)$ is a globally smooth, $2\pi\tau$-periodic solution of the pressureless equations (2.1a)*. Then $w(t,\cdot)$ is also $2\pi\tau$-periodic.*

Proof. Let $\phi := \nabla \times \mathbf{u}_1 + \tau^{-1}$ denote the so-called relative vorticity. By (2.1a) it satisfies the same equation that w does, namely,

$$\partial_t \phi + \nabla \cdot (\mathbf{u}_1 \phi) = 0.$$

Coupled with (3.2), it is easy to verify that the ratio w/ϕ satisfies a transport equation

$$\big(\partial_t + \mathbf{u}_1 \cdot \nabla\big)\frac{w}{\phi} = 0,$$

which in turn implies that w/ϕ remains constant along the trajectories Γ_0 in (2.2). But (2.1a) tells us that $\mathbf{u}_1{}' = \frac{J}{\tau}\mathbf{u}_1$, yielding $\mathbf{u}_1(t, x(t)) = e^{\frac{t}{\tau}J}\mathbf{u}_0(x_0)$. We integrate to find $x(2\pi\tau) = x(0)$, namely, the trajectories come back to their initial positions at $t = 2\pi\tau$. Therefore

$$\frac{w}{\phi}(2\pi\tau, x_0) = \frac{w}{\phi}(0, x_0) \quad \text{for all } x_0\text{'s}.$$

Since the above argument is time invariant, it implies that $w/\phi(t,\cdot)$ is $2\pi\tau$-periodic. The conclusion follows from the fact that $\mathbf{u}_1(t,\cdot)$ and thus $\phi(t,\cdot)$ are $2\pi\tau$-periodic. ∎

Equipped with this lemma we conclude the following.

THEOREM 3.2. *Consider the mass equation* (3.1) *on a two-dimensional torus, $\mathbb{T}^2$, linearized around the pressureless velocity field $\mathbf{u}_1$ and subject to subcritical initial data $(h_0, \mathbf{u}_0) \in H^m(\mathbb{T}^2)$ with $m > 5$. It admits a globally smooth solution,*

$h_2(t,\cdot) \in H^{m-1}(\mathbb{T}^2)$, which is $2\pi\tau$-periodic in time, and the following upper bounds hold uniformly in time,

$$|h_2(t,\cdot)|_\infty \leq \widehat{C}_0 \left(1 + \frac{\tau}{\sigma}\right), \tag{3.3a}$$

$$\|h_2(t,\cdot)\|_{m-1} \leq C_0 \left(1 + \frac{\tau}{\sigma}\right). \tag{3.3b}$$

Proof. Apply Lemma 3.1 with $w := \sigma^{-1} + h_2$ to (3.1) to conclude that h_2 is also $2\pi\tau$-periodic. We turn to examining the regularity of h_2. First, its L^∞ bound (3.3a) is studied using the L^∞ estimate for scalar transport equations, which yields an inequality for $|h_2|_\infty = |h_2(t,\cdot)|_\infty$,

$$\frac{d}{dt}|h_2|_\infty \leq |\nabla \cdot \mathbf{u}_1|_\infty (\sigma^{-1} + |h_2|_\infty).$$

Combined with the L^∞ estimate of $\nabla \mathbf{u}_1$ in (2.4a), this Gronwall inequality implies

$$|h_2|_\infty \leq e^{\widehat{C}_0 t}|h_0|_\infty + \frac{1}{\sigma}\left(e^{\widehat{C}_0 t} - 1\right).$$

As before, due to the $2\pi\tau$-periodicity of h_2 and the subcritical condition $\tau \leq \tau_c$, we can replace the first t on the right with τ_c, the second t with $2\pi\tau$, and then (3.3a) follows.

For the H^{m-1} estimate (3.3b), we use the energy method and the Gagliardo–Nirenberg inequality to obtain a similar inequality for $|h_2|_{m-1} = |h_2(t,\cdot)|_{m-1}$,

$$\frac{d}{dt}\|h_2\|_{m-1} \lesssim_m |\nabla \mathbf{u}_1|_\infty \|h_2\|_{m-1} + \left(\frac{1}{\sigma} + |h_2|_\infty\right)\|\mathbf{u}_1\|_m.$$

Applying the estimate on $\mathbf{u}_1$ in (2.4) and the L^∞ estimate on h_2 in (3.3a), we find the above inequality shares a similar form as the previous one. Thus the estimate (3.3b) follows by the same periodicity and subcriticality argument as for (3.3a). We note in passing the *linear* dependence of C_0 on $\|(h_0, \mathbf{u}_0)\|_m$. ∎

To continue with the second approximation, we turn to the approximate momentum equation (1.3b) with $j = 2$,

$$\partial_t \mathbf{u}_2 + \mathbf{u}_2 \cdot \nabla \mathbf{u}_2 + \frac{1}{\sigma}\nabla h_2 - \frac{1}{\tau} J \mathbf{u}_2 = 0. \tag{3.4}$$

The following *splitting* approach will lead to a simplified linearization of (3.4) which is "close" to (3.4) and still maintains the nature of our methodology. The idea is to treat the nonlinear term and the pressure term in (3.4) separately, resulting in two systems for $\widetilde{\mathbf{v}} \approx \mathbf{u}_2$ and $\widehat{\mathbf{v}} \approx \mathbf{u}_2$,

$$\partial_t \widetilde{\mathbf{v}} + \widetilde{\mathbf{v}} \nabla \cdot \widetilde{\mathbf{v}} - \frac{1}{\tau} J \widetilde{\mathbf{v}} = 0, \tag{3.5a}$$

$$\partial_t \widehat{\mathbf{v}} + \frac{1}{\sigma}\nabla h_2 - \frac{1}{\tau} J \widehat{\mathbf{v}} = 0, \tag{3.5b}$$

subject to the same initial data $\widetilde{\mathbf{v}}(0,\cdot) = \widehat{\mathbf{v}}(0,\cdot) = \mathbf{u}_0(\cdot)$.

The first system (3.5a), ignoring the pressure term, is identified as the pressureless system (2.1) and therefore is solved as

$$\widetilde{\mathbf{v}} = \mathbf{u}_1,$$

while the second system (3.5b), ignoring the nonlinear advection term, is solved using Duhamel's principle,

$$\begin{aligned}\widehat{\mathbf{v}}(t,\cdot) &= e^{tJ/\tau}\left(\mathbf{u}_0(t,\cdot) - \int_0^t \frac{e^{-sJ/\sigma}}{\sigma}\nabla h_2(s,\cdot)\,ds\right)\\ &\approx e^{tJ/\tau}\left(\mathbf{u}_0(t,\cdot) - \int_0^t \frac{e^{-sJ/\sigma}}{\sigma}\nabla h_2(t,\cdot)\,ds\right)\\ &= e^{tJ/\tau}\mathbf{u}_0(t,\cdot) + \frac{\tau}{\sigma}J(I - e^{tJ/\tau})\nabla h_2(t,\cdot).\end{aligned}$$

Here, we make an approximation by replacing $h_2(s,\cdot)$ with $h_2(t,\cdot)$ in the integrand, which introduces an error of order τ, taking into account the $2\pi\tau$ periodicity of $h(t,\cdot)$.

Now, synthesizing the two solutions listed above, we make a correction to $\widehat{\mathbf{v}}$ by replacing $e^{tJ/\tau}\mathbf{u}_0$ with $\mathbf{u}_1$. This gives the final form of our approximate velocity field $\mathbf{u}_2$ (with tolerable abuse of notations)

$$\mathbf{u}_2 := \mathbf{u}_1 + \frac{\tau}{\sigma}J(I - e^{tJ/\tau})\nabla h_2(t,\cdot). \tag{3.6a}$$

A straightforward computation shows that this velocity field, $\mathbf{u}_2$, satisfies the approximate momentum equation

$$\partial_t \mathbf{u}_2 + \mathbf{u}_1\cdot\nabla\mathbf{u}_2 + \frac{1}{\sigma}\nabla h_2 - \frac{1}{\tau}{\mathbf{u}_2}^{\perp} = R, \tag{3.6b}$$

where

(3.6c)

$$\begin{aligned} R &:= \frac{\tau}{\sigma}J(I - e^{tJ/\tau})(\partial_t + \mathbf{u}_1\cdot\nabla)\nabla h_2(t,\cdot)\\ \text{(by (3.1))}\qquad &= -\frac{\tau}{\sigma}J(I - e^{tJ/\tau})\left[(\nabla\mathbf{u}_1)^\top\nabla h_2 + \nabla\left(\left(\frac{1}{\sigma} + h_2\right)\nabla\cdot\mathbf{u}_1\right)\right].\end{aligned}$$

Combining Theorem 3.2 on $h_2(t,\cdot)$ with the Gagliardo–Nirenberg inequality, we arrive at the following corollary on periodicity and regularity of $\mathbf{u}_2$.

Corollary 3.3. *Consider the velocity field $\mathbf{u}_2$ in* (3.6) *subject to subcritical initial data $(h_0,\mathbf{u}_0)\in H^m(\mathbb{T}^2)$ with $m>5$. Then, $\mathbf{u}_2(t,\cdot)$ is $2\pi\tau$-periodic in time, and the following upper bound, uniformly in time, holds:*

$$\|\mathbf{u}_2 - \mathbf{u}_1\|_{m-2} \le C_0\frac{\tau}{\sigma}\left(1 + \frac{\tau}{\sigma}\right).$$

In particular, since $\|\mathbf{u}_1\|_m \le C_0$ for subcritical τ, we conclude that $\mathbf{u}_2(t,\cdot)$ has the Sobolev regularity,

$$\|\mathbf{u}_2\|_{m-2} \le C_0\left(1 + \frac{\tau}{\sigma} + \frac{\tau^2}{\sigma^2}\right).$$

We close this section by noting that the second iteration led to an approximate RSW system linearized around the pressureless velocity field $\mathbf{u}_1$, (i.e., system (3.1),(3.6)), which governs our improved, $2\pi\tau$-periodic approximation, $(h_2(t,\cdot),\mathbf{u}_2(t,\cdot))\in H^{m-1}(\mathbb{T}^2)\times H^{m-2}(\mathbb{T}^2)$.

4. Long-time existence of approximate periodic solutions.

4.1. The shallow-water equations. How close is $(h_2(t,\cdot),\mathbf{u}_2(t,\cdot))$ to the exact solution $(h(t,\cdot),\mathbf{u}(t,\cdot))$? Below we shall show that their distance, measured in $H^{m-3}(\mathbb{T}^2)$, does not exceed $\frac{e^{C_0 t}\delta}{1-e^{C_0 t}\delta}$. Thus for sufficiently small δ, the RSW solution $(h,\mathbf{u})$ is *"approximate periodic"* which in turn implies its long-time stability. This is the content of our main result.

THEOREM 4.1. *Consider the RSW equations on a fixed two-dimensional torus,*

$$\partial_t h + \mathbf{u}\cdot\nabla h + \left(\frac{1}{\sigma}+h\right)\nabla\cdot\mathbf{u} = 0, \tag{4.1a}$$

$$\partial_t \mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} + \frac{1}{\sigma}\nabla h - \frac{1}{\tau}J\mathbf{u} = 0 \tag{4.1b}$$

subject to subcritical initial data $(h_0,\mathbf{u}_0)\in H^m(\mathbb{T}^2)$ with $m>5$ and $\alpha_0 := \min(1+\sigma h_0(\cdot))>0$. Let

$$\delta = \frac{\tau}{\sigma^2}$$

denote the ratio between the Rossby number τ and the squared Froude number σ, with subcritical $\tau \le \tau_c(\nabla\mathbf{u}_0)$ so that (2.1b) holds. Assume $\sigma\le 1$ for a substantial amount of pressure forcing in (4.1b). Then, there exists a constant C_0, depending only on m, τ_c, α_0, and in particular depending linearly on $\|(h_0,\mathbf{u}_0)\|_m$, such that the RSW equations admit a smooth, "approximate periodic" solution in the sense that there exists a nearby $2\pi\tau$-periodic solution, $(h_2(t,\cdot),\mathbf{u}_2(t,\cdot))$, such that

$$\|p(t,\cdot)-p_2(t,\cdot)\|_{m-3} + \|\mathbf{u}(t,\cdot)-\mathbf{u}_2(t,\cdot)\|_{m-3} \le \frac{e^{C_0 t}\delta}{1-e^{C_0 t}\delta}. \tag{4.2}$$

Here p is the "normalized height" such that $1+\frac{1}{2}\sigma p = \sqrt{1+\sigma h}$, and correspondingly, p_2 satisfies $1+\frac{1}{2}\sigma p_2 = \sqrt{1+\sigma h_2}$.

It follows that the lifespan of the RSW solution $t \lesssim t_\delta := \ln(\delta^{-1})$ is prolonged due to the rapid rotation $\delta \ll 1$, and in particular, it tends to infinity when $\delta\to 0$.

Proof. We compare the solution of the RSW system (4.1a), (4.1b) with the solution $(h_2,\mathbf{u}_2)$ of approximate RSW system (3.1), (3.6). To this end, we rewrite the latter in the equivalent form,

(4.3a)

$$\partial_t h_2 + \mathbf{u}_2\cdot\nabla h_2 + \left(\frac{1}{\sigma}+h_2\right)\nabla\cdot\mathbf{u}_2 = (\mathbf{u}_2-\mathbf{u}_1)\cdot\nabla h_2 + \left(\frac{1}{\sigma}+h_2\right)\nabla\cdot(\mathbf{u}_2-\mathbf{u}_1),$$

(4.3b)

$$\partial_t \mathbf{u}_2 + \mathbf{u}_2\cdot\nabla\mathbf{u}_2 + \frac{1}{\sigma}\nabla h_2 - \frac{1}{\tau}J\mathbf{u}_2 = (\mathbf{u}_2-\mathbf{u}_1)\cdot\nabla\mathbf{u}_2 + R.$$

The approximate system differs from the exact one, (4.1a), (4.1b), in the residuals on the RHS of (4.3a), (4.3b). We will show that they have an amplitude of order δ. In particular, the comparison in the rotationally dominant regime, $\delta \ll 1$, leads to a long-time existence of a smooth RSW solution, which remains "nearby" the time-periodic solution, $(h_2,\mathbf{u}_2)$. To show that $(h_2,\mathbf{u}_2)$ is indeed an approximate solution for the RSW equations, we proceed as follows.

We first symmetrize both systems so that we can employ the standard energy method for nonlinear hyperbolic systems. To this end, we set the new variable ("normalized height") p such that $1+\frac{1}{2}\sigma p=\sqrt{1+\sigma h}$. Compressing notation with $\mathbf{U}:=(p,\mathbf{u})^\top$, we transform (4.1a), (4.1b) into the *symmetric hyperbolic* quasi-linear system

$$\partial_t\mathbf{U}+B(\mathbf{U},\nabla\mathbf{U})+K[\mathbf{U}]=0. \tag{4.4}$$

Here $B(\mathbf{F},\nabla\mathbf{G}):=A_1(\mathbf{F})\mathbf{G}_x+A_2(\mathbf{F})\mathbf{G}_y$, where A_1,A_2 are bounded linear functions with values being symmetric matrices, and $K[\mathbf{F}]$ is a skew-symmetric linear operator so that $\langle K[\mathbf{F}],\mathbf{F}\rangle=0$. By standard energy arguments, (see, e.g., [12, 13, 15]), the symmetric form of (4.4) yields an exact RSW solution $\mathbf{U}$, which stays smooth for finite time $t\lesssim 1$. The essence of our main theorem is that for small δ's, rotation *prolongs* the lifespan of classical solutions up to $t\sim\mathcal{O}(\ln\delta^{-1})$. To this end, we symmetrize the approximate system (4.3a), (4.3b) using a new variable p_2 such that $1+\frac{1}{2}\sigma p_2=\sqrt{1+\sigma h_2}$. Compressing notation with $\mathbf{U_2}:=(p_2,\mathbf{u}_2)^\top$, we have

$$\partial_t\mathbf{U_2}+B(\mathbf{U_2},\nabla\mathbf{U_2})+K(\mathbf{U_2})=\mathbf{R}, \tag{4.5}$$

where the residual $\mathbf{R}$ is given by

$$\mathbf{R}:=\left[\begin{array}{lll}(\mathbf{u}_2-\mathbf{u}_1)\cdot\nabla p_2 & + & \left(\frac{2}{\sigma}+p_2\right)\nabla\cdot(\mathbf{u}_2-\mathbf{u}_1)\\ (\mathbf{u}_2-\mathbf{u}_1)\cdot\nabla\mathbf{u}_2 & - & R\end{array}\right],$$

with R defined as in (3.6c). We will show $\mathbf{R}$ is small which in turn, using the symmetry of (4.4) and (4.5), will imply that $\|\mathbf{U}-\mathbf{U_2}\|_{m-3}$ is equally small. Indeed, thanks to the fact that $H^{m-3}(\mathbb{T}^2)$ is an algebra for $m>5$, every term in the above expression is upper-bounded in H^{m-3}, by the quadratic products of the terms $\|\mathbf{u}_1\|_m,\|p_2\|_{m-1},\|\mathbf{u}_2\|_{m-2},\|\mathbf{u}_2-\mathbf{u}_1\|_{m-2}$, up to a factor of $\mathcal{O}(1+\frac{1}{\sigma})$. The Sobolev regularity of these terms, $\mathbf{u}_1,\mathbf{u}_2$, and p_2, is guaranteed, respectively, in Corollary 2.2, Corollary 3.3, and Theorem 3.2. Moreover, the nonvacuum condition, $1+\sigma h_0\geq\alpha_0>0$, implies that $1+\sigma h_2$ remains uniformly bounded from below, and by standard arguments (carried out in the appendix), $\|p_2\|_{m-2}\leq C_0(1+\tau/\sigma)$. Summing up, the residual $\mathbf{R}$ does not exceed,

$$\|\mathbf{R}\|_{m-3}\leq C_0^2\left(\delta+\frac{\tau}{\sigma}+\cdots+\frac{\tau^4}{\sigma^4}\right)\lesssim C_0^2\delta \tag{4.6}$$

for subcritical $\tau\in(0,\tau_c)$ and under scaling assumptions $\delta<1$, $\sigma<1$.

We now claim that the same $\mathcal{O}(\delta)$-upperbound holds for the error $\mathbf{E}:=\mathbf{U_2}-\mathbf{U}$, for a long time, $t\lesssim t_\delta$. Indeed, subtracting (4.4) from (4.5), we find the error equation

$$\partial_t\mathbf{E}+B(\mathbf{E},\nabla\mathbf{E})+K[\mathbf{E}]=-B(\mathbf{U_2},\nabla\mathbf{E})-B(\mathbf{E},\nabla\mathbf{U_2})+\mathbf{R}.$$

By the standard energy method using integration by parts and Sobolev inequalities, while utilizing the symmetric structure of B and the skew-symmetry of K, we arrive at

$$\frac{d}{dt}\|\mathbf{E}\|_{m-3}^2\lesssim_m\|\mathbf{E}\|_{m-3}^3+\|\mathbf{U_2}\|_{m-2}\|\mathbf{E}\|_{m-3}^2+\|\mathbf{R}\|_{m-3}\|\mathbf{E}\|_{m-3}.$$

Using the regularity estimates of $\mathbf{U_2}=(p_2,\mathbf{u}_2)^\top$ and the upper bounds on $\mathbf{R}$ in (4.6), we end up with an energy inequality for $\|\mathbf{E}(t,\cdot)\|_{m-3}$,

$$\frac{d}{dt}\|\mathbf{E}\|_{m-3}\lesssim_m\|\mathbf{E}\|_{m-3}^2+C_0\|\mathbf{E}\|_{m-3}+C_0^2\delta,\quad\|\mathbf{E}(0,\cdot)\|_{m-3}=0.$$

A straightforward integration of this forced Riccati equation (consult, for example, [16, section 5]) shows that the error $\|\mathbf{E}\|_{m-3}$ does not exceed

$$\|\mathbf{U}(t,\cdot)-\mathbf{U_2}(t,\cdot)\|_{m-3} \leq \frac{e^{C_0 t}\delta}{1-e^{C_0 t}\delta}. \tag{4.7}$$

In particular, the RSW equations admits an "approximate periodic" $H^{m-3}(\mathbb{T}^2)$-smooth solution for $t \leq \frac{1}{C_0}\ln(\delta^{-1})$ for $\delta \ll 1$. $\square$

Remark 4.1. The estimate on the actual height function h follows by applying the Gagliardo–Nirenberg inequality to $h-h_2 = p(1+\frac{\sigma}{4}p)-p_2(1+\frac{\sigma}{4}p_2) = (p-p_2)(1+\frac{\sigma}{4}(p-p_2)+\frac{\sigma}{2}p_2)$,

$$\|h(t,\cdot)-h_2(t,\cdot)\|_{m-3} \lesssim \frac{e^{C_0 t}\delta}{(1-e^{C_0 t}\delta)^2}.$$

Our result is closely related to observations of near inertial oscillations (NIOs) in oceanography (see, e.g., [22]). These NIOs are mostly seen after a storm blows over the oceans. They exhibit almost periodic dynamics with a period consistent with the Coriolis force and stay stable for about 20 days, which is a long-time scale relative to many oceanic processes such as the storm itself. This observation agrees with our theoretical result regarding the stability and periodicity of RSW solutions. In terms of physical scales, our rotationally dominant condition, $\delta = \frac{gH}{fLU} \ll 1$, provides a physical characterization of this phenomenon. Indeed, NIOs are triggered when storms pass by (large U) and only a thin layer of the oceans is reactive (small aspect ratio H/L). Upon using the multilayer model ([18, section 6.16]), we consider scales $f = 10^{-4}s^{-1}, L = 10^5 m, H = 10^2 m, U = 1ms^{-1}, g = 0.01ms^{-2}$ (reduced gravity due to density stratification—consult [18, section 1.3]). With this parameter setting $\delta = 0.1$, and Theorem 4.1 implies the existence of a smooth, approximate periodic solution over a time scale of $\ln(\delta^{-1})L/U \approx 2$ days. We note in passing that most cyclonic storms on the Northern Hemisphere rotate clockwise, yielding a negative vorticity, $\omega_0 = \partial_y u_0 - \partial_x v_0 < 0$, which is a preferred scenario of the subcritical condition (2.1b) assumed in Theorem 4.1.

4.2. The isentropic gasdynamics. In this section we extend Theorem 4.1 to rotational two-dimensional Euler equations for isentropic gas,

$$\partial_t \rho + \nabla\cdot(\rho\mathbf{u}) = 0, \tag{4.8a}$$

$$\partial_t \mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} + \rho^{-1}\nabla\widetilde{p}(\rho) - f\mathbf{u}^{\perp} = 0. \tag{4.8b}$$

Here, $\mathbf{u} := (u^{(1)}, u^{(2)})^{\top}$ is the velocity field, ρ is the density, and $\widetilde{p} = \widetilde{p}(\rho)$ is the pressure, which for simplicity is taken to be that of a polytropic gas, given by the γ-power law,

$$\widetilde{p}(\rho) = A\rho^{\gamma}. \tag{4.8c}$$

The particular case $A = g/2, \gamma = 2$ corresponds to the RSW equations (1.1a), (1.1b). The following argument for long term existence of the two-dimensional rapidly rotating isentropic equations applies, with minor modifications, to the more general pressure laws, $\widetilde{p}(\rho)$, which induce the hyperbolicity of (4.8a).

We first transform the isentropic Euler equations (4.8a) into their nondimensional form,

$$\partial_t \rho + \mathbf{u} \cdot \nabla \rho + \left(\frac{1}{\sigma} + \rho \right) \nabla \cdot \mathbf{u} = 0,$$
$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \frac{1}{\sigma^2} \nabla (1 + \sigma\rho)^{\gamma - 1} - \frac{1}{\tau} J\mathbf{u} = 0,$$

where the Mach number σ plays the same role as the Froude number in the RSW equation. In order to utilize the technique developed in the previous section, we introduce a new variable h by setting $1+\sigma h = (1+\sigma\rho)^{\gamma-1}$, so that the new variables, $(h, \mathbf{u})$, satisfy

$$\partial_t h + \mathbf{u} \cdot \nabla h + (\gamma - 1) \left(\frac{1}{\sigma} + h \right) \nabla \cdot \mathbf{u} = 0, \tag{4.9a}$$

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \frac{1}{\sigma} \nabla h - \frac{1}{\tau} J\mathbf{u} = 0. \tag{4.9b}$$

This is an analogue to the RSW equations (4.1a), (4.1b), except for the additional factor $(\gamma - 1)$ in the mass equation (4.9a). We can therefore duplicate the steps which led to Theorem 4.1 to obtain a long-time existence for the rotational Euler equations (4.9a), (4.9b). We proceed as follows.

An approximate solution is constructed in two steps. First, we use the $2\pi\tau$-periodic pressureless solution ($h_1 \equiv$ constant, $\mathbf{u}_1(t, \cdot)$) for subcritical initial data $(h_0, \mathbf{u}_0)$. Second, we construct a $2\pi\tau$-periodic solution $(h_2(t, \cdot), \mathbf{u}_2(t\cdot))$ as the solution to an approximate system of the isentropic equations, *linearized* around the pressureless velocity $\mathbf{u}_1$,

$$\partial_t h_2 + \mathbf{u}_1 \cdot \nabla h_2 + (\gamma - 1) \left(\frac{1}{\sigma} + h_2 \right) \nabla \cdot \mathbf{u}_1 = 0,$$
$$\mathbf{u}_2 := \mathbf{u}_1 + \frac{\tau}{\sigma} J \left(I - e^{tJ/\tau} \right) \nabla h_2(t, \cdot).$$

In the final step, we compare $(h, \mathbf{u})$ with the $2\pi\tau$-periodic approximate solution, $(h_2, \mathbf{u}_2)$. To this end, we symmetrize the corresponding systems using $\mathbf{U} = (p, \mathbf{u})^\top$ with the normalized density function p satisfying $1+\frac{1}{2}\sqrt{\frac{1}{\gamma-1}}\sigma p = \sqrt{1+\sigma h}$. Similarly, the approximate system is symmetrized with the variables $\mathbf{U_2} = (p_2, \mathbf{u}_2)$, where $1+\frac{1}{2}\sqrt{\frac{1}{\gamma-1}}\sigma p_2 = \sqrt{1+\sigma h_2}$. We conclude with the following.

THEOREM 4.2. *Consider the rotational isentropic equations on a fixed two-dimensional torus,* (4.9a), (4.9a)*, subject to subcritical initial data* $(\rho_0, \mathbf{u}_0) \in H^m(\mathbb{T}^2)$ *with* $m > 5$ *and* $\alpha_0 := \min(1 + \sigma\rho_0(\cdot)) > 0$.
Let

$$\delta = \frac{\tau}{\sigma^2}$$

denote the ratio between the Rossby and the squared Mach numbers, with subcritical $\tau \le \tau_c(\nabla \mathbf{u}_0)$ *so that* (2.1b) *holds. Assume* $\sigma < 1$ *for a substantial amount of pressure in* (4.9a)*. Then, there exists a constant* C_0*, depending only on* m, $\|(\rho_0, \mathbf{u}_0)\|_m$, τ_c*, and* α_0 *such that the RSW equations admit a smooth, "approximate periodic" solution in the sense that there exists a nearby* $2\pi\tau$*-periodic solution* $(\rho_2(t, \cdot), \mathbf{u}_2(t, \cdot))$ *such that*

$$\|p(t, \cdot) - p_2(t, \cdot)\|_{m-3} + \|\mathbf{u}(t, \cdot) - \mathbf{u}_2(t, \cdot)\|_{m-3} \le \frac{e^{C_0 t}\delta}{1 - e^{C_0 t}\delta}. \tag{4.10}$$

Here, p is the normalized density function satisfying $1+\sigma p = (1+\sigma\rho)^{\frac{\gamma-1}{2}}$, and p_2 results from the same normalization for ρ_2.

It follows that the lifespan of the isentropic solution $t \lesssim t_\delta := 1+\ln(\delta^{-1})$ is prolonged due to the rapid rotation $\delta \ll 1$, and in particular, it tends to infinity when $\delta \to 0$.

Remark 4.2. For the actual density functions, $\rho - \rho_2 = \frac{1}{\sigma}[(1+\sigma p)^{\frac{2}{\gamma-1}} - (1+\sigma p_2)^{\frac{2}{\gamma-1}}] = \int_0^1 C_\gamma [1+\sigma(\theta(p-p_2)+p_2)]^{\frac{2}{\gamma}-1}\, d\theta$,

$$\|\rho(t,\cdot) - \rho_2(t,\cdot)\|_{m-3} \lesssim \frac{e^{C_0 t}\delta}{(1-e^{C_0 t}\delta)^{\frac{2}{\gamma-1}}},$$

in the physically relevant regime $\gamma \in (1,3)$.

4.3. The ideal gasdynamics. We turn our attention to the full Euler equations in the two-dimensional torus,

$$\begin{gathered}\partial_t \rho + \nabla\cdot(\rho\mathbf{u}) = 0,\\ \partial_t \mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} + \rho^{-1}\nabla\widetilde{p}(\rho,S) = fJ\mathbf{u},\\ \partial_t S + \mathbf{u}\cdot\nabla S = 0,\end{gathered}$$

where the pressure law is given as a function of the density, ρ and the specific entropy S, $\widetilde{p}(\rho,S) := \rho^\gamma e^S$. It can be symmetrized by defining a new variable—the "normalized" pressure function,

$$p := \frac{\sqrt{\gamma}}{\gamma-1}\widetilde{p}^{\frac{\gamma-1}{2\gamma}},$$

and by replacing the density equation (4.11a) by a (normalized) pressure equation, so that the above system is recast into an equivalent and symmetric form (see, e.g., [12, 11])

$$\begin{gathered}e^S\partial_t p + e^S\mathbf{u}\cdot\nabla p + C_\gamma e^S p\nabla\cdot\mathbf{u} = 0,\\ \partial_t\mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} + C_\gamma e^S p\nabla p = fJu, \qquad C_\gamma := \frac{\gamma-1}{2},\\ \partial_t S + \mathbf{u}\cdot\nabla S = 0.\end{gathered}$$

It is the exponential function, e^S, involved in *triple* products such as $e^S p\nabla p$, that makes the ideal gas system a nontrivial generalization of the RSW and isentropic gas equations.

We then proceed to the nondimensional form by substitution,

$$\mathbf{u} \to \mathrm{U}u', \quad p \to \mathrm{P}(1+\sigma p'), \quad S = \ln(p\rho^{-\gamma}) \to \ln(\mathrm{P}\mathrm{R}^{-\gamma}) + \sigma S'.$$

After discarding all the primes, we arrive at a nondimensional system

$$e^{\sigma S}\partial_t p + e^{\sigma S}\mathbf{u}\cdot\nabla p + C_\gamma\left(\frac{e^{\sigma S}-1}{\sigma} + e^{\sigma S}p\right)\nabla\cdot\mathbf{u} = -C_\gamma\frac{1}{\sigma}\nabla\cdot\mathbf{u}, \tag{4.11a}$$

$$\partial_t\mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} + C_\gamma\left(\frac{e^{\sigma S}-1}{\sigma} + e^{\sigma S}p\right)\nabla p = -C_\gamma\frac{1}{\sigma}\nabla p + \frac{1}{\tau}Ju, \tag{4.11b}$$

$$\partial_t S + \mathbf{u}\cdot\nabla S = 0, \tag{4.11c}$$

where σ and τ are, respectively, the Mach and the Rossby numbers. With abbreviated notation, $\mathbf{U} := (p, \mathbf{u}, S)^\top$, the equations above amount to a symmetric hyperbolic system written in the compact form,

$$A_0(S)\partial_t \mathbf{U} + A_1(\mathbf{U})\partial_x \mathbf{U} + A_2(\mathbf{U})\partial_y \mathbf{U} = K[\mathbf{U}]. \tag{4.12}$$

Here, $A_i (i = 0, 1, 2)$ are symmetric-matrix-valued functions, *nonlinear* in $\mathbf{U}$, and in particular, A_0 is always positive definite. The linear operator K is skew-symmetric so that $\langle K[\mathbf{U}], \mathbf{U}\rangle = 0$.

Two successive approximations are then constructed based on the iterations (1.3), starting with $j = 1$,

$$\begin{aligned} p_1 &\equiv \text{constant}, \\ \partial_t \mathbf{u}_1 + \mathbf{u}_1 \cdot \nabla \mathbf{u}_1 &= \frac{1}{\tau} J \mathbf{u}_1, \\ S_1 &\equiv \text{constant}. \end{aligned}$$

Identified as the pressureless solution, $\mathbf{u}_1$ is used to linearize the system, resulting in the following approximation:

$$\partial_t p_2 + \mathbf{u}_1 \cdot \nabla p_2 + C_\gamma p_2 \nabla \cdot \mathbf{u}_2 = -C_\gamma \frac{1}{\sigma} \nabla \cdot \mathbf{u}_2, \tag{4.13a}$$

$$\mathbf{u}_2 - \mathbf{u}_1 = \frac{\tau}{\sigma} J(I - e^{tJ/\tau}) C_\gamma e^{\sigma S_2}(1 + \sigma p_2)\nabla p_2, \tag{4.13b}$$

$$\partial_t S_2 + \mathbf{u}_1 \cdot \nabla S_2 = 0. \tag{4.13c}$$

The $2\pi\tau$-periodicity and global regularity of $\mathbf{U}_2 := (p_2, \mathbf{u}_2, S_2)^\top$ follow along the same lines outlined for the RSW equations in section 3 (and therefore omitted), together with the following nonlinear estimate for $e^{\sigma S}$:

$$\|e^{\sigma S} - 1\|_m = \left\| \sum_{j=1}^{\infty} \frac{(\sigma S)^j}{j!} \right\|_m \lesssim_m \sum_{j=1}^{\infty} \frac{(C_m |\sigma S|_\infty)^{j-1}}{j!} \|\sigma S\|_m = \frac{e^{C_m|\sigma S|_\infty} - 1}{C_m |\sigma S|_\infty} \|\sigma S\|_m;$$

for the latter, we apply recursively the Gagliardo–Nirenberg inequality to typical terms $\|(\sigma S)^j\|_m$. Notice the entropy variables (both the exact and approximate ones) satisfy a transport equation and therefore are conserved along particle trajectories, which implies that the L^∞ norm of the entropy variable is an invariant. Thus we arrive at an estimate

$$\|e^{\sigma S} - 1\|_m \leq \sigma \widehat{C}_0 \|S\|_m. \tag{4.14}$$

Of course, the same type of estimate holds for the approximate entropy, S_2.

Finally, we subtract the approximate system (4.13) from the exact system (4.12), arriving at an error equation for $\mathbf{E} := \mathbf{U} - \mathbf{U}_2$ that shares the same form as the RSW system in section 4.1, except that $A_i(\mathbf{U}) - A_i(\mathbf{U}_2) \neq A_i(\mathbf{U} - \mathbf{U}_2)$ due to nonlinearity, which is essentially quadratic in the sense that [1]

$$\begin{aligned} \|A_i(\mathbf{U}) - A_i(\mathbf{U}_2)\|_n &\lesssim \|\mathbf{U} - \mathbf{U}_2\|_n^2 + \|\mathbf{U} - \mathbf{U}_2\|_n, \qquad i = 0, 1, 2, \\ A_i(\mathbf{U}) - A_i(\mathbf{U}_2)\|_{W^{1,\infty}} &\lesssim \|\mathbf{U} - \mathbf{U}_2\|_{W^{1,\infty}}^2 + \|\mathbf{U} - \mathbf{U}_2\|_{W^{1,\infty}}, \qquad i = 0, 1, 2. \end{aligned}$$

[1] Consider a typical term of A_i, e.g., $e^{\sigma S} p$. Applying (4.14) together with the Gagliardo–Nirenberg inequality to $e^{\sigma S} - e^{\sigma S_2} = e^{\sigma S_2}(e^{\sigma(S - S_2) - 1})$, we can show $\|e^{\sigma S} - e^{\sigma S_2}\|_n \lesssim \|S - S_2\|_n$. The estimate on $\|e^{\sigma S} p - e^{\sigma S_2} p_2\|_n$ then follows by applying identity $ab - a_2 b_2 = (a - a_2)(b - b_2) + (a - a_2) b_2 + a_2(b - b_2)$ together with the triangle inequality and the Gagliardo–Nirenberg inequality. Here regularity of S_2 and p_2 is a priori known.

where $n > 2$. This additional nonlinearity manifests itself as three more multiplications in the energy inequality,

$$\frac{d}{dt}\|\mathbf{E}\|_{m-3} \lesssim \|\mathbf{E}\|_{m-3}^5 + \cdots + \|\mathbf{E}\|_{m-3} + \delta, \qquad \|\mathbf{E}(0,\cdot)\|_{m-3} = 0,$$

whose solution (developed around a simple root of the quintic polynomial on the right) has the same asymptotic behavior as for the quadratic Riccati equations derived in the previous sections.

THEOREM 4.3. *Consider the (symmetrized) rotational Euler equations on a fixed two-dimensional torus* (4.11) *subject to subcritical initial data* $(p_0, \mathbf{u}_0, S_0) \in H^m(\mathbb{T}^2)$ *with* $m > 5$. *Let*

$$\delta = \frac{\tau}{\sigma^2}$$

denote the ratio between the Rossby and the squared Mach numbers, with subcritical $\tau \le \tau_c(\nabla \mathbf{u}_0)$ *so that* (2.1b) *holds. Assume* $\sigma < 1$ *for a substantial amount of pressure forcing in* (4.11b). *Then, there exists a constant* C, *depending only on* m, $\|(p_0, \mathbf{u}_0, S_0)\|_m$, τ_c, *such that the ideal gas equations admit a smooth, "approximate periodic" solution in the sense that there exists a nearby* $2\pi\tau$*-periodic solution,* $(p_2(t,\cdot), \mathbf{u}_2(t,\cdot), S_2(t,\cdot))$ *such that*

(4.15)

$$\|p(t,\cdot) - p_2(t,\cdot)\|_{m-3} + \|\mathbf{u}(t,\cdot) - \mathbf{u}_2(t,\cdot)\|_{m-3} + \|S(t,\cdot) - S_2(t,\cdot)\|_{m-3} \le \frac{e^{C_0 t}\delta}{1 - e^{C_0 t}\delta}.$$

It follows that the lifespan of the ideal gas solution, $t \lesssim t_\delta := \ln(\delta^{-1})$, *is prolonged due to the rapid rotation* $\delta \ll 1$, *and in particular, it tends to infinity when* $\delta \to 0$.

5. Appendix. Staying away from vacuum. We will show the following proposition on the new variable p_2 defined in section 4.1.

PROPOSITION 5.1. *Let* p_2 *satisfy*

$$1 + \frac{1}{2}\sigma p_2 = \sqrt{1 + \sigma h_2}, \tag{5.1}$$

where h_2 *is defined as in* (3.1), *that is,*

$$\partial_t h_2 + \mathbf{u}_1 \cdot \nabla h_2 + \left(\frac{1}{\sigma} + h_2\right) \nabla \cdot \mathbf{u}_1 = 0 \tag{5.2}$$

subject to initial data $h_2(0,\cdot) = h_0(\cdot)$ *that satisfies the nonvacuum condition* $1 + \sigma h_0(\cdot) \ge \alpha_0 > 0$. *Then,*

$$|p_2|_\infty \le \widehat{C}_0 \left(1 + \frac{\tau}{\sigma}\right),$$

$$\|p_2\|_n \le C_0 \left(1 + \frac{\tau}{\sigma}\right).$$

The proof of this proposition follows in two steps. First, we show that the L^∞ and H^n norms of $p_2(0,\cdot)$ are dominated by $h_2(0,\cdot)$ due to the nonvacuum condition. Second, we derive the equation for p_2 and obtain regularity estimates using similar techniques from section 4.1.

Step 1. For simplicity, we use $p := p_2(0,\cdot)$ and $h := h_2(0,\cdot)$.
Differentiation of (5.1) yields

$$p = \frac{2h}{\sqrt{1+\sigma h}+1}, \quad \nabla p = \frac{\nabla h}{\sqrt{1+\sigma h}}.$$

Clearly, $|p|_\infty \le |h|_\infty$. The above identities, together with the nonvacuum condition, imply

$$\|p\|_1 \le 2\|h\|_1 \quad \text{and} \quad |\nabla p|_{L^\infty} \le \frac{|\nabla h|_{L^\infty}}{\sqrt{\alpha_0}}.$$

For higher derivatives of p, we use the following recursive relation. Rewrite (5.1) as $p + \frac{1}{4}\sigma p^2 = h$ and then take the kth derivative on both sides,

$$D^k p + \frac{1}{4}\sigma 2pD^k p + \frac{1}{4}\sigma\left(D^k(q^2) - 2pD^k p\right) = D^k h$$

so that taking the L^2 norm of this equation yields

$$I - II := \left\|\left(1+\frac{1}{2}\sigma p\right)D^k p\right\|_0 - \frac{1}{4}\sigma\left\|D^k(q^2) - 2pD^k p\right\|_0 \le \|D^k h\|_0.$$

Furthermore, we find $I \ge \sqrt{\alpha_0}\|D^k p\|_0$ by (5.1) and the nonvacuum condition. We also find $II \lesssim_n |\nabla p|_\infty \|p\|_{|k|-1}$ by the Gagliardo–Nirenberg inequalities. Thus we arrive at a recursive relation

$$\|p\|_{|k|} \le \widehat{C}_0(\|p\|_{|k|-1} + \|h\|_{|k|}),$$

which implies that the H^n norm of $p_2(0,\cdot) = p$ is dominated by $\|h_2(0,\cdot)\|_n = \|h\|_n$.

Step 2. We derive an equation for p_2 using relation (5.1) and equation (5.2),

$$\partial_t p_2 + 2\mathbf{u}_1\cdot\nabla p_2 + \left(\frac{1}{\sigma} + p_2\right)\nabla\cdot\mathbf{u}_1 = 0.$$

This equation resembles the formality of the approximate mass equation (3.1) for h_2, and thus we apply a similar technique to arrive at the same regularity estimate for p_2,

$$|p_2(t,\cdot)|_\infty \le \widehat{C}_0\left(1+\frac{\tau}{\sigma}\right),$$
$$\|p_2(t,\cdot)\|_n \le C_0\left(1+\frac{\tau}{\sigma}\right).$$

REFERENCES

[1] A. Babin, A. Mahalov, and B. Nicolaenko, *Resonances and regularity for Boussinesq equations*, Russian J. Math. Phys., 4 (1996), pp. 417–428.

[2] A. Babin, A. Mahalov, B. Nicolaenko, and Y. Zhou, *On the asymptotic regimes and the strongly stratified limit of rotating Boussinesq equations*, Theor. Comput. Fluid Dyn., 9 (1997), pp. 223–251.

[3] A. Babin, A. Mahalov, and B. Nicolaenko, *On the nonlinear baroclinic waves and adjustment of pancake dynamics*, Theor. Comput. Fluid Dyn., 11 (1998), pp. 215–235.

[4] A. Babin, A. Mahalov, and B. Nicolaenko, *Fast singular oscillating limits and global regularity for the 3D primitive equations of geophysics*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 201–222.

[5] A. Babin, A. Mahalov, and B. Nicolaenko, *Fast singular oscillating limits of stably-stratified 3D Euler and Navier-Stokes equations and geostrophic wave fronts*, in Large-Scale Atmosphere-Ocean Dynamics, Vol. I, Cambridge University Press, Cambridge, 2002, pp. 126–201.

[6] J.-Y. Chemin, B. Desjardins, I. Gallagher, and E. Grenier, *Mathematical Geophysics*, Oxford Lecture Ser., Math. Appl. 32, The Clarendon Press, Oxford University Press, Oxford, UK, 2006.

[7] P. Embid and A. Majda, *Averaging over fast gravity waves for geophysical flows with arbitrary potential vorticity*, Comm. Partial Differential Equations, 21 (1996), pp. 619–658.

[8] P. Embid and A. Majda, *A low Froude number limiting dynamics for stably stratified flow with small or finite Rossby numbers*, Geophys. Astrophys. Fluid Dyn., 87 (1998), pp. 1–50.

[9] I. Gallagher, *Asymptotic of the solutions of hyperbolic equations with a skew-symmetric perturbation*, J. Differential Equations, 150 (1998), pp. 363–384.

[10] M. Ghil and S. Childress, *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory, and Climate Dynamics*, Appl. Math. Sci. 60, Springer-Verlag, New York, 1987.

[11] M. Grassin, *Global smooth solutions to Euler equations for a perfect gas*, Indiana Univ. Math. J., 47 (1998), pp. 1397–1432

[12] T. Kato, *The Cauchy problem for quasi-linear symmetric hyperbolic systems*, Arch. Ration. Mech. Anal., 58 (1975), pp. 181–205.

[13] S. Klainerman and A. Majda, *Compressible and incompressible fluids*, Comm. Pure Appl. Math., 35 (1982), pp. 629–651.

[14] H.-O. Kreiss, *Problems with different time scales for partial differential equations*, Comm. Pure Appl. Math., 33 (1980), pp. 399–439.

[15] H.-O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier–Stokes Equations*, Reprint of the 1989 edition, Classics Appl. Math. 47, SIAM, Philadelphia, 2004.

[16] H. Liu and E. Tadmor, *Semi-classical limit of the nonlinear Schrödinger-Poisson equation with sub-critical initial data*, Methods Appl. Anal., 9 (2002), pp. 517–532.

[17] H. Liu and E. Tadmor, *Rotation prevents finite-time breakdown*, Phys. D, 188 (2004), pp. 262–276.

[18] J. Pedlosky, *Geophysical Fluid Dynamics*, Springer-Verlag, Berlin, 1992.

[19] S. Schochet, *Fast singular limits of hyperbolic PDEs*, J. Differential Equations, 114 (1994), pp. 476–512.

[20] E. Tadmor, *Hyperbolic systems with different time scales*, Comm. Pure Appl. Math., 35 (1982), pp. 839–866.

[21] E. Tadmor and D.-M. Wei, *On the global regularity of subcritical Euler-Poisson equations with pressure*, J. Eur. Math. Soc. (JEMS), to appear.

[22] W. R. Youn and M. Ben Jelloul, *Propagation of near-inertial oscillations through a geostrophic flow*, J. Marine Research, 55 (1997), pp. 735–766.

[23] V. Zeitlin, G. M. Reznik, and M. Ben Jelloul, *Nonlinear theory of geostrophic adjustment.* I. *Rotating shallow-water model*, J. Fluid Mech., 445 (2001), pp. 93–120.

[24] V. Zeitlin, G. M. Reznik, and M. Ben Jelloul, *Nonlinear theory of geostrophic adjustment.* II. *Two-layer and continuously stratified primitive equations*, J. Fluid Mech., 491 (2003), pp. 207–228.

SIAM J. MATH. ANAL.
Vol. 39, No. 5, pp. 1686–1692

ON SOME PROPERTIES OF TRAVELING WATER WAVES WITH VORTICITY*

EUGEN VARVARUCA†

Abstract. We prove that for a large class of vorticity functions the crests of any corresponding traveling gravity water wave of finite depth are necessarily points of maximal horizontal velocity. We also show that for waves with nonpositive vorticity the pressure everywhere in the fluid is larger than the atmospheric pressure. A related a priori estimate for waves with nonnegative vorticity is also given.

Key words. water waves, vorticity, maximum principle

AMS subject classifications. 76B15, 35R35, 35J65

DOI. 10.1137/070697513

1. Introduction.

In this article we consider the classical hydrodynamical problem concerning traveling two-dimensional gravity water waves with vorticity. This problem has attracted considerable interest in recent years, starting with the systematic study of Constantin and Strauss [8] on periodic waves of finite depth.

The problem arises from the following physical situation. A wave of permanent form moves with constant speed on the surface of an incompressible flow, the bottom of the flow domain being horizontal. With respect to a frame of reference moving with the speed of the wave, the flow is steady and occupies a fixed region Ω in the upper half of the (x, y)-plane, which lies between the real axis $\mathcal{B} := \{(x, 0) : x \in \mathbb{R}\}$ and some a priori unknown free surface $\mathcal{S} := \{(x, \eta(x)) : x \in \mathbb{R}\}$, where η is a periodic function. Since the fluid is incompressible, the flow can be described by a (relative) stream function ψ which is periodic in the horizontal direction and satisfies the following equations and boundary conditions:

$$\Delta\psi = -\gamma(\psi) \quad \text{in } \Omega, \tag{1.1a}$$

$$\psi = B \quad \text{on } \mathcal{B}, \tag{1.1b}$$

$$\psi = 0 \quad \text{on } \mathcal{S}, \tag{1.1c}$$

$$|\nabla\psi|^2 + 2gy = Q \quad \text{on } \mathcal{S}, \tag{1.1d}$$

$$\psi_y < 0 \quad \text{in } \Omega, \tag{1.1e}$$

where B, g, and Q are positive constants. Equation (1.1a) involves a vorticity function $\gamma : [0, B] \to \mathbb{R}$ and expresses the fact that the vorticity of the flow $\omega := -\Delta\psi$ and the stream function ψ are functionally dependent. Equations (1.1b) and (1.1c) mean that the bottom and the free surface are streamlines, while (1.1d) means that the pressure at the surface of the flow is a constant. The relative velocity of the fluid particles is given by $(\psi_y, -\psi_x)$. The requirement (1.1e) means that the horizontal velocity of each fluid particle is smaller than the speed of the wave and is motivated both by field observations and by laboratory experiments; see [8] for references. It is customary [8]

*Received by the editors July 17, 2007; accepted for publication (in revised form) October 9, 2007; published electronically January 30, 2008. This work was supported by the EPSRC grant EP/D505402/1.
http://www.siam.org/journals/sima/39-5/69751.html

†Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom (mapev@maths.bath.ac.uk).

to assume that the constants g, B and the vorticity function γ are given. The problem consists in determining the curves $\mathcal{S}$ for which there exists a function ψ in Ω which satisfies (1.1a)–(1.1e) for some value of the parameter Q. For a full justification of the equivalence between the problem of seeking solution triples $(\mathcal{S}, \psi, Q)$ of (1.1) and that of seeking traveling-wave solutions of the two-dimensional Euler equations, the reader is referred to [8].

When $\gamma \equiv 0$, the corresponding flow is called irrotational. Nowadays the mathematical theory dealing with this situation contains a wealth of results, mostly obtained during the last three decades, concerning the existence of large-amplitude solutions and their properties. Global bifurcation theories were given for various types of waves (periodic or solitary of finite depth; periodic of infinite depth) by Keady and Norbury [14] and by Amick and Toland [1, 2]. Moreover, it was shown by Toland [20] and McLeod [16] that in the closure of these continua of solutions there exist waves with stagnation points at their crests, a stagnation point being one at which the relative fluid velocity is zero, i.e., $|\nabla\psi| = 0$. The existence of such waves, called "extreme waves," was predicted by Stokes [18], who also conjectured that their profiles necessarily have corners with included angle of 120° at the crests. This conjecture was proved independently by Amick, Fraenkel, and Toland [3] and by Plotnikov [17]. Recently, the method of [3] was simplified and extended in [22].

On the other hand, when $\gamma \not\equiv 0$, the flow is called rotational or with vorticity, and significant advances in the corresponding mathematical theory have been made only in the last few years. The existence of global continua of smooth solutions was proved by Constantin and Strauss [8] for the periodic finite depth problem, and by Hur [13] for the related problem of periodic waves of infinite depth. For the solutions found in [8, 13] the wave profiles have exactly one crest and one trough per minimal period, are monotone between crests and troughs, and have a vertical axis of symmetry. (The symmetry assumption is in fact not a restriction since, for any vorticity function, any wave profile with the above monotonicity properties is necessarily symmetric [5, 11].) Of particular significance is the fact that the continuum of solutions found in [8] contains waves for which the values of $\max_{\overline{\Omega}} \psi_y$ are arbitrarily close to 0. Thus it is natural to expect that, as in the irrotational case, waves with stagnation points, referred to above as "extreme waves," exist for many vorticity functions, and that they can be obtained as limits, in a suitable sense, of certain sequences of regular waves found in [8]. In the case of constant vorticity, numerical evidence [15, 19] strongly points to the existence of extreme waves for any negative vorticity and for small positive vorticity, and also indicates that, for large positive vorticity, continua of solutions bifurcating from a line of trivial solutions develop into overhanging profiles (a situation which is not possible in the irrotational case; see [23] for references) and do not approach extreme waves. Further references to numerical investigations of waves with vorticity can be found in [15].

One of the questions addressed in this article concerns the location of the points at which the maximum over $\overline{\Omega}$ of the relative horizontal velocity ψ_y is attained for smooth waves with vorticity. In the irrotational case, the crests of the wave are the only such points; see Toland [21]. Very recently, Constantin and Strauss [9, Theorem 4.1] showed that this is also the case for the waves in the continuum in [8] under the assumption that γ is a nonpositive constant which satisfies a smallness condition involving B and g. Here we prove, with a novel approach, a slightly weaker result under substantially more general assumptions. Namely, for wave profiles with finitely many local extrema on a period, if the vorticity function γ satisfies $\gamma \leq 0$ and $\gamma' \geq 0$ everywhere on $[0, B]$, then any point of maximal relative horizontal velocity must lie on the free surface and the crests are necessarily such points. An immediate consequence of this result is that, whenever $\gamma \leq 0$ and $\gamma' \geq 0$, the continuum of solutions in [8] contains waves for which the values of $|\nabla\psi|$ at their crests are arbitrarily close to 0. Thus in this case the existence of waves with stagnation points at their crests is to be particularly expected.

Another contribution of this article is that we establish some new a priori bounds for waves corresponding to vorticity functions γ which do not change sign, without any assumptions on γ'. When $\gamma \leq 0$, the estimate in question means that the pressure everywhere in the fluid is larger than the atmospheric pressure. This estimate is the main ingredient in the proof of the previously mentioned result concerning the location of the points where $\max_{\overline{\Omega}} \psi_y$ is attained. When $\gamma \geq 0$, a slightly different, but related, estimate is given. Both these estimates play an essential role in the investigation in [24] concerning the existence of extreme waves with vorticity and the Stokes conjecture.

The proofs here are based on simple applications of the maximum principle [12, Chapters 2 and 3]. Analogous results to those of this article hold in the case of periodic rotational waves of infinite depth. They will be presented, together with some applications, in a subsequent article.

Of the many other directions in which the theory of traveling gravity water waves, with or without vorticity, has seen recent progress and is currently being further developed, we mention here only a few: variational formulations [7], stability [10], and properties of the fluid particle trajectories [4, 6].

2. The main results. We always deal with classical solutions of (1.1), in the sense that $\gamma \in C^1([0,B])$, $\eta \in C^3(\mathbb{R})$, $\psi \in C^3(\overline{\Omega})$. We assume that η is a periodic function of minimal period $2L$, and that ψ is $2L$-periodic in the horizontal direction. However, we do not assume that η has exactly one local maximum and one local minimum per minimal period.

Let $\hat{\Gamma} : [0,B] \to \mathbb{R}$ be given by

$$\hat{\Gamma}(s) = \int_0^s \gamma(t)\,dt \quad \text{for all } s \in [0,B]. \tag{2.1}$$

(Note that in [8] a function Γ is considered which is related to $\hat{\Gamma}$ by $\hat{\Gamma}(s) = -\Gamma(-s)$. The quantity of interest both here and there is $\hat{\Gamma}(\psi)$, which is denoted there by $-\Gamma(-\psi)$; we find our notation more convenient.) Let us also consider the function $R : \overline{\Omega} \to \mathbb{R}$ given by

$$R = \frac{1}{2}|\nabla\psi|^2 + gy - \frac{1}{2}Q + \hat{\Gamma}(\psi). \tag{2.2}$$

The function R is (up to a constant) the negative of the pressure in the fluid domain; see [8].

Our next result shows that when γ is everywhere nonpositive the pressure in the fluid domain is larger than the atmospheric pressure.

Theorem 2.1. *Suppose that $\gamma(s) \leq 0$ for all $s \in [0,B]$. Then $R \leq 0$ in $\overline{\Omega}$.*

Remark 2.2. Under the much more restrictive assumptions that

$$\gamma \leq 0, \quad \gamma' \leq 0 \quad \text{and} \quad -\psi_y(x,0)\gamma(B) \geq -g \text{ for all } x \in \mathbb{R},$$

the conclusion of Theorem 2.1 was previously obtained in [9, Example 3.1].

The importance of the inequality $R \le 0$ in $\overline{\Omega}$ in relation to the monotonicity of ψ_y along the free surface $\mathcal{S}$ was first recognized for waves with vorticity by Constantin and Strauss [9, Proposition 3.4]. We give here a slightly more general statement of their result and a somewhat more direct proof.

THEOREM 2.3. *Let $\eta : \mathbb{R} \to \mathbb{R}$ be such that there exists $N \in \mathbb{N}$ and points $x_0 < x_1 < \cdots < x_{2N} = x_0 + 2L$ with the property that $\eta'(x_j) = 0$ for all $j \in \{0, \ldots, 2N\}$, η is strictly increasing on $[x_{2j}, x_{2j+1}]$ for all $j \in \{0, \ldots, N-1\}$, and η is strictly decreasing on $[x_{2j-1}, x_{2j}]$ for all $j \in \{1, \ldots, N\}$. Suppose that $R \le 0$ in $\overline{\Omega}$. Then the function $x \mapsto \psi_y(x, \eta(x))$ is increasing on $[x_{2j}, x_{2j+1}]$ for all $j \in \{0, \ldots, N-1\}$ and decreasing on $[x_{2j-1}, x_{2j}]$ for all $j \in \{1, \ldots, N\}$. Therefore, $\max_{\mathcal{S}} \psi_y$ is attained at the points of maximal height on $\mathcal{S}$.*

The preceding result leads with little effort to one concerning the location of the points where $\max_{\overline{\Omega}} \psi_y$ is attained.

THEOREM 2.4. *Let $\eta : \mathbb{R} \to \mathbb{R}$ be as in Theorem 2.3. Suppose that $\gamma(s) \le 0$ and $\gamma'(s) \ge 0$ for all $s \in [0, B]$. Then any point at which $\max_{\overline{\Omega}} \psi_y$ is attained lies on $\mathcal{S}$, and the crests of the wave are necessarily such points.*

Remark 2.5. For a more restrictive class of wave profiles and under the assumption that γ is a nonpositive constant which satisfies

$$g^2 \ge 2g(-2B\gamma^3)^{1/2} - 2B\gamma^3,$$

Constantin and Strauss [9, Theorem 4.1] proved that the crests of the wave are the only points at which $\max_{\overline{\Omega}} \psi_y$ is attained. This slightly stronger conclusion does not seem to be readily obtainable by the methods used in the proof of Theorem 2.4.

The next result gives a new estimate in the case when γ is everywhere nonnegative, which is in the same spirit as that of Theorem 2.1. Let us consider the function $T : \overline{\Omega} \to \mathbb{R}$ given by

$$T := R - \varpi\psi, \tag{2.3}$$

where R is given by (2.2) and

$$\varpi := \frac{1}{2} \max_{s \in [0,B]} \gamma(s). \tag{2.4}$$

THEOREM 2.6. *Suppose that $\gamma(s) \ge 0$ for all $s \in [0, B]$. Then $T \le 0$ in $\overline{\Omega}$.*

3. Proofs of the main results. A simple calculation shows that, everywhere in $\overline{\Omega}$,

$$R_x = \psi_y\psi_{xy} - \psi_x\psi_{yy}, \tag{3.1a}$$
$$R_y = \psi_x\psi_{xy} - \psi_y\psi_{xx} + g, \tag{3.1b}$$

and

$$\Delta R = 2\psi_{xy}^2 - 2\psi_{xx}\psi_{yy}. \tag{3.2}$$

Proof of Theorem 2.1. Note that $R = 0$ everywhere on the free surface $\mathcal{S}$. We claim that the maximum of R over $\overline{\Omega}$ must be attained on $\mathcal{S}$.

Observe first that, since $R_y = g > 0$ everywhere on the bottom $\mathcal{B}$, $\max_{\overline{\Omega}} R$ cannot be attained anywhere on $\mathcal{B}$.

Suppose now that $\max_{\overline{\Omega}} R$ is attained at some point A in Ω. Then necessarily

$$R_x(A) = 0, \quad R_y(A) = 0, \quad \Delta R(A) \leq 0.$$

It follows from this, (3.1), and (3.2) that

$$\psi_y(A)\psi_{xy}(A) = \psi_x(A)\psi_{yy}(A), \tag{3.3a}$$

$$\psi_x(A)\psi_{xy}(A) < \psi_y(A)\psi_{xx}(A), \tag{3.3b}$$

$$\psi_{xy}^2(A) \leq \psi_{xx}(A)\psi_{yy}(A). \tag{3.3c}$$

Since (1.1e) holds, it follows that $\psi_y(A) < 0$. We now distinguish two cases, depending on whether or not $\psi_{yy}(A) = 0$.

If $\psi_{yy}(A) = 0$, then (3.3a) implies that $\psi_{xy}(A) = 0$. It then follows from (3.3b) that $\psi_{xx}(A) < 0$, and hence $\gamma(\psi(A)) = -\Delta\psi(A) > 0$, which contradicts the assumption that $\gamma(s) \leq 0$ for all $s \in [0, B]$.

If $\psi_{yy}(A) \neq 0$, then it follows from (3.3a) and (3.3b) that

$$\frac{\psi_y(A)\psi_{xy}^2(A)}{\psi_{yy}(A)} < \psi_y(A)\psi_{xx}(A).$$

It then follows from this and (3.3c) that $\psi_{yy}(A) < 0$. We now deduce from (3.3c) that $\psi_{xx}(A) \leq 0$, and therefore $\gamma(\psi(A)) = -\Delta\psi(A) > 0$, which again contradicts the assumption that $\gamma(s) \leq 0$ for all $s \in [0, B]$.

We conclude that the maximum of R over $\overline{\Omega}$ must be attained on $\mathcal{S}$, which implies that $R \leq 0$ in $\overline{\Omega}$. This completes the proof of Theorem 2.1. □

Proof of Theorem 2.3. The proof is based on a remarkable, though straightforward to verify, identity observed by Toland [21] in the irrotational case and by Constantin and Strauss [9] in the general case:

$$\frac{d}{dx}\left[\frac{1}{2}\psi_y^2(x, \eta(x))\right] = R_x(x, \eta(x)) \quad \text{for all } x \in \mathbb{R}. \tag{3.4}$$

Since $R \leq 0$ in $\overline{\Omega}$ and $R = 0$ on $\mathcal{S}$, the required result concerning the monotonicity of $x \mapsto \psi_y(x, \eta(x))$ is immediate from (3.4). It follows that

$$\max_{\mathcal{S}} \psi_y = \max_{j \in \{0,\dots,N-1\}} \psi_y(x_{2j+1}, \eta(x_{2j+1})). \tag{3.5}$$

But for every $j \in \{0, \dots, 2N\}$, $\psi_x(x_j, \eta(x_j)) = 0$ and therefore

$$\psi_y(x_j, \eta(x_j)) = -(Q - 2g\eta(x_j))^{1/2}.$$

Hence $\max_{\mathcal{S}} \psi_y$ is attained at the points of maximal height on $\mathcal{S}$. This completes the proof of Theorem 2.3. □

Proof of Theorem 2.4. It follows from (1.1a) that

$$\Delta\psi_y = -\gamma'(\psi)\psi_y \quad \text{in } \Omega.$$

Since $\psi_y < 0$ in Ω and $\gamma'(s) \geq 0$ for all $s \in [0, B]$, it follows immediately from the maximum principle that $\max_{\overline{\Omega}} \psi_y$ cannot be attained anywhere in Ω.

We now show that $\max_{\overline{\Omega}} \psi_y$ cannot be attained anywhere on $\mathcal{B}$ either. This is trivial when $\gamma(B) < 0$, since then $\psi_{yy} = -\gamma(B) > 0$ everywhere on $\mathcal{B}$. When

$\gamma(B) = 0$, we use a reflection argument. Let $\tilde{\gamma} : [0, 2B] \to \mathbb{R}$ be an extension of γ such that $\tilde{\gamma}(s) = -\gamma(2B - s)$ for all $s \in (B, 2B]$. Let Ω^R be the reflection of Ω into $\mathcal{B}$,

$$\widetilde{\Omega} := \Omega \cup \mathcal{B} \cup \Omega^R,$$

and $\tilde{\psi} : \widetilde{\Omega} \to \mathbb{R}$ be an extension of ψ such that $\tilde{\psi}(x, y) = 2B - \psi(x, -y)$ for all $(x, y) \in \Omega^R$. Then it is easily checked that $\tilde{\gamma} \in C^1([0, 2B])$, $\tilde{\psi} \in C^3(\widetilde{\Omega})$ and

$$\Delta\tilde{\psi} = -\tilde{\gamma}(\tilde{\psi}) \quad \text{in } \widetilde{\Omega}.$$

Since $\tilde{\psi}_y < 0$ in $\widetilde{\Omega}$ and $\tilde{\gamma}'(s) \geq 0$ for all $s \in [0, 2B]$, the maximum principle yields the required result.

We conclude that $\max_{\overline{\Omega}} \psi_y$ can only be attained on $\mathcal{S}$. Next note that, since $\gamma(s) \leq 0$ for all $s \in [0, B]$, Theorem 2.1 shows that $R \leq 0$ in $\overline{\Omega}$. An application of Theorem 2.3 now yields that $\max_{\overline{\Omega}} \psi_y$ is attained at the crests of the wave. This completes the proof of Theorem 2.4. ∎

Proof of Theorem 2.6. Note first that $T_y = g - \varpi\psi_y > 0$ everywhere on $\mathcal{B}$, so that the maximum of T over $\overline{\Omega}$ cannot be attained anywhere on $\mathcal{B}$.

Next note from (3.2) that

$$\Delta R \geq -\frac{1}{2}\gamma^2(\psi) \quad \text{in } \Omega.$$

Since

$$\Delta T = \Delta R + \varpi\gamma(\psi),$$

it is immediate, upon using (2.4) and the assumption that $\gamma(s) \geq 0$ for all $s \in [0, B]$, that T is a subharmonic function in Ω. Therefore, the maximum of T over $\overline{\Omega}$ cannot be attained anywhere in Ω.

We conclude that $\max_{\overline{\Omega}} T$ must be attained somewhere on $\mathcal{S}$. Since $T = 0$ everywhere on $\mathcal{S}$, it follows that $T \leq 0$ in $\overline{\Omega}$. This completes the proof of Theorem 2.6. ∎

Acknowledgments. I am grateful to John Toland and Markus Lilli for many useful discussions, and to the anonymous referees for suggestions which helped improve the presentation.

REFERENCES

[1] C. J. Amick and J. F. Toland, *On solitary water-waves of finite amplitude*, Arch. Rational Mech. Anal., 76 (1981), pp. 9–95.
[2] C. J. Amick and J. F. Toland, *On periodic water-waves and their convergence to solitary waves in the long-wave limit*, Philos. Trans. Roy. Soc. London Ser. A, 303 (1981), pp. 633–669.
[3] C. J. Amick, L. E. Fraenkel, and J. F. Toland, *On the Stokes conjecture for the wave of extreme form*, Acta Math., 148 (1982), pp. 193–214.
[4] A. Constantin, *The trajectories of particles in Stokes waves*, Invent. Math., 166 (2006), pp. 523–535.
[5] A. Constantin, M. Ehrnström, and E. Wahlèn, *Symmetry of steady periodic gravity water waves with vorticity*, Duke Math. J., 3 (2007), pp. 591–603.
[6] A. Constantin and J. Escher, *Particle trajectories in solitary water waves*, Bull. Amer. Math. Soc., 44 (2007), pp. 423–431.
[7] A. Constantin, D. Sattinger, and W. Strauss, *Variational formulations for steady water waves with vorticity*, J. Fluid Mech., 548 (2006), pp. 151–163.

[8] A. Constantin and W. Strauss, *Exact steady periodic water waves with vorticity*, Comm. Pure Appl. Math., 57 (2004), pp. 481–527.

[9] A. Constantin and W. Strauss, *Rotational steady water waves near stagnation*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 365 (2007), pp. 2227–2239.

[10] A. Constantin and W. Strauss, *Stability properties of steady water waves with vorticity*, Comm. Pure Appl. Math., 60 (2007), pp. 911–950.

[11] M. Ehrnström, *Deep-water waves with vorticity: Symmetry and rotational behaviour*, Discrete Contin. Dyn. Syst., 19 (2007), pp. 483–491.

[12] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[13] V. M. Hur, *Global bifurcation theory of deep-water waves with vorticity*, SIAM J. Math. Anal., 37 (2006), pp. 1482–1521.

[14] G. Keady and J. Norbury, *On the existence theory for irrotational water waves*, Math. Proc. Camb. Philos. Soc., 83 (1978), pp. 137–157.

[15] J. Ko and W. Strauss, *Large-amplitude steady rotational water waves*, Eur. J. Mech. B Fluids, to appear.

[16] J. B. McLeod, *The Stokes and Krasovskii conjectures for the wave of greatest height*, Stud. Appl. Math., 98 (1997), pp. 311–334. (In preprint form: Univ. of Wisconsin MRC Report no. 2041, 1979).

[17] P. I. Plotnikov, *Proof of the Stokes conjecture in the theory of surface waves*, Dinamika Splosh. Sredy, No. 57 (1982), pp. 41–76 (in Russian). English translation: Stud. Appl. Math., 108 (2002), pp. 217–244.

[18] G. G. Stokes, *Considerations relative to the greatest height of oscillatory irrotational waves which can be propagated without change of form*, in Math. and Phys. Papers I, Cambridge University Press, Cambridge, 1880, pp. 225–228.

[19] A. F. Teles da Silva and D. H. Peregerine, *Steep, steady surface waves on water of finite depth with constant vorticity*, J. Fluid Mech., 195 (1988), pp. 281–302.

[20] J. F. Toland, *On the existence of a wave of greatest height and Stokes's conjecture*, Proc. Roy. Soc. London Ser. A, 363 (1978), pp. 469–485.

[21] J. F. Toland, *Stokes waves*, Topol. Methods Nonlinear Anal., 7 (1996), pp. 1–48; *Errata*, 8 (1996), pp. 412–414.

[22] E. Varvaruca, *Singularities of Bernoulli free boundaries*, Comm. Partial Differential Equations, 31 (2006), pp. 1451–1477.

[23] E. Varvaruca, *Bernoulli free-boundary problems in strip-like domains and a property of permanent waves on water of finite depth*, Proc. Roy. Soc. Edinburgh Sect. A, to appear; also available online at http://arXiv.org/abs/0708.4371.

[24] E. Varvaruca, *On the existence of extreme waves and the Stokes conjecture with vorticity*, submitted; also available online at http://arXiv.org/abs/0707.2224.

SIAM J. MATH. ANAL.
Vol. 39, No. 5, pp. 1693–1709

EXISTENCE AND UNIQUENESS OF SOLUTIONS TO A NONLOCAL EQUATION WITH MONOSTABLE NONLINEARITY*

JÉRÔME COVILLE†, JUAN DÁVILA‡, AND SALOMÉ MARTÍNEZ‡

Abstract. Let $J \in C(\mathbb{R})$, $J \geq 0$, $\int_{\mathbb{R}} J = 1$ and consider the nonlocal diffusion operator $\mathcal{M}[u] = J \star u - u$. We study the equation $\mathcal{M}u + f(x, u) = 0$, $u \geq 0$, in $\mathbb{R}$, where f is a KPP-type nonlinearity, periodic in x. We show that the principal eigenvalue of the linearization around zero is well defined and that a nontrivial solution of the nonlinear problem exists if and only if this eigenvalue is negative. We prove that if, additionally, J is symmetric, then the nontrivial solution is unique.

Key words. nonlocal dispersal, monostable, existence and uniqueness, convolution operator

AMS subject classifications. 45G10, 47G20, 35B50

DOI. 10.1137/060676854

1. Introduction.

Reaction-diffusion equations have been used to describe a variety of phenomena in combustion theory, bacterial growth, nerve propagation, epidemiology, and spatial ecology [13, 12, 15, 19]. However, in many situations, such as in population ecology, dispersal is better described as a long range process rather than as a local one, and integral operators appear as a natural choice. Let us mention in particular the seminal work of Kolmogorov, Petrovsky, and Piskunov [16], who in 1937 introduced a model for the dispersion of gene fractions involving a nonlocal linear operator and a nonlinearity of the form $u(1-u)$, which many authors now call a KPP-type nonlinearity.

Nonlocal dispersal operators usually take the form $\mathcal{M}[u] = \int_{\mathbb{R}^N} k(x, y)u(y)dy - u(x)$, where $k \geq 0$ and $\int_{\mathbb{R}^N} k(y, x)dy = 1$ for all $x \in \mathbb{R}^N$. They have been mainly used in discrete time models [17], while continuous time versions have also been recently considered in population dynamics [14, 18]. Steady state and travelling wave solutions for single equations have been studied in the case $k(x, y) = J(x - y)$, with J even, for some specific reaction nonlinearities in [1, 10, 8, 2, 6, 21].

In this work we restrict ourselves to one dimension and take

$$k(x, y) = J(x - y).$$

We are interested in the existence/nonexistence and uniqueness of solutions of the following problem:

$$\mathcal{M}[u] + f(x, u) = 0 \quad \text{in } \mathbb{R}, \tag{1.1}$$

where $f(x, u)$ is a KPP-type nonlinearity, periodic in x, and

$$\mathcal{M}[u] := J \star u - u. \tag{1.2}$$

*Received by the editors December 6, 2006; accepted for publication (in revised form) July 5, 2007; published electronically February 8, 2008. This work was partially supported by the Ecos-Conicyt project C05E04, Fondecyt grants 1050754 and 1050725, Nucleus Millenium P04-069-F, Information and Randomness, and by FONDAP grant for Applied Mathematics, Chile.
http://www.siam.org/journals/sima/39-5/67685.html

†Centro de Modelamiento Matemático, UMI 2807 CNRS-Universidad de Chile, Blanco Encalada 2120, 7 Piso, Santiago, Chile. Current address: Max Planck Institut for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany (coville@mis.mpg.de).

‡Departamento de Ingeniería Matemática, Universidad de Chile, Blanco Encalada 2120, 5 Piso, Santiago, Chile (jdavila@dim.uchile.cl, samartin@dim.uchile.cl).

We assume that J satisfies

$$J \in C(\mathbb{R}), \quad J \geq 0, \quad \int_{\mathbb{R}} J = 1, \tag{1.3}$$

$$\text{there exist } a < 0 < b \text{ such that } J(a) > 0,\ J(b) > 0. \tag{1.4}$$

On f we assume that

$$\begin{cases} f \in C(\mathbb{R} \times [0,\infty)) \text{ and is differentiable with respect to } u, \\ \text{for each } u,\ f(\cdot,u) \text{ is periodic with period } 2R, \\ f_u(\cdot,0) \text{ is Lipschitz}, \\ f(\cdot,0) \equiv 0 \text{ and } f(x,u)/u \text{ is decreasing with respect to } u, \\ \text{there exists } M > 0 \text{ such that } f(x,u) \leq 0 \text{ for all } u \geq M \text{ and all } x. \end{cases} \tag{1.5}$$

The model example of such a nonlinearity is

$$f(x,u) = u(a(x) - u),$$

where $a(x)$ is periodic and Lipschitz.

In a recent work, Berestycki, Hamel, and Roques [2] studied the analogue of (1.1) with a divergence operator in a periodic setting. More precisely, they considered

$$-\nabla \cdot (A(x)\nabla u) = f(x,u), \quad x \in \mathbb{R}^N, \quad u \geq 0, \tag{1.6}$$

where $A(x)$ is a symmetric matrix of class $C^{1,\alpha}$, periodic with respect to all variables and uniformly elliptic, and f is C^1 and satisfies (1.5). They showed existence of nontrivial solutions provided the linearization of the equation around zero has a negative first periodic eigenvalue.

We prove the following result.

THEOREM 1.1. *Assume J satisfies* (1.3), (1.4) *and f satisfies* (1.5). *Then there exists a nontrivial, periodic solution of* (1.1) *if and only if*

$$\lambda_1(\mathcal{M} + f_u(x,0)) < 0,$$

where λ_1 is the principal eigenvalue of the linear operator $-(\mathcal{M} + f_u(x,0))$ in the set of $2R$-periodic continuous functions. Moreover, if $\lambda_1 \geq 0$, then any nonnegative bounded solution is identically zero.

To prove Theorem 1.1, we first need to show that the principal periodic eigenvalue of $-(\mathcal{M} + f_u(x,0))$ is well defined. Let us introduce some notation:

$$C_{per}(\mathbb{R}) = \{u : \mathbb{R} \to \mathbb{R} \mid u \text{ is continuous and } 2R\text{-periodic}\},$$

$$C^{0,1}_{per}(\mathbb{R}) = \{u : \mathbb{R} \to \mathbb{R} \mid u \text{ is Lipschitz and } 2R\text{-periodic}\}.$$

THEOREM 1.2. *Suppose $a(x) \in C^{0,1}_{per}(\mathbb{R})$. Then the operator $-(\mathcal{M} + a(x))$ has a unique principal eigenvalue λ_1 in $C_{per}(\mathbb{R})$; that is, there is a unique $\lambda_1 \in \mathbb{R}$ such that*

$$\mathcal{M}[\phi_1] + a(x)\phi_1 = -\lambda_1\phi_1 \quad \text{in } \mathbb{R} \tag{1.7}$$

admits a positive solution $\phi_1 \in C_{per}(\mathbb{R})$. Moreover, λ_1 is simple, that is, the space of $C_{per}(\mathbb{R})$ solutions to (1.7) *is one dimensional.*

In [2] the authors proved that (1.6) has at most one nontrivial bounded solution, and that it has to be periodic. A similar result is true for the nonlocal problem (1.1), but this time we need J to be symmetric, that is,

$$J(x) = J(-x) \quad \text{for all } x \in \mathbb{R}. \tag{1.8}$$

Note, however, that for the existence result, Theorem 1.1, we do not need this condition.

THEOREM 1.3. *Assume J satisfies* (1.3), (1.4), (1.8) *and f satisfies* (1.5). *Let u be a nonnegative, bounded solution to* (1.1) *and let λ_1 be the principal eigenvalue of the operator $-(\mathcal{M} + f_u(x, 0))$ with periodic boundary conditions.*

(a) *If $\lambda_1 < 0$, then either $u \equiv 0$ or $u \equiv p$, where p is the positive periodic solution of Theorem* 1.1.

(b) *If $\lambda_1 \geq 0$, then $u \equiv 0$.*

Part (b) of the preceding theorem is already covered in Theorem 1.1 and does not depend on the symmetry of J.

When f is independent of x and satisfies (1.5), the principal eigenvalue of $-(\mathcal{M} + f'(0))$ is given by $\lambda_1 = -f'(0)$ and ϕ_1 is just a constant. Thus in this case Theorem 1.1 says that a bounded, nonnegative, nontrivial solution exists if and only if $f'(0) > 0$, and this solution is just the constant u_0 such that $f(u_0) = 0$. Assuming that J is symmetric, Theorem 1.3 then implies that the constant u_0 is the unique solution in the class of nonnegative, bounded functions.

Recently, considering a nonperiodic nonlinearity f, Berestycki, Hamel, and Rossi [3] analyzed the analogue of Theorem 1.3 for general elliptic operators in $\mathbb{R}^N$, finding sufficient conditions that ensure existence and uniqueness of a positive bounded solution. It is natural to ask whether the periodicity of f and the symmetry of J are crucial hypotheses in Theorem 1.3. We believe that this is the case, since a general nonlocal operator such as (1.2) may contain a transport term, and a standing wave connecting the steady states of the system could appear. We shall investigate further this issue in a forthcoming work.

Hypothesis (1.4) implies that the operator $\mathcal{M}$ satisfies the strong maximum principle. Suppose, for instance, that J satisfies (1.3), (1.4). If $u \in C(\mathbb{R})$ satisfies $\mathcal{M}[u] \geq 0$ in $\mathbb{R}$, then u cannot achieve a global maximum without being constant (see [9]). However, we will need the following version.

THEOREM 1.4. *Assume J satisfies* (1.3), (1.4) *and let $c \in L^\infty(\mathbb{R})$. If $u \in L^\infty(\mathbb{R})$ satisfies $u \leq 0$ a.e. and $\mathcal{M}[u] + c(x)u \geq 0$ a.e. in $\mathbb{R}$, then $\operatorname{ess\,sup}_K u < 0$ for all compact $K \subset \mathbb{R}$ or $u = 0$ a.e. in $\mathbb{R}$.*

If f satisfies the stronger hypothesis that, for any x, $f(x, u)$ is concave with respect to u, then actually the periodic solution p of Theorem 1.1 is continuous. To see this notice that from the strong maximum principle, Theorem 1.4, $J \star p > 0$ in $\mathbb{R}$. The concavity of f with respect to u implies that for any x the map $u \mapsto u - f(x, u)$ is strictly increasing whenever $u - f(x, u) > 0$. Then from the continuity of $J \star p$ and (1.1), which can be rewritten as in the form $J \star p = p - f(x, p)$, we deduce that p is continuous.

In section 2 we review some spectral theory and give the argument of Theorem 1.2. Then we prove Theorem 1.1 in section 3 and the uniqueness result, Theorem 1.3(a), in section 4. We leave for an appendix a proof of Theorem 1.4.

2. Some spectral theory. In this section we deal with the principal eigenvalue problem (1.7). Before stating our result, let us recall some basic spectral results for positive operators due to Edmunds, Potter, and Stuart [11] which are extensions of the Krein–Rutmann theorem for positive noncompact operators.

A cone in a real Banach space X is a nonempty closed set K such that for all $x, y \in K$ and all $\alpha \geq 0$ one has $x + \alpha y \in K$, and if $x \in K$, $-x \in K$, then $x = 0$. A cone K is called reproducing if $X = K - K$. A cone K induces a partial ordering in X by the relation $x \leq y$ if and only if $x - y \in K$. A linear map or operator $T : X \to X$ is called positive if $T(K) \subseteq K$. The dual cone K^* is the set of functionals $x^* \in X^*$ which are positive, that is, such that $x^*(K) \subset [0, \infty)$.

If $T : X \to X$ is a bounded linear map on a complex Banach space X, its essential spectrum (according to Browder [5]) consists of those λ in the spectrum of T such that at least one of the following conditions holds: (1) the range of $\lambda I - T$ is not closed, (2) λ is a limit point of the spectrum of T, (3) $\cup_{n=1}^{\infty} \ker((\lambda I - T)^n)$ is infinite dimensional. The radius of the essential spectrum of T, denoted by $r_e(T)$, is the largest value of $|\lambda|$ with λ in the essential spectrum of T. For more properties of $r_e(T)$ see [20].

Theorem 2.1 (Edmunds, Potter, and Stuart [11]). *Let K be a reproducing cone in a real Banach space X, and let $T \in \mathcal{L}(X)$ be a positive operator such that $T^p(u) \geq cu$ for some $u \in K$ with $\|u\| = 1$, some positive integer p, and some positive number c. Then if $c^{\frac{1}{p}} > r_e(T)$, T has an eigenvector $v \in K$ with associated eigenvalue $\rho \geq c^{\frac{1}{p}}$ and T^* has an eigenvector $v^* \in K^*$ corresponding to the eigenvalue ρ.*

A proof of this theorem can be found in [11]. If the cone K has nonempty interior and T is strongly positive, i.e., $u \geq 0$, $u \neq 0$ implies $Tu \in \operatorname{int}(K)$, then ρ is the unique $\lambda \in \mathbb{R}$ for which there exists nontrivial $v \in K$ such that $Tv = \lambda v$ and ρ is simple; see [22].

Proof of Theorem 1.2. For convenience, in this proof we write the eigenvalue problem

$$\mathcal{M}[u] + a(x)u = -\lambda u$$

in the form

$$\mathcal{L}[u] + b(x)u = \mu u, \tag{2.1}$$

where

$$\mathcal{L}[u] = J \star u, \quad b(x) = a(x) + k, \quad \mu = -\lambda + 1 + k,$$

and $k > 0$ is a constant such that $\inf_{[-R,R]} b > 0$.

Observe that $\mathcal{L} : C_{per(\mathbb{R})} \to C_{per}(\mathbb{R})$ is compact ($C_{per}(\mathbb{R})$ is endowed with the norm $\|u\|_{L^\infty([-R,R])}$). Indeed, let $u_n \in C_{per}(\mathbb{R})$ be a bounded sequence, say $\|u_n\|_{L^\infty([-R,R])} \leq B$. Let $\epsilon > 0$ and let A be large enough so that $\int_{|x| \geq A} J \leq \epsilon$. Since J is uniformly continuous in $[-R - 2A, R + 2A]$ there is $\delta > 0$ such that $|J(z_1) - J(z_2)| \leq \frac{\epsilon}{2(A+R)}$ for $z_1, z_2 \in [-R - 2A, R + 2A]$ with $|z_1 - z_2| \leq \delta$. Then for $x_1, x_2 \in [-R, R]$,

$$\begin{aligned}
|\mathcal{L}[u_n](x_1) - \mathcal{L}[u_n](x_2)| &\leq \int_{\mathbb{R}} |J(x_1 - y) - J(x_2 - y)| \, |u_n(y)| \, dy \\
&\leq 2B\epsilon + B \int_{-R-A}^{R+A} |J(x_1 - y) - J(x_2 - y)| \, dy \\
&\leq 3B\epsilon.
\end{aligned}$$

This shows that $\mathcal{L}[u_n]$ is equicontinuous, and therefore by the Arzelà–Ascoli theorem, $\mathcal{L}[u_n]$ is relatively compact.

Let us now establish some useful lemma.

LEMMA 2.2. *Suppose $b(x) \in C^{0,1}(\mathbb{R})$ is $2R$-periodic, $b(x) > 0$, and let $\sigma := \max_{[-R,R]} b(x)$. Then there exist $p \in \mathbb{N}, \delta > 0$, and $u \in C_{per}(\mathbb{R})$, $u \geq 0$, $u \not\equiv 0$, such that*

$$\mathcal{L}^p u + b(x)^p u \geq (\sigma^p + \delta)u.$$

Observe that the proof of Theorem 1.2 will then easily follow from the above lemma. Indeed, if the lemma holds, then since u and b are nonnegative and $\mathcal{L}$ is a positive operator, we easily see that

$$(\mathcal{L} + b(x))^p[u] \geq \mathcal{L}^p[u] + b(x)^p u \geq (\sigma^p + \delta)u.$$

Using the compactness of the operator $\mathcal{L}$, we have $r_e(\mathcal{L} + b(x)) = r_e(b(x))=\sigma$, and thus $(\sigma^p + \delta)^{\frac{1}{p}} > r_e(\mathcal{L} + b(x))$ and Theorem 2.1 applies. Finally, we observe that the principal eigenvalue is simple since the cone of positive $2R$-periodic functions has nonempty interior and, for a sufficiently large p, the operator $(\mathcal{L} + b)^p$ is strongly positive. $\square$

Let us now turn our attention to the proof of the above lemma.

Proof of Lemma 2.2. Recall that for $p \in \mathbb{N} \setminus \{0\}$, $J \star^p u := J \star (J \star^{p-1} u)$ is well defined by induction and satisfies $J \star^p u = \mathcal{J}_p \star u$ with $\mathcal{J}_p$ defined as follows:

$$\mathcal{J}_p := \underbrace{J \star J \star \cdots \star J \star J}_{p \text{ times}}.$$

By (1.4) it follows that there exists $p \in \mathbb{N}$ such that $\inf_{(-2R-1,2R+1)} \mathcal{J}_p > 0$. Using the definition of $\mathcal{L}$, a short computation shows that

$$\mathcal{L}^p[u] := \int_{-R}^{R} \widetilde{\mathcal{J}}_p(x,y)u(y)\,dy$$

with $\widetilde{\mathcal{J}}_p(x,y) = \sum_{k\in\mathbb{Z}} \mathcal{J}_p(x + 2kR - y)$. Following the idea of Hutson et al. [14], consider now the following function:

$$v(x) := \begin{cases} \frac{\eta(x)}{b^p(x_0)-b^p(x)+\gamma} & \text{in } \Omega_{2\epsilon} := (x_0 - 2\epsilon, x_0 + 2\epsilon), \\ 0 & \text{elsewhere,} \end{cases}$$

where $x_0 \in (-R,R)$ is a point of maximum of $b(x)$, $\epsilon > 0$ is chosen such that $(x_0 - 2\epsilon, x_0 + 2\epsilon) \subset (-R,R)$, γ is a positive constant that we will define later on, and η is a smooth function such that $0 \leq \eta \leq 1$, $\eta(x) = 1$ for $|x - x_0| \leq \epsilon$, $\eta(x) = 0$ for $|x - x_0| \geq 2\epsilon$. Let us compute $\mathcal{L}^p[v] + b^p(x)v - \sigma^p v$:

$$\begin{aligned} \mathcal{L}^p[v] + b^p(x)v - \sigma^p v &= \int_{x_0-\epsilon}^{x_0+\epsilon} \widetilde{\mathcal{J}}_p(x,y)\frac{dy}{b^p(x_0) - b^p(y) + \gamma} + \int_{\Omega_{2\epsilon}\setminus\Omega_\epsilon} \widetilde{\mathcal{J}}_p(x,y)v(y)\,dy \\ &\quad + (b^p(x) - b^p(x_0))v \\ &\geq \int_{x_0-\epsilon}^{x_0+\epsilon} \widetilde{\mathcal{J}}_p(x,y)\frac{dy}{b^p(x_0) - b^p(y) + \gamma} + (b^p(x) - b^p(x_0))v \\ &\geq \int_{x_0-\epsilon}^{x_0+\epsilon} \widetilde{\mathcal{J}}_p(x,y)\frac{dy}{b^p(x_0) - b^p(y) + \gamma} - 1. \end{aligned}$$

Using that $\inf_{(-2R-1,2R+1)} \mathcal{J}_p > 0$, it follows that $\widetilde{\mathcal{J}}_p(x,y) \geq c > 0$ for $x, y \in (-R,R)$. Hence

$$\int_{x_0-\epsilon}^{x_0+\epsilon} \widetilde{\mathcal{J}}_p(x,y) \frac{dy}{b^p(x_0)-b^p(y)+\gamma} \geq c \int_{x_0-\epsilon}^{x_0+\epsilon} \frac{dy}{k|x_0-y|+\gamma},$$

where k is the Lipschitz constant for b^p. Using this inequality in the above estimate yields

$$\mathcal{L}^p[v] + b^p(x)v - \sigma^p v \geq c \int_{x_0-\epsilon}^{x_0+\epsilon} \frac{dy}{k|x_0-y|+\gamma} - 1.$$

Therefore we have

$$\begin{aligned}\mathcal{L}^p[v] + b^p(x)v - (\sigma^p+\delta)v &\geq \frac{2c}{k}\log\left(1+\frac{k\epsilon}{\gamma}\right) - 1 - \delta v \\ &\geq \frac{2c}{k}\log\left(1+\frac{k\epsilon}{\gamma}\right) - 1 - \frac{\delta}{\gamma}.\end{aligned}$$

Choosing now $\gamma > 0$ small so that $\frac{2c}{k}\log(1+\frac{k\epsilon}{\gamma}) - 1 > \frac{1}{2}$ and $\delta = \frac{\gamma}{4}$, we end up with

$$\mathcal{L}^p[v] + b^p(x)v - (\sigma^p+\delta)v \geq \frac{1}{4} > 0. \qquad \square$$

3. Existence of solutions.

Proof of Theorem 1.1. We follow the argument developed by Berestycki, Hamel, and Roques in [2].

First assume that $\lambda_1 < 0$. From Theorem 1.2 there exists a positive eigenfunction ϕ_1 such that

$$\mathcal{M}[\phi_1] + f_u(x,0)\phi_1 = -\lambda_1\phi_1 \geq 0.$$

Computing $\mathcal{M}[\epsilon\phi_1] + f(x,\epsilon\phi_1)$, it follows that

$$\begin{aligned}\mathcal{M}[\epsilon\phi_1] + f(x,\epsilon\phi_1) &= f(x,\epsilon\phi_1) - f_u(x,0)\epsilon\phi_1 - \lambda_1\epsilon\phi_1 \\ &= -\lambda_1\epsilon\phi_1 + o(\epsilon\phi_1) > 0.\end{aligned}$$

Therefore, for $\epsilon > 0$ small, $\epsilon\phi_1$ is a periodic subsolution of (1.1). By definition of f, any constant M sufficiently large is a periodic supersolution of the problem. Choosing M so large that $\epsilon\phi_1 \leq M$ and using a basic iterative scheme yields the existence of a positive periodic solution u of (1.1).

Let us now turn our attention to the nonexistence setting and assume that $\lambda_1 \geq 0$. Let u be a bounded nonnegative solution of (1.1). Observe that $\gamma\phi_1$ is a periodic supersolution for any positive γ. Indeed,

$$\begin{aligned}\mathcal{M}[\gamma\phi_1] + f(x,\gamma\phi_1) &< \mathcal{M}[\gamma\phi_1] + f_u(x,0)\gamma\phi_1 \\ &\leq -\lambda_1\gamma\phi_1 \leq 0.\end{aligned}$$

Since $\phi_1 \geq \delta$ for some positive δ we may define the following quantity:

$$\gamma^* := \inf\{\gamma > 0 | u \leq \gamma\phi_1\}.$$

We have the following claim.

CLAIM 3.1. $\gamma^* = 0$.

Observe that we end the proof of the theorem by proving the above claim.

Proof of the claim. Assume that $\gamma^* > 0$. Since $v := u - \gamma^*\phi_1$ satisfies $v \le 0$ in $\mathbb{R}$ and

$$\mathcal{M}[v] + c(x)v \ge 0 \quad \text{in } \mathbb{R},$$

where $c(x) = \frac{f(x,u)-f(x,\gamma^*\phi_1)}{v}$ by the strong maximum principle, Theorem 1.4, we have the following possibilities:

- either $u \equiv \gamma^*\phi_1$, or
- there exists a sequence of points $(x_n)_{n\in\mathbb{N}}$ such that $|x_n| \to +\infty$ and $\lim_{n\to+\infty} \gamma^*\phi_1(x_n) - u(x_n) = 0$.

In the first case we get the following contradiction:

$$0 = \mathcal{M}[\gamma^*\phi_1] + f(x, \gamma^*\phi_1) < \mathcal{M}[\gamma^*\phi_1] + f_u(x,0)\gamma^*\phi_1 \le 0.$$

Hence $\gamma^* = 0$.

In the second case we argue as follows. Let $(y_n)_{n\in\mathbb{N}}$ be a sequence of points satisfying, for all n, $y_n \in [-R, R]$ and $x_n - y_n \in 2R\mathbb{Z}$. Up to extraction of a subsequence, $y_n \to \bar{y}$. Now consider the following sequence of functions $u_n := u(. + x_n)$, $\phi_n := \phi_1(. + x_n)$, and $w_n := \gamma^*\phi_n - u_n$ so that $w_n > 0$ in $\mathbb{R}$. Since $\mathcal{M}$ is translation invariant and f is periodic, u_n and $\phi_n > 0$ satisfy

$$\mathcal{M}[u_n] + f(x + y_n, u_n) = 0 \quad \text{in } \mathbb{R},$$

$$\mathcal{M}[\gamma^*\phi_n] + f_u(x + y_n, 0)\gamma^*\phi_n \le 0 \quad \text{in } \mathbb{R}.$$

It follows that

$$J \star w_n \le a_n(x)w_n,$$

where

$$a_n(x) = 1 - \frac{\gamma^* f_u(x + y_n, 0)\phi_n - f(x + y_n, u_n)}{\gamma^*\phi_n - u_n}.$$

Since $w_n > 0$ we see that a_n is well defined and $a_n \ge 0$. Using that $f(x,u)/u$ is nonincreasing with respect to u we have $f(x, \gamma^*\phi_n) \le \gamma^* f_u(x,0)\phi_n$. This implies

$$\frac{\gamma^* f_u(x + y_n, 0)\phi_n - f(x + y_n, u_n)}{\gamma^*\phi_n - u_n} \ge \frac{f(x + y_n, \gamma^*\phi_n) - f(x + y_n, u_n)}{\gamma^*\phi_n - u_n} \ge -C.$$

Thus

$$0 \le a_n \le C + 1 \quad \text{in } \mathbb{R} \quad \text{for all } n,$$

with C independent of n. Observe that

$$J \star w_n(0) = a_n(0)(\gamma^*\phi_1(x_n) - u(x_n)) \to 0,$$

which implies

$$\int_{\mathbb{R}} J(-y)w_n(y)\,dy \to 0 \quad \text{as } n \to +\infty.$$

Similarly,

$$J \star J \star w_n(0) = J \star (a_n w_n)(0) = \int_{\mathbb{R}} J(-y) a_n(y) w_n(y)\, dy,$$

but

$$\int_{\mathbb{R}} J(-y) a_n(y) w_n(y)\, dy \le \|a_n\|_{L^\infty} \int_{\mathbb{R}} J(-y) w_n(y)\, dy \to 0.$$

Hence

$$J \star J \star w_n(0) = \int_{\mathbb{R}} (J \star J)(-y) w_n(y)\, dy \to 0 \quad \text{as } n \to +\infty.$$

Defining

$$\mathcal{J}_k := \underbrace{J \star \cdots \star J}_{k \text{ times}},$$

we see that for any fixed $k \in \mathbb{N}$,

$$\int_{\mathbb{R}} \mathcal{J}_k(-y) w_n(y)\, dy \to 0 \quad \text{as } n \to +\infty.$$

By (1.4) the support of $\mathcal{J}_k$ increases to all of $\mathbb{R}$ as $k \to +\infty$. Thus we may find a new subsequence such that $w_n \to 0$ a.e. in $\mathbb{R}$ as $n \to +\infty$. Since ϕ_1 is periodic and continuous, $\phi_n(x) \to \bar{\phi}(x)$ uniformly with respect to x, where $\bar{\phi}(x) = \phi(x + \bar{y})$. Hence $\bar{u}(x) = \lim_{n\to+\infty} u_n(x)$ exists a.e. and is given by $\bar{u}(x) = \gamma^* \bar{\phi}$. By dominated convergence, $\bar{u}$ is a solution to

$$\mathcal{M}[\bar{u}] + f(x + \bar{y}, \bar{u}) = 0,$$

while by uniform convergence

$$\mathcal{M}[\gamma^* \bar{\phi}] + f_u(x + \bar{y}, 0)\gamma^* \bar{\phi} \le 0 \quad \text{in } \mathbb{R}.$$

Since $\bar{u} = \gamma^* \bar{\phi}$ it follows that $f(x + \bar{y}, \gamma^* \bar{\phi}) \equiv f_u(x + \bar{y}, 0)\gamma^* \bar{\phi}$. This contradicts the fact that $f(x, u)/u$ is decreasing in u. Hence, $\gamma^* = 0$. ∎

4. Uniqueness when *J* is symmetric. Throughout this section we assume that J is symmetric. For the proof of Theorem 1.3 we follow the ideas in [2].

Proof of Theorem 1.3. Part (b) of this theorem is contained in Theorem 1.1 so we concentrate on part (a).

Let p denote the positive periodic solution to (1.1) constructed in Theorem 1.1 and let $u \ge 0$, $u \not\equiv 0$ be a bounded solution. We will prove that $u \equiv p$.

We show first that $u \le p$. Set

$$\gamma^* := \inf\{\gamma > 0 \mid u \le \gamma p\}.$$

Note that γ^* is well defined because u is bounded and p is bounded below by a positive constant. We claim that

$$\gamma^* \le 1.$$

Suppose that $\gamma^* > 1$ and note that $u \le \gamma^* p$. By Theorem 1.4 either $u \equiv \gamma^* p$ or $\operatorname{ess\,inf}_K(\gamma^* p - u) > 0$ for all compact $K \subset \mathbb{R}$. The first possibility leads to $f(x, \gamma^* p) = \gamma^* f(x,p)$ for all $x \in \mathbb{R}$, which is not possible if $\gamma^* > 1$. In the second case there exists a sequence $(x_n)_{n\in\mathbb{N}}$ such that $|x_n| \to +\infty$ and $\lim_{n\to+\infty} \gamma^* p(x_n) - u(x_n) = 0$. Let $(y_n)_{n\in\mathbb{N}}$ be a sequence satisfying $y_n \in [-R,R]$ and $x_n - y_n = k_n 2R$ for some $k_n \in \mathbb{Z}$. We may assume that $y_n \to \bar{y}$. Let $u_n := u(.+x_n)$, which satisfies

$$\mathcal{M}[u_n] + f(x+y_n, u_n) = 0.$$

Let $w_n = \gamma^* p(.+y_n) - u_n \ge 0$. Then $w_n > 0$ in $\mathbb{R}$ and

$$J \star w_n = a_n(x) w_n,$$

where

$$a_n(x) = 1 - \frac{\gamma^* f(x+y_n, p(x+y_n)) - f(x+y_n, u_n(x))}{\gamma^* p(x+y_n) - u_n(x)}.$$

Since $w_n > 0$ we deduce that a_n is well defined and $a_n \ge 0$. Using that $f(x,u)/u$ is nonincreasing with respect to u and the fact that $\gamma^* > 1$, we have $f(x,\gamma^* p) \le \gamma^* f(x,p)$. This implies

$$\frac{\gamma^* f(x,p) - f(x,u)}{\gamma^* p - u} \ge \frac{f(x,\gamma^* p) - f(x,u)}{\gamma^* p - u} \ge -C.$$

Thus

$$0 \le a_n \le C+1 \quad \text{in } \mathbb{R} \quad \text{for all } n,$$

with C independent of n. Observe that

$$J \star w_n(0) = a_n(0)(\gamma^* p(y_n) - u(x_n)) = a_n(0)(\gamma^* p(x_n) - u(x_n)) \to 0,$$

which implies

$$\int_{\mathbb{R}} J(-y) w_n(y)\, dy \to 0 \quad \text{as } n \to +\infty.$$

Similarly,

$$J \star J \star w_n(0) = J \star (a_n w_n)(0) = \int_{\mathbb{R}} J(-y) a_n(y) w_n(y)\, dy,$$

but

$$\int_{\mathbb{R}} J(-y) a_n(y) w_n(y)\, dy \le \|a_n\|_{L^\infty} \int_{\mathbb{R}} J(-y) w_n(y)\, dy \to 0.$$

Hence

$$J \star J \star w_n(0) = \int_{\mathbb{R}} (J \star J)(-y) w_n(y)\, dy \to 0 \quad \text{as } n \to +\infty.$$

Defining

$$\mathcal{J}_k := \underbrace{J \star \cdots \star J}_{k \text{ times}},$$

we see that for all $k \in \mathbb{N}$,

$$\int_{\mathbb{R}} \mathcal{J}_k(-y)w_n(y)\,dy \to 0 \quad \text{as } n \to +\infty.$$

Hypothesis (1.4) implies that the support of $\mathcal{J}_k$ converges to all of $\mathbb{R}$ as $k \to +\infty$. Therefore, for a subsequence, $w_n \to 0$ a.e. in $\mathbb{R}$ as $n \to +\infty$. Since p is periodic, for possibly a new subsequence $p(x+y_n) \to p(x+\bar{y})$ a.e. Hence, $\bar{u}(x) = \lim_{n\to+\infty} u_n(x)$ exists a.e. and by dominated convergence, $\bar{u}$ is a solution to

$$\mathcal{M}[\bar{u}] + f(x+\bar{y}, \bar{u}) = 0. \tag{4.1}$$

But since $w_n \to 0$ a.e. we have $\bar{u} = \gamma^* p(\cdot + \bar{y})$. Thus $\gamma^* p(\cdot+\bar{y})$ is a solution to (4.1), which is impossible for $\gamma^* > 1$ as argued before.

The proof that $p \le u$ is analogous, but a key point is to prove first that under the conditions of Theorem 1.3 any nontrivial, nonnegative solution is bounded below by a positive constant. This is the content of Proposition 4.1. $\square$

PROPOSITION 4.1. *Assume that J satisfies* (1.3), (1.4), *and* (1.8), *f satisfies* (1.5), *and that the operator $-(\mathcal{M} - f_u(x,0))$ has a negative principal periodic eigenvalue. Suppose that u is a nonnegative, bounded solution to* (1.1). *Then $u \equiv 0$ or there exists a constant $c > 0$ such that*

$$u(x) \ge c \quad \textit{for all } x \in \mathbb{R}.$$

The basic tool to prove Proposition 4.1, following an idea in [2], is to study the principal eigenvalue of the linearized operator in bounded domains. More precisely, let $\Omega = (-r, +r)$ and $a\colon \Omega \to \mathbb{R}$ be Lipschitz. We consider the eigenvalue problem in Ω with "Dirichlet boundary condition" in the following sense:

$$\begin{cases} \mathcal{M}[\varphi] + a(x)\varphi = -\lambda\varphi & \text{in } \Omega, \\ \varphi(x) = 0 \quad \text{for all } x \notin \Omega, \\ \varphi|_{\overline{\Omega}} \text{ is continuous.} \end{cases} \tag{4.2}$$

We show that the principal eigenvalue for (4.2) exists and converges to the principal periodic eigenvalue as $r \to +\infty$. The first step is to establish variational characterizations of these eigenvalues, which is the argument that requires the symmetry of J.

LEMMA 4.2. *Let $\Omega \subset \mathbb{R}$ be a bounded open interval. Assume that J satisfies* (1.3), (1.4), *and* (1.8), *and let $a\colon \Omega \to \mathbb{R}$ be Lipschitz. Then there exists a smallest λ_1 such that* (4.2) *has a nontrivial solution. This eigenvalue is simple and the eigenfunctions are of constant sign in Ω. Moreover,*

$$\lambda_1 = \min_{\varphi \in C(\overline{\Omega})} -\frac{\int_\Omega (\mathcal{M}[\widetilde{\varphi}] + a(x)\varphi)\varphi}{\int_\Omega \varphi^2}, \tag{4.3}$$

where $\widetilde{\varphi}$ denotes the extension by 0 *of φ to $\mathbb{R}$ and the minimum is attained.*

The statement and the proof are analogous to those of Theorem 3.1 in [14] except that here we do not assume that $J(0) > 0$. A different formula for the principal eigenvalue with a Dirichlet boundary condition appears in [7], where it is used to characterize the rate of decay of solutions to a linear evolution equation.

Proof. Define the operator $X[\varphi] = \int_\Omega J(x-y)\,\varphi(y)\,dy$ for $\varphi \in C(\overline{\Omega})$. Then $X : C(\overline{\Omega}) \to C(\overline{\Omega})$ is compact. Let $c_0 > 0$ be such that $\inf_\Omega a(x) + c_0 > 0$ and define $\widetilde{a} = a + c_0$. The eigenvalue problem (4.2) is equivalent to the following: find $\varphi \in C(\overline{\Omega})$ and $\lambda \in \mathbb{R}$ such that

$$X[\varphi] + \widetilde{a}\varphi = (-\lambda + 1 + c_0)\varphi \quad \text{in } \Omega.$$

A calculation similar to Lemma 2.2 shows that there exists an integer p, $u \in C(\overline{\Omega})$, and $\delta > 0$ such that

$$(X + \widetilde{a})^p u \geq \left(\left(\max_{\overline{\Omega}} \widetilde{a}\right)^p + \delta\right) u \quad \text{in } \Omega. \tag{4.4}$$

Using Theorem 2.1 we deduce that the operator $X + \widetilde{a}$ has a unique principal eigenvalue $\rho > 0$ and a principal eigenvector $\varphi_1 \in C(\overline{\Omega})$. Let $\lambda = 1 + c_0 - \rho$ so that $X[\varphi_1] + a(x)\varphi_1 = (1-\lambda)\varphi_1$. From (4.4) we deduce that σ_+ defined by

$$\sigma_+ = \sup_{\varphi \in C(\overline{\Omega})} \frac{\int_\Omega (X[\varphi] + a(x)\varphi)\varphi}{\int_\Omega \varphi^2} \tag{4.5}$$

satisfies

$$\sigma_+ \geq 1 - \lambda > \max_{\overline{\Omega}} a. \tag{4.6}$$

Now, using the same argument as in [14] we deduce that the supremum in (4.5) is achieved. Indeed, it is standard [4] that the spectrum of $\hat{X} + a(x)$ is to the left of σ_+ and that there exists a sequence $\varphi_n \in C(\overline{\Omega})$ such that $\|\varphi_n\|_{L^2(\Omega)} = 1$ and $\|(X + a(x) - \sigma_+)\varphi_n\|_{L^2(\Omega)} \to 0$ as $n \to +\infty$. By compactness of $X : L^2(\Omega) \to C(\overline{\Omega})$ for a subsequence, $\lim_{n\to+\infty} X[\varphi_n]$ exists in $C(\overline{\Omega})$. Then, using (4.6), we see that $\varphi_n \to \varphi$ in $L^2(\Omega)$ for some φ and $(X + a)\varphi = \sigma_+\varphi$. This equation implies $\varphi \in C(\overline{\Omega})$, and hence σ_+ is a principal eigenvalue for the operator X and by uniqueness of this eigenvalue we have $\sigma_+ = 1 - \lambda$. ◻

LEMMA 4.3. *Assume that J satisfies* (1.3), (1.4), *and* (1.8) *and that $a\colon \mathbb{R} \to \mathbb{R}$ is a $2R$-periodic, Lipschitz function. Then the principal eigenvalue of the operator $-(\mathcal{M} + a(x))$ in $C_{per}(\mathbb{R})$ is given by*

$$\lambda_1(a) = \inf_{\|\varphi\|_{L^2(\mathbb{R})}=1} -\int_{\mathbb{R}} (\mathcal{M}[\varphi] + a(x)\varphi)\varphi \tag{4.7}$$

$$= \min_{\varphi \in C_{per}(\mathbb{R})} -\frac{\int_{-R}^{R} (\mathcal{M}[\varphi] + a(x)\varphi)\varphi}{\int_{-R}^{R} \varphi^2}. \tag{4.8}$$

Proof. By Theorem 1.2 we know that there exists a unique principal eigenvalue $\lambda_1(a)$ of the operator $-(\mathcal{M} + a)$ in $C_{per}(\mathbb{R})$. Let $\phi_1 \in C_{per}(\mathbb{R})$ denote a positive eigenfunction associated with $\lambda_1(a)$. We normalize ϕ_1 such that

$$\int_{-R}^{R} \phi_1^2 = 2R. \tag{4.9}$$

On the other hand, the quantity

$$\widetilde{\lambda}_1(a) = \inf_{\varphi \in C_{per}(\mathbb{R})} -\frac{\int_{-R}^{R} (\mathcal{M}[\varphi] + a(x)\varphi)\varphi}{\int_{-R}^{R} \varphi^2}$$

is also an eigenvalue of $-(\mathcal{M}+a)$ on $C_{per}(\mathbb{R})$ with a positive eigenfunction. By uniqueness of the principal eigenvalue, $\lambda_1(a)=\widetilde{\lambda}_1(a)$.

We claim that

$$\inf_{\|\varphi\|_{L^2(\mathbb{R})}=1} -\int_{\mathbb{R}} (\mathcal{M}[\varphi]+a(x)\varphi)\varphi \le \lambda_1(a).$$

Indeed, for $r>0$ let $\eta_r\in C_0^\infty(\mathbb{R})$ be such that $0\le \eta_r\le 1$, $\eta_r(x)=1$ for $|x|\le r$, $\eta_r(x)=0$ for $|x|\ge r+1$. It will be sufficient to show that

$$\lim_{r\to+\infty} \frac{\int_{\mathbb{R}}(\mathcal{M}[\phi_1\eta_r]+a\phi_1\eta_r)\phi_1\eta_r}{\int_{\mathbb{R}}(\phi_1\eta_r)^2} = -\lambda_1(a). \tag{4.10}$$

By (4.9) we have

$$\int_{\mathbb{R}} (\phi_1\eta_r)^2 = 2r+O(1) \quad \text{ as } r\to+\infty. \tag{4.11}$$

Let $0<\theta<1$. Then

$$\begin{aligned} |\mathcal{M}[\phi_1](x)-\mathcal{M}[\phi_1\eta_r]| &\le \|\phi_1\|_{L^\infty}\int_{|x-z|\ge r}|J(z)|\,dz \\ &\le \|\phi_1\|_{L^\infty}\int_{|z|\ge(1-\theta)r}|J(z)|\,dz \quad \text{for all } |x|\le\theta r \\ &= o(1) \quad \text{uniformly for all } |x|\le \theta r. \end{aligned} \tag{4.12}$$

We split the integral

$$\int_{\mathbb{R}}(\mathcal{M}[\phi_1\eta_r]+a\phi_1\eta_r)\phi_1\eta_r = \int_{|x|\le\theta r}\dots dx + \int_{|x|\ge\theta r}\dots dx. \tag{4.13}$$

Using $\eta_r(x)=1$ for $|x|\le\theta r$ and (4.12) we see that

$$\begin{aligned} \int_{|x|\le\theta r}(\mathcal{M}[\phi_1\eta_r]+a\phi_1\eta_r)\phi_1\eta_r &= \int_{|x|\le\theta r}(\mathcal{M}[\phi_1\eta_r]+a\phi_1)\phi_1 \\ &= \int_{|x|\le\theta r}(\mathcal{M}[\phi_1]+a\phi_1+o(1))\phi_1 \\ &= -2\theta\lambda_1(a)r+o(r) \quad \text{ as } r\to+\infty. \end{aligned}$$

The second integral in (4.13) is bounded by

$$\left|\int_{|x|\ge\theta r}(\mathcal{M}[\phi_1\eta_r]+a\phi_1\eta_r)\phi_1\eta_r\right| \le C(1-\theta)r. \tag{4.14}$$

Thus from (4.11)–(4.14) we conclude that

$$\left|\frac{\int_{\mathbb{R}}(\mathcal{M}[\phi_1\eta_r]+a\phi_1\eta_r)\phi_1\eta_r}{\int_{\mathbb{R}}(\phi_1\eta_r)^2}+\lambda_1(a)\right| \le C(1-\theta)+o(1),$$

which proves (4.10).

To establish (4.7) it remains to verify that

$$\lambda_1(a) \le -\frac{\int_{\mathbb{R}}(\mathcal{M}[\varphi]+a(x)\varphi)\varphi}{\int_{\mathbb{R}}\varphi^2} \quad \text{for all } \varphi\in C_c(\mathbb{R}). \tag{4.15}$$

By uniqueness of the principal eigenvalue we have

$$\lambda_1(a) = \inf_{\varphi\in C_{per}(\Omega_k)} -\frac{\int_{-kR}^{kR}(\mathcal{M}[\varphi]+a(x)\varphi)\varphi}{\int_{-kR}^{kR}\varphi^2}, \tag{4.16}$$

where

$$\Omega_k = (-kR, kR) \quad \text{ for } k\ge 1$$

and $C_{per}(\Omega_k)$ is the set of continuous $2kR$-periodic functions on $\mathbb{R}$.

Fix $\varphi\in C_c(\mathbb{R})$ and consider k large enough so that $\operatorname{supp}(\varphi)\subseteq\Omega_k$. Consider now φ_k the $4kR$-periodic extension of φ. Since $\varphi_k\in C_{per}(\Omega_{2k})$, (4.16) yields

$$\lambda_1(a) \le -\frac{\int_{-2kR}^{2kR}(\mathcal{M}[\varphi_k]+a(x)\varphi_k)\varphi_k}{\int_{-2kR}^{2kR}\varphi_k^2} = -\frac{\int_{\mathbb{R}}(\mathcal{M}[\varphi_k]+a(x)\varphi)\varphi}{\int_{\mathbb{R}}\varphi^2}. \tag{4.17}$$

For $|x|\le kR$ we have

$$|\mathcal{M}[\varphi_k](x)-\mathcal{M}[\varphi](x)| \le \|\varphi\|_{L^\infty}\int_{|y|\ge 2kR}|J(x-y)|\,dy \le \|\varphi\|_{L^\infty}\int_{|z|\ge kR}|J(z)|\,dz.$$

Hence

$$\lim_{k\to+\infty}\int_{\mathbb{R}}(\mathcal{M}[\varphi_k]+a(x)\varphi)\varphi = \int_{\mathbb{R}}(\mathcal{M}[\varphi]+a(x)\varphi)\varphi. \tag{4.18}$$

Thanks to (4.17) and (4.18), we conclude the validity of (4.15). ◻

LEMMA 4.4. *Assume J satisfies* (1.3), (1.4), *and* (1.8) *and that $a\colon\mathbb{R}\to\mathbb{R}$ is a $2R$-periodic, Lipschitz function. Let $\lambda_{r,y}$ be the principal eigenvalue of* (4.2) *for*

$$\Omega_{r,y} = B_r(y)$$

and let $\lambda_1(a)$ denote the principal eigenvalue of $-(\mathcal{M}+a(x))$ in $C_{per}(\mathbb{R})$. Then

$$\lim_{r\to+\infty}\lambda_{r,y} = \lambda_1(a).$$

Moreover, the applications $y\mapsto\lambda_{r,y}$ and $y\mapsto\varphi_{r,y}$ are periodic. The periodicity of the application $y\mapsto\varphi_{r,y}$ is understood as follows:

$$\varphi_{r,y+2R}(x) = \varphi_{r,y}(x-2R).$$

Proof. For convenience we write

$$\lambda_r = \lambda_{r,y}$$

and let φ_r be a positive eigenfunction of (4.2) in Ω_r.

By the variational characterization (4.3) we see that $r \mapsto \lambda_r$ is nonincreasing, and hence $\lim_{r\to+\infty} \lambda_r$ exists. Moreover, using (4.7) we have

$$\lambda_r \geq \lambda_1(a) \quad \text{for all } r > 0. \tag{4.19}$$

Let $\phi_1 \in C_{per}(\mathbb{R})$ be a positive eigenfunction of $-(\mathcal{M} + a(x))$ with eigenvalue $\lambda_1(a)$ normalized such that

$$\int_{-R}^{R} \phi_1^2 = 2R.$$

Let $\eta_r \in C_0^\infty(\mathbb{R})$ be such that $0 \leq \eta \leq 1$,

$$\eta_r(x) = 1 \text{ for } |x - y| \leq r - 1, \qquad \eta_r(x) = 0 \text{ for } |x - y| \geq r$$

and such that $\|\eta_r\|_{C^2(\mathbb{R})} \leq C$ with C independent of r. Arguing in the same way as in the proof of Lemma 4.3 we obtain

$$\lim_{r\to+\infty} \frac{\int_{\mathbb{R}}(\mathcal{M}[\phi_1\eta_r] + a\phi_1\eta_r)\phi_1\eta_r}{\int_{\mathbb{R}}(\phi_1\eta_r)^2} = -\lambda_1(a).$$

Since

$$\lambda_r \leq -\frac{\int_{\mathbb{R}}(\mathcal{M}[\phi_1\eta_r] + a\phi_1\eta_r)\phi_1\eta_r}{\int_{\mathbb{R}}(\phi_1\eta_r)^2}$$

we conclude that

$$\lim_{r\to+\infty} \lambda_r \leq \lambda_1(a).$$

This and (4.19) prove the desired result.

Let us now show the periodicity of the applications $y \mapsto \lambda_{r,y}$ and $y \mapsto \varphi_{r,y}$. Replace y by $y+2R$ in the above problem (4.2) and let us denote by $\lambda_{r,y+2R}$ and $\varphi_{r,y+2R}$ the corresponding principal eigenvalue and the associated positive eigenfunction:

$$\mathcal{M}[\varphi_{r,y+2R}] + a(x)\varphi_{r,y+2R} = -\lambda_{r,y+2R}\varphi_{r,y+2R} \quad \text{ in } B_r(y + 2R).$$

We take the following normalization:

$$\int_{\Omega_{r,y+2R}} \varphi_{r,y+2R}^2(x)\,dx = 1.$$

Let us defined $\psi(x) := \varphi_{r,y+2R}(x + 2R)$ for any $x \in B_r(y)$. A short computation shows that

$$\mathcal{M}[\psi](x) = \mathcal{M}[\varphi]_{r,y+2R}(x + 2R).$$

Therefore, using the periodicity of $a(x)$, we have

$$\mathcal{M}[\psi](x) + a(x + 2R)\psi(x) = \lambda_{r,y+2R}\psi \quad \text{in } B_r(y),$$

$$\mathcal{M}[\psi](x) + a(x)\psi(x) = \lambda_{r,y+2R}\psi \quad \text{in } B_r(y).$$

Thus, $\lambda_{r,y+2R}$ is a principal eigenvalue of the problem (4.2) with $\Omega_{r,y} = B_r(y)$. Hence, by uniqueness of the principal eigenvalue we have $\lambda_{r,y} = \lambda_{r,y+2R}$ and $\psi = \gamma\varphi_{r,y}$ for some positive γ. Using the normalization, it follows that $\gamma = 1$. Therefore, $\varphi_{r,y}(x) = \varphi_{r,y+2R}(x+2R)$; in other words

$$\varphi_{r,y+2R}(x) = \varphi_{r,y}(x-2R). \qquad \square$$

Remark 4.5. The proof of Lemma 4.4 yields the slightly stronger conclusion that the convergence

$$\lim_{r\to+\infty} \lambda_{r,y} = \lambda_1(a)$$

is uniform with respect to $y \in \mathbb{R}$, since $\lambda_{r,y}$ is continuous in y.

Proof of Proposition 4.1. Let $u \geq 0$ be a bounded solution to (1.1) such that $u \not\equiv 0$. By the strong maximum principle (Theorem 1.4) we must have $\inf_K u > 0$ for compact sets $K \subset \mathbb{R}$.

Given $y \in \mathbb{R}$ and $r > 0$ we write $\Omega_{r,y} = (y-r, y+r)$, $\lambda_{r,y}$ the principal eigenvalue of $-(\mathcal{M} + f_u(x,0))$ with Dirichlet boundary condition in $\Omega_{r,y}$ as in (4.2), and $\varphi_{r,y}$ a positive Dirichlet eigenfunction normalized so that

$$\int_{\Omega_{r,y}} \varphi_{r,y}^2 = 1.$$

Since the principal eigenvalue $\lambda_1 := \lambda_1(f_u(x,0))$ of $-(\mathcal{M} + f_u(x,0))$ with periodic boundary conditions is negative by hypothesis, by Lemma 4.4 and Remark 4.5 we may fix $r > 0$ large enough so that

$$\lambda_{r,y} < \lambda_1/2 \quad \text{for all } y \in \mathbb{R}.$$

Note that for $x \in \Omega_{r,y}$,

$$\begin{aligned} \mathcal{M}[\gamma\varphi_{r,y}] + f(x,\gamma_{r,y}) &= -\lambda_{r,y}\gamma\varphi_{r,y} - f_u(x,0)\gamma\varphi_{r,y} + f(x,\gamma\varphi_{r,y}) \\ &\geq -\lambda_1/2\gamma\varphi_{r,y} - f_u(x,0)\gamma\varphi_{r,y} + f(x,\gamma\varphi_{r,y}) \\ &\geq 0 \end{aligned}$$

if $0 \leq \gamma \leq \gamma_0$ with γ_0 fixed suitably small. For $x \notin \Omega_{y,r}$ we have $\varphi_{y,r}(x) = 0$ and $\mathcal{M}[\varphi_{r,y}] \geq 0$. Thus

$$\mathcal{M}[\gamma\varphi_{r,y}] + f(x,\gamma\varphi_{r,y}) \geq 0 \quad \text{in } \mathbb{R} \tag{4.20}$$

for all $0 < \gamma < \gamma_0$.

We claim that

$$\gamma_0\varphi_{r,y} \leq u \quad \text{in } \mathbb{R} \qquad \text{for all } y \in \mathbb{R}. \tag{4.21}$$

This proves the proposition because there is a positive constant c such that $\varphi_{r,y}(y) \geq c$ for all $y \in \mathbb{R}$ since the application $y \mapsto \varphi_{r,y}$ is periodic and $\varphi_{r,y}(y) > 0$ for any $y \in [-2R, 2R]$.

Now, to prove (4.21) fix $y \in \mathbb{R}$ and set

$$\gamma^* = \sup\{\, \gamma > 0 \,/\, \gamma\varphi_{r,y} \leq u \text{ in } \mathbb{R}\}.$$

Since $\inf_K u > 0$ for compact sets $K \subset \mathbb{R}$ and $\varphi_{r,y}$ has compact support we see that $\gamma^* > 0$. Assume that $\gamma^* < \gamma_0$. Then by (4.20), $\gamma^* \varphi_{r,y}$ is a subsolution of (1.1) while u is a solution. By the strong maximum principle (Theorem 1.4) either $\gamma^* \varphi_{r,y} \equiv u$ in $\mathbb{R}$ or $\inf_K (u - \gamma^* \varphi_{r,y}) > 0$ for compact sets $K \subset \mathbb{R}$. The former case is impossible because u is strictly positive, while the latter case yields a contradiction with the definition of γ^*. It follows that $\gamma^* \geq \gamma_0$ as desired. ∎

Appendix. In this appendix we give a short proof of Theorem 1.4. We assume that J satisfies (1.3), (1.4), $c \in L^\infty(\mathbb{R})$, and $u \in L^\infty(\mathbb{R})$ satisfies

$$\begin{aligned} &u \leq 0 \quad \text{a.e. in } \mathbb{R}, \\ &\mathcal{M}[u] + cu \geq 0 \quad \text{a.e. in } \mathbb{R}. \end{aligned} \tag{A.1}$$

For $\epsilon > 0$ define

$$u_\epsilon(x) = \frac{1}{2\epsilon} \int_{x-\epsilon}^{x+\epsilon} u.$$

Then u_ϵ is continuous in $\mathbb{R}$, $u_\epsilon \leq 0$, and $u_\epsilon \to u$ a.e. as $\epsilon \to 0$. There are two cases:

(1) for any closed interval I one has $\limsup_{\epsilon \to 0} \sup_I u_\epsilon < 0$, or

(2) for some closed interval I one has $\limsup_{\epsilon \to 0} \sup_I u_\epsilon = 0$.

If case (1) occurs, we see that for all closed intervals I we have $\operatorname{ess\,sup}_I u < 0$. Assume case (2) holds. Let I be a closed interval and $\epsilon_n \to 0$ be such that $\lim_{n \to +\infty} u_{\epsilon_n}(x_n) = 0$, where $x_n \in I$ is such that $\sup_I u_{\epsilon_n} = u_{\epsilon_n}(x_n)$. Integrating (A.1) from $x_n - \epsilon_n$ to $x_n + \epsilon_n$ and dividing by $2\epsilon_n$, we have

$$J \star u_{\epsilon_n}(x_n) \geq u_{\epsilon_n}(x_n) - \frac{1}{2\epsilon_n} \int_{x_n - \epsilon_n}^{x_n + \epsilon_n} cu.$$

But, since $u \leq 0$ a.e.,

$$\left| \frac{1}{2\epsilon_n} \int_{x_n - \epsilon_n}^{x_n + \epsilon_n} cu \right| \leq -\|c\|_{L^\infty} u_{\epsilon_n}(x_n) \to 0.$$

Hence

$$\liminf_{n \to +\infty} J \star u_{\epsilon_n}(x_n) \geq 0.$$

We may assume that $x_n \to x \in I$. Then by dominated convergence,

$$J \star u_{\epsilon_n}(x_n) = \int_{\mathbb{R}} J(x_n - y) u_{\epsilon_n}(y)\, dy \to \int_{\mathbb{R}} J(x - y) u(y)\, dy.$$

This shows that $u = 0$ a.e. in $x - \operatorname{supp}(J)$. Now, for any x_1 in the interior of $x - \operatorname{supp}(J)$ we have $J \star u(x_1) \geq 0$, which shows that $u = 0$ a.e. in $x - 2\operatorname{supp}(J)$, where $2\operatorname{supp}(J) = \operatorname{supp}(J) + \operatorname{supp}(J)$. Note that assumption (1.4) implies that $k \operatorname{supp}(J)$ covers all of $\mathbb{R}$ as $k \to +\infty$, where $k \operatorname{supp}(J)$ is defined inductively as $(k-1)\operatorname{supp}(J) + \operatorname{supp}(J)$. Repeating the previous argument we deduce that $u = 0$ a.e. in $\mathbb{R}$. ∎

REFERENCES

[1] P. W. Bates, P. C. Fife, X. Ren, and X. Wang, *Traveling waves in a convolution model for phase transitions*, Arch. Ration. Mech. Anal., 138 (1997), pp. 105–136.
[2] H. Berestycki, F. Hamel, and L. Roques, *Analysis of the periodically fragmented environment model*. I. *Species persistence*, J. Math. Biol., 51 (2005), pp. 75–113.
[3] H. Berestycki, F. Hamel, and L. Rossi, *Liouville-type results for semilinear elliptic equations in unbounded domains*, Ann. Mat. Pura Appl. (4), 186 (2007), pp. 469–507.
[4] H. Brezis, *Analyse Fonctionnelle. Théorie et Applications*, Masson, Paris, 1983.
[5] F. E. Browder, *On the spectral theory of elliptic differential operators*. I, Math. Ann., 142 (1960/1961), pp. 22–130.
[6] J. Carr and A. Chmaj, *Uniqueness of travelling waves for nonlocal monostable equations*, Proc. Amer. Math. Soc., 132 (2004), pp. 2433–2439.
[7] E. Chasseigne, M. Chaves, and J. D. Rossi, *Asymptotic behavior for nonlocal diffusion equations*, J. Math. Pures Appl. (9), 86 (2006), pp. 271–291.
[8] A. Chmaj and X. Ren, *Homoclinic solutions of an integral equation: Existence and stability*, J. Differential Equations, 155 (1999), pp. 17–43.
[9] J. Coville, *Maximum principles, sliding techniques and applications to nonlocal equations*, Electron. J. Differential Equations, 2007, no. 67, pp. 1–23.
[10] J. Coville and L. Dupaigne, *Propagation speed of travelling fronts in non local reaction-diffusion equations*, Nonlinear Anal., 60 (2005), pp. 797–819.
[11] D. E. Edmunds, A. J. B. Potter, and C. A. Stuart, *Noncompact positive operators*, Proc. Roy. Soc. London Ser. A, 328 (1972), pp. 67–81.
[12] P. Fife, *Mathematical Aspects of Reacting and Diffusing Systems*, Lecture Notes in Biomath. 28, Springer-Verlag, Berlin, New York, 1979.
[13] R. A. Fisher, *The Genetical Theory of Natural Selection. A Complete Variorum Edition*, H. Bennett, ed., Oxford University Press, Oxford, 1999.
[14] V. Hutson, S. Martinez, K. Mischaikow, and G. T. Vickers, *The evolution of dispersal*, J. Math. Biol., 47 (2003), pp. 483–517.
[15] Y. I. Kanel, *Certain problems on equations in the theory of burning*, Dokl. Akad. Nauk SSSR, 136 (1961), pp. 277–280 (in Russian); Soviet Math. Dokl., 2 (1961), pp. 48–51 (in English).
[16] A. N. Kolmogorov, I. G. Petrovsky, and N. S. Piskunov, *Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique*, Bulletin Université d'État à Moscow (Bjul. Moskowskogo Gos. Univ.), Série Internationale, Section A, 1 (1937), pp. 1–26; *Study of the difussion equation with growth of the quantity of matter and its application to a biological problem*, in Dynamics of Curved Fronts, Perspec. Phys., R. Pelcé, ed. Academic Press, Boston, 1988, pp. 105–130.
[17] M. Kot, M. Lewis, and P. Van den Driessche, *Dispersal data and the spread of invading organisms*, Ecology, 77 (1996), pp. 2027–2042.
[18] F. Lutscher, E. Pachepsky, and M. A. Lewis, *The effect of dispersal patterns on stream populations*, SIAM J. Appl. Math., 65 (2005), pp. 1305–1327.
[19] J. D. Murray, *Mathematical Biology*. I. *An Introduction*, 3rd ed., Interdiscip. Appl. Math. 17, Springer-Verlag, New York, 2002.
[20] R. D. Nussbaum, *The radius of the essential spectrum*, Duke Math. J., 37 (1970), pp. 473–478.
[21] K. Schumacher, *Travelling-front solutions for integro-differential equations*. I, J. Reine Angew. Math., 316 (1980), pp. 54–70.
[22] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*. I. *Fixed-Point Theorems*, Springer-Verlag, New York, 1986.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 1711–1739

STRUCTURE OF THE LINEARIZED GRAVITATIONAL VLASOV–POISSON SYSTEM CLOSE TO A POLYTROPIC GROUND STATE*

MOHAMMED LEMOU†, FLORIAN MÉHATS‡, AND PIERRE RAPHAËL§

Abstract. We deal in this paper with a generalized gravitational Vlasov–Poisson system that covers the three- and four-dimensional cases as well as the three-dimensional ultrarelativistic case. This system admits polytropic stationary solutions which are orbitally stable. We study in this paper the linear system obtained after a linearization close to these ground states and prove that the linearized flow displays at most algebraic instabilities. The heart of the proof is the derivation of a positivity property for the linearized Hamiltonian that implies a "quantitative" proof of the orbital stability statement. Our strategy follows the analysis by Weinstein [*SIAM J. Math. Anal.*, 16 (1985), pp. 472–491], who obtained similar results for the nonlinear Schrödinger equation that turned out to be fundamental preliminary properties for the further description of the fine qualitative properties of the Hamiltonian system.

Key words. gravitation, kinetic equations, Vlasov–Poisson, linearization, polytrope, stability, variational analysis

AMS subject classifications. 37K05, 37K45, 49S05, 82C40, 85A05

DOI. 10.1137/060673709

1. Introduction.

1.1. The gravitational Vlasov–Poisson system.

We consider in this paper the following generalized gravitational Vlasov–Poisson system:

$$\text{(1.1)}\quad \text{(VP)}\quad \begin{cases} \partial_t f + |v|^{\alpha-2} v \cdot \nabla_x f - E_f \cdot \nabla_v f = 0, \quad (t,x,v) \in \mathbb{R}_+ \times \mathbb{R}^N \times \mathbb{R}^N, \\ f(t=0,x,v) = f_0(x,v) \geq 0 \end{cases}$$

in the range of parameters

$$(N,\alpha) \in \{(3,1),(3,2),(4,2)\} \tag{1.2}$$

and where we denote, for a given distribution $f \geq 0$,

$$E_f(x) = \nabla_x \phi_f\,, \quad \phi(x) = -\frac{1}{N(N-2)\omega_N}\frac{1}{|x|^{N-2}} \star \rho_f, \quad \rho_f(x) = \int_{\mathbb{R}^N} f(x,v)\,dv, \tag{1.3}$$

ω_N being the volume of the unit ball in $\mathbb{R}^N$ ($\omega_3 = \frac{4\pi}{3}$ and $\omega_4 = \frac{\pi^2}{2}$). Our range of parameters (1.2) covers the three following situations.

(1) The three-dimensional gravitational Vlasov–Poisson system $(N,\alpha) = (3,2)$ which describes the mechanical state of a stellar system subject to its own gravity (see, for instance, [4, 9]) and whose classical solutions are global in time; see Lions and Perthame [28], Pfaffelmoser [33], Schaeffer [38].

*Received by the editors October 30, 2006; accepted for publication (in revised form) July 24, 2007; published electronically February 15, 2008.
http://www.siam.org/journals/sima/39-6/67370.html
†Université Paul Sabatier, MIP, F-31062 Toulouse Cedex 9, France (lemou@mip.ups-tlse.fr).
‡IRMAR, Université de Rennes 1, 35042 Rennes Cedex, France (florian.mehats@univ-rennes1.fr).
§Départment de Mathématique, Faculté des Sciences d'Orsay, Université Paris-Sud, F-91405 Orsay Cedex, France (pierre.raphael@math.u-psud.fr), and CNRS, 75794 Paris Cedex 16, France.

(2) Classical calculations show that this model should be correct only for low velocities and if high velocities occur, special relativistic corrections should be introduced; see Van Kampen and Felderhof [41] and Glassey and Schaeffer [11, 12]. A more accurate model is then provided by the relativistic three-dimensional Vlasov–Poisson system

(1.4)
$$\text{(RVP)} \quad \begin{cases} \partial_t f + \dfrac{v}{\sqrt{1+|v|^2}} \cdot \nabla_x f - E_f \cdot \nabla_v f = 0, \quad (t,x,v) \in \mathbb{R}_+ \times \mathbb{R}^3 \times \mathbb{R}^3, \\ f(t=0,x,v) = f_0(x,v) \geq 0. \end{cases}$$

A major difference with the three-dimensional (VP) is that this system may develop finite time blowup singularities (see [10, 11]), and a preliminary model problem is given by the three-dimensional ultrarelativistic (VP) system which is (1.1) with $(N,\alpha) = (3,1)$.

(3) The four-dimensional Vlasov–Poisson system $(N,\alpha) = (4,2)$ is a fundamental mathematical model for the study of the singularity formation (see [22, 23]), which shares a critical structure similar to (1.4) but admits extra fundamental invariances, and in particular an explicit pseudoconformal symmetry.

A natural space to study the (VP) system is the energy space

$$\mathcal{E} = \{f \geq 0 \ \text{ with } \ |f|_{\mathcal{E}} = |f|_{L^1} + |f|_{L^p} + ||v|^\alpha f|_{L^1} < +\infty\}$$

for $p_{crit} < p < +\infty$, where

$$p_{crit} = \frac{N\alpha + (\alpha+2)N - N^2}{2\alpha + (\alpha+2)N - N^2} = \begin{cases} 9/7 & \text{for } \ (N,\alpha) = (3,2), \\ 2 & \text{for } \ (N,\alpha) = (4,2), \\ 3/2 & \text{for } \ (N,\alpha) = (3,1). \end{cases} \tag{1.5}$$

Recall that (1.1) satisfies *formally* some conservation laws: $\forall\, q \in [1,p]$ the L^q norm of a solution f is independent of time, as well as the Hamiltonian defined by

$$\mathcal{H}(f) = \frac{1}{\alpha} \int_{\mathbb{R}^{2N}} |v|^\alpha f - \frac{1}{2} \int_{\mathbb{R}^N} |E_f|^2. \tag{1.6}$$

Moreover, a large group of symmetries leaves (1.1) invariant: if $f(t,x,v)$ solves (1.1), then $\forall (t_0, x_0, \lambda_0, \mu_0) \in \mathbb{R} \times \mathbb{R}^N \times \mathbb{R}_+^* \times \mathbb{R}_+^*$, so does

$$\frac{\mu_0^{N-\alpha}}{\lambda_0^2} f\left(\frac{t+t_0}{\lambda_0 \mu_0^{\alpha-1}}, \frac{x+x_0}{\lambda_0}, \mu_0 v \right). \tag{1.7}$$

The case $\alpha = 2$ also enjoys the Galilean invariance: if $f(t,x,v)$ solves (1.1), then $\forall v_0 \in \mathbb{R}^N$, so does $f(t, x+v_0 t, v+v_0)$.

In the classical case corresponding to $\alpha = 2$, the existence of weak solutions for (1.1) in the energy space $\mathcal{E}$ is due to Horst and Hunze [19] and Diperna and Lions [7, 8]. These solutions verify an upper bound on the Hamiltonian

$$\mathcal{H}(f(t)) \leq \mathcal{H}(f_0) \tag{1.8}$$

and the exact conservation of the L^q norm

$$\forall 1 \leq q \leq p, \qquad |f(t)|_{L^q} = |f_0|_{L^q}. \tag{1.9}$$

In the ultrarelativistic case $\alpha = 1$, we are not aware of any result concerning the Cauchy theory for (1.1).

In the energy space, we have the interpolation estimate

$$\forall f \in \mathcal{E}, \qquad |E_f|_{L^2}^2 \le C_p \, ||v|^\alpha f|_{L^1}^{\theta_1} \, |f|_{L^p}^{\theta_2} \, |f|_{L^1}^{\theta_3} \tag{1.10}$$

with

$$\theta_1 = \frac{N-2}{\alpha}, \qquad \theta_2 = \frac{(N-2)p}{N(p-1)}, \qquad \theta_3 = 2 - \theta_1 - \theta_2. \tag{1.11}$$

Note that we have $0 < \theta_i < 2$ for $i = 1, 2, 3$ in the range of parameters (1.2) and $p_{crit} < p < +\infty$. In particular, for $(N, \alpha) = (3, 2)$, $\theta_1 = \frac{1}{2}$ and thus the bound on the Hamiltonian (1.8) and the conservation of the L^1, L^p norms imply a uniform bound on the kinetic energy, hence the existence of a global weak solution to (1.1); on the contrary for $(N, \alpha) \in \{(3, 1), (4, 2)\}$, $\theta_1 = 1$ and a blowup can indeed occur from a classical virial identity; see [11]. The blowup problem in this case is of critical nature in the sense that the strength of the kinetic and the potential energy in the Hamiltonian is the same; see [22] for a further discussion on this problem.

1.2. Linear and nonlinear stability. Our aim in this paper is to study the properties of the linear flow close to a specific class of stationary solutions, the so-called polytropic ground states. This is a classical problem related to the question of the linear and nonlinear stability of the stationary solutions which has been addressed in a number of works for the case of the three-dimensional gravitational Vlasov–Poisson system (1.1) for $(N, \alpha) = (3, 2)$.

A large class of stationary solutions to (1.1) for $(N, \alpha) = (3, 2)$ of the form $f(t, x, v) = F(e)$, where $e = \frac{|v|^2}{2} + \phi(x)$, has been constructed in [2] by solving the associated nonlinear radial ODE. Two classical strategies then emerge to prove the nonlinear stability of such solutions: variational techniques for those stationary solutions that can be obtained as minimizers of a well-chosen functional; direct linearization techniques using the conservation of the Hamiltonian and coercivity properties of the linearized energy.

The first approach has been used in particular by Guo [13], Guo and Rein [14, 15, 16], and Rein [35] where part of these steady states including the polytropes have been obtained as minimizers of appropriately chosen energy-Casimir functionals under a constraint of prescribed mass. As observed in [22] (see also Sánchez and Soler [37]), a direct application of the original concentration technique introduced by Lions in [26, 27] allows one to recover the orbital stability of a two-parameter family of ground states—while the energy-Casimir technique covers only one-parameter families—in the energy space by proving the strong relative compactness up to space translation of the minimizing sequences of the problem

$$\min_{f \ge 0, \quad |f|_{L^1} = M_1, \quad |j(f)|_{L^1} = M_j} \mathcal{H}(f) \tag{1.12}$$

for a large class of convex functions j. Note that the two-parameter family is in correspondence with the two-parameter scaling invariance of the Vlasov–Poisson system. Here a difficulty arises, however, which is that uniqueness for (1.12) is known only in two special cases: (i) when $j(f) = f^p$, which is the case of polytropes where uniqueness follows directly from the scaling invariance of the polytropic equation (1.17);

(ii) when the minimizer of (1.12) is also a minimizer of the following one-constraint minimization problem:

$$\min_{f\geq 0,\quad |f|_{L^1}+|j(f)|_{L^1}=M} \mathcal{H}(f), \tag{1.13}$$

which can easily be proved to hold for a large subclass of solutions to (1.12), and then one may use Schaeffer's uniqueness result [39]. Using extra scaling invariances in the case of the polytrope, we have the following result which was proved for $\alpha = 2$ and $N = 3, 4$ in [22] and easily adapts to $(N, \alpha) = (3, 1)$.

PROPOSITION 1.1 (variational characterization of the ground state [22]). *Let (N, α) satisfy (1.2), $p \in (p_{crit}, +\infty)$, and $(\theta_i)_{1\leq i\leq 3}$ be given by (1.11). The minimization problem*

$$\inf_{f\in\mathcal{E},\, f\neq 0} \frac{||v|^\alpha f|_{L^1}^{\theta_1}\, |f|_{L^p}^{\theta_2}\, |f|_{L^1}^{\theta_3}}{|E_f|_{L^2}^2} \tag{1.14}$$

is attained on the four-parameter family

$$\gamma Q\left(\frac{x-x_0}{\lambda}, \mu v\right), \qquad (\gamma, \lambda, \mu, x_0) \in \mathbb{R}_+^* \times \mathbb{R}_+^* \times \mathbb{R}_+^* \times \mathbb{R}^N. \tag{1.15}$$

Here Q is the polytropic ground state

$$Q_{\alpha,p,N}(x,v) = \begin{cases} \left(-1 - \dfrac{|v|^\alpha}{\alpha} - \phi_Q(x)\right)^{\frac{1}{p-1}} & \text{for } \frac{|v|^\alpha}{\alpha} + \phi_Q(x) < -1, \\ 0 & \text{for } \frac{|v|^\alpha}{\alpha} + \phi_Q(x) > -1, \end{cases} \tag{1.16}$$

where ϕ_Q is the unique nontrivial radial solution to

$$-\frac{1}{r^{N-1}}\frac{d}{dr}(r^{N-1}\phi_Q') + \gamma_{\alpha,p,N}(-1-\phi_Q)_+^{\frac{1}{p-1}+\frac{N}{\alpha}} = 0, \qquad \phi(r) \to 0 \quad \text{as} \quad r \to +\infty, \tag{1.17}$$

$\gamma(\alpha, p, N)$ is given by

$$\gamma_{\alpha,p,N} = \sigma_N \int_0^1 (\alpha t)^{\frac{N-\alpha}{\alpha}} (1-t)^{\frac{1}{p-1}}\, dt. \tag{1.18}$$

For $(N, \alpha) = (3, 2)$, Q is moreover orbitally stable in the energy space by the flow of (1.1) and orbital stability up to an additional scaling invariance holds as well for $(N, \alpha) = (4, 2)$; see [22], Sánchez and Soler [37], and also Hadzic [18].

THEOREM 1.2 (orbital stability of the ground state for $(N, \alpha) = (3, 2)$). *Let $(N, \alpha) = (3, 2)$ and $p_{crit} < p < +\infty$. Then $\forall\ \eta > 0$, there exists $\delta(\eta) > 0$ such that the following holds true. Let $f_0 \in \mathcal{E}$ with*

$$\mathcal{H}(f_0) - \mathcal{H}(Q) \leq \delta(\eta), \qquad |f_0|_{L^1} \leq |Q|_{L^1} + \delta(\eta), \qquad |f_0|_{L^p} \leq |Q|_{L^p} + \delta(\eta),$$

and let $f(t) \in L^\infty([0, +\infty), \mathcal{E}_p)$ be a weak solution to (1.1) satisfying (1.9) and (1.8). Then there exists a translation shift $x(t) \in \mathbb{R}^N$ such that

$$\forall t \geq 0, \quad |f(t, x + x(t), v) - Q|_{\mathcal{E}} < \eta.$$

A similar statement holds in the critical case $(N, \alpha) = (4, 2)$ up to an additional time-dependent rescaling of the solution; see [22] for precise statements.

A different strategy to attack the question of the nonlinear stability is to consider coercivity properties of the linearized Hamiltonian as already performed in the pioneering works by Antonov [1]. Let us for simplicity restrict our attention to the case of the polytropes of Proposition 1.1. Consider the energy-Casimir functional which is formally conserved by the flow of (1.1):

$$\mathcal{H}_C(f) = \frac{|f|_{L^p}^p}{p} + |f|_{L^1} + \mathcal{H}(f) = \int_{\mathbb{R}^{2N}} \left(\frac{|f|^p}{p} + f + \frac{|v|^\alpha}{\alpha} f \right) - \frac{1}{2} \int_{\mathbb{R}^N} |E_f|^2. \tag{1.19}$$

Then this functional is continuously differentiable on $\mathcal{E}$, and Q is a critical point in the following sense: let

$$K = \mathrm{Supp}(Q) = \left\{ (x,v) \in \mathbb{R}^N \times \mathbb{R}^N \quad \text{such that} \quad \frac{|v|^\alpha}{\alpha} + \phi_Q(x) + 1 \le 0 \right\},$$

then from (1.16) and (1.19),

$$\forall f \in \mathcal{C}_0^\infty(K), \quad d\mathcal{H}_C(Q) f = \int_{\mathbb{R}^{2N}} \left(Q^{p-1} + 1 + \frac{|v|^\alpha}{\alpha} + \phi_Q \right) f = 0.$$

The Hessian on $C_0^\infty(K)$ is given by

$$d^2\mathcal{H}_C(Q)(f,f) = (p-1) \int_K Q^{p-2} f^2 - \int_{\mathbb{R}^N} |E_f|^2. \tag{1.20}$$

The understanding of the coercivity properties of this quadratic form in sufficiently strong norms will allow us to prove from a simple bootstrap argument the nonlinear stability of the polytrope. Of course this approach can be generalized to any stationary solution and in particular provides a strategy to prove nonlinear stability without any variational structure. This problem has been addressed in several places in both the physics and mathematics literature. It is known that this quadratic form will be coercive for a well-chosen class of perturbations called "admissible" perturbations; see, for example, [40, 20, 32]. A similar approach has been used recently by Guo and Rein [17] to prove conditional stability for the King-type steady states of the Vlasov–Poisson system and by Rein and Hadzic [36] for the relativistic gravitational Vlasov–Poisson system. However, this kind of structure requires us to restrict the class of the perturbation theory, whereas the perturbations authorized in the present paper are in an open set of the energy space, which contains in particular these "admissible" perturbations. A different approach was developed by Wan [42] which obtains coercivity results for a large class of quadratic forms similar to (1.20), which imply the proof of the nonlinear stability of ground states for a large class of nonvariational problems. However, the specific case of the polytropes or more generally the solutions to (1.12) do not enter this theory due to their lack of $\mathcal{C}^1$ regularity on the boundary of their domain. Eventually, the linearized Vlasov–Poisson system close to a large class of ground states was also considered in Batt, Morrison, and Rein [3], but their analysis is restricted to stationary solutions for which the quadratic form of the linearized energy-Casimir functional is the sum of two positive terms and thus directly coercive. In particular, none of the ground states obtained from variational techniques in the energy space in, for example, [16] or [22] is covered by this analysis.

Let us stress the fact that for the polytropes which have a nice variational characterization, the sharp understanding of the coercivity properties of the quadratic form (1.20) allows a *quantification* of the orbital stability statement which is crucial for the further understanding of the properties of the flow of (1.1) close to Q. By "sharp" we mean a precise understanding of the instability directions. This situation is similar to the one for the nonlinear Schrödinger equation $iu_t = -\Delta u - |u|^{p-1}u$ or the Korteweg–de Vries (KdV) equation $u_t + (u_{xx} + u^p)_x = 0$. Indeed, both these Hamiltonian systems admit, for a suitable range of the parameter p, ground-state-type stationary solutions which are orbitally stable; see Cazenave and Lions [6]. For these two systems, another proof of the orbital stability has been given by Weinstein [43, 44] by linearizing the conservation laws around the ground state and studying the coercivity properties of the obtained quadratic forms. Moreover, this work provided a preliminary investigation of the dispersive structure of the linearized operator close to the ground state. The obtained estimates are the *starting point* of a number of recent works regarding the dynamical stability of some specific solutions to these systems; see, for example, Martel, Merle, and Tsai [30] for the stability of the multisolitary waves for the KdV equation, or Bourgain and Wang [5] and Merle and Raphaël [31] for the stability of some nonlinear blowup dynamics for the nonlinear Schrödinger (NLS) system.

1.3. Statement of the results. Our aim in this paper is to adapt for (1.1) Weinstein's analysis in [43], which is the starting point for the further investigation of the nonlinear dynamics of (1.1). In a forthcoming work [24], we will, in particular, prove the existence and the stability of self-similar solutions for the three-dimensional relativistic Vlasov–Poisson system (1.4), and the proof will partly rely on the understanding of the linearized operator close to the ground state as studied in this paper. More generally, our aim, as with the NLS system, is to be able to *quantify* the orbital stability statement of Theorem 1.2, and the obtained estimates are one of the keys to further understand the *dynamical* couplings induced by the flow near Q. Let us consider the quadratic form (1.20) obtained by linearizing the energy-Casimir functional near the polytrope Q:

$$d^2\mathcal{H}_C(Q)(f,f) = (p-1)\int_K Q^{p-2}\, f^2 - \int_{\mathbb{R}^N} |E_f|^2.$$

Even though this quadratic form is not positive on its domain, we claim that we can deduce from the variational structure of Q given by Proposition 1.1 the sharp coercive structure of this quadratic form. More precisely, let us denote by

$$(f,g) = \int_K fg\, dxdv$$

the $L^2(K, dxdv)$ scalar product and consider on K the weighted L^2 measure associated with Q :

$$d\mu = Q^{p-2}\, dxdv.$$

For $f \in L^2(K, d\mu)$, we introduce the linear operator

$$\mathcal{M}f = \left((p-1)Q^{p-2}f + \phi_f\right)\mathbf{1}_K, \tag{1.21}$$

related to the quadratic form

$$(\mathcal{M}f, f) = (p-1)\int_K Q^{p-2}\, f^2 - \int_{\mathbb{R}^N} |E_f|^2, \tag{1.22}$$

and claim the following theorem.

THEOREM 1.3 (coercivity of the linearized energy-Casimir functional). *Let* (N, α) *satisfy* (1.2) *and* $p_{crit} < p < +\infty$. *Then the quadratic form* $(\mathcal{M}f, f)$ *defined by* (1.22) *is continuous and self-adjoint on* $L^2(K, d\mu)$ *and there exists a universal constant* $\delta = \delta(N, \alpha, p) > 0$ *such that* $\forall f \in L^2(K, d\mu)$, *we have*

(i) *if* $N \neq \alpha + 2$,

$$(\mathcal{M}f, f) \geq \delta \int_K f^2\, Q^{p-2} dx dv - \frac{1}{\delta}\left\{\left(f, \frac{|v|^\alpha}{\alpha} + \phi_Q\right)^2 + \sum_{i=1}^N (f, x_i)^2\right\};$$

(ii) *if* $N = \alpha + 2$,

$$(\mathcal{M}f, f) \geq \delta \int_K f^2\, Q^{p-2} dx dv - \frac{1}{\delta}\left\{\left(f, \frac{|v|^\alpha}{\alpha} + \phi_Q\right)^2 + \sum_{i=1}^N (f, x_i)^2 + (f, |v|^{2-\alpha}|x|^2)^2\right\}.$$

Following [44], Theorem 1.3 provides a *quantitative* proof of the orbital stability of the ground state Q. Let us stress again the fact that this improvement is one of the key ingredients of the *nonlinear dynamical* analysis of the three-dimensional relativistic Vlasov–Poisson system in the forthcoming work [24].

The quadratic form $(\mathcal{M}f, f)$ is intimately related to the linearized Vlasov–Poisson system which is obtained by linearizing (1.1) around Q:

$$\text{(LVP)} \quad \begin{cases} \partial_t f + \mathcal{L} f = 0, \\ f(t=0, x, v) = f_0(x, v) \end{cases} \tag{1.23}$$

with

$$\mathcal{L} f = |v|^{\alpha-2} v \cdot \nabla_x f - E_Q \cdot \nabla_v f - E_f \cdot \nabla_v Q. \tag{1.24}$$

For a specific set of initial data, the linearized energy-Casimir functional $(\mathcal{M}f, f)$ is conserved by the flow of (1.24), and this allows us to prove that the linearized system (1.23) displays at most algebraic instabilities. More precisely, consider the space

$$\mathcal{L}\mathcal{E} = \{f \in L^1_{loc}(\mathbb{R}^{2N}) \ \text{ with } \ f\mathbf{1}_K \in L^2(K, d\mu) \ \text{ and } \ f\mathbf{1}_{K^c} \in \mathcal{E}\}, \tag{1.25}$$

where $K^c = \mathbb{R}^{2N} \setminus K$. Then we have the following theorem.

THEOREM 1.4 (algebraic instability for the linearized equations). *Let* $(N, \alpha) \in \{(3, 2), (4, 2)\}$ *and* $f_0 \in \mathcal{L}\mathcal{E}$. *Then* (1.23) *admits a unique solution* $f(t) = e^{-t\mathcal{L}} f_0 \in C(\mathbb{R}_+, \mathcal{L}\mathcal{E})$. *Moreover, we have the following estimates.*

(i) *General dynamics: There holds the growth estimate*

$$\forall t \in \mathbb{R}_+, \qquad \left|e^{-t\mathcal{L}} f_0\right|_{\mathcal{L}\mathcal{E}} \leq C\,(1 + t^k)\,|f_0|_{\mathcal{L}\mathcal{E}} \tag{1.26}$$

with $k = 2$ *for* $N = 3$, $k = 3$ *for* $N = 4$.

(ii) *Dynamics on* K: *There holds a decomposition* $L^2(K, d\mu) = M \oplus S$ *where the spaces* M, S *are both invariant through the flow* $e^{-t\mathcal{L}}$ *and* S *is finite-dimensional. Moreover, we have,* $\forall t \in \mathbb{R}_+$,

$$\forall g_0 \in M, \quad \left|e^{-t\mathcal{L}} g_0\right|_{L^2(K,d\mu)} \leq C\,|g_0|_{L^2(K,d\mu)}\,, \tag{1.27}$$

$$\forall g_0 \in S, \quad \left|e^{-t\mathcal{L}} g_0\right|_{L^2(K,d\mu)} \leq \begin{cases} C\,(1+t)\,|g_0|_{L^2(K,d\mu)} & \textit{for} \quad N = 3, \\ C\,(1+t^2)\,|g_0|_{L^2(K,d\mu)} & \textit{for} \quad N = 4. \end{cases}$$

In fact, we have a complete understanding of the dispersive properties of the flow $e^{-t\mathcal{L}}$. On the support of K, the decomposition $L^2(K, d\mu) = M \oplus S$ is explicit; see Lemma 3.4. S is the so-called finite-dimensional "flag" space which contains the algebraic instabilities generated by the large group of symmetries (1.7). On the contrary, the linear dynamics are bounded on M according to (1.27). Note that *no further dispersion holds* due to the fact that the quadratic form $(\mathcal{M}f, f)$ is conserved by the flow for $\mathrm{Supp}(f) \subset K$ and K is a compact set.

Now let $\mathrm{Supp}(f_0) \subset K^c$; then the solution decomposes into a part supported on K and a part supported outside K. For this last part, the flow (1.23) reduces exactly to the linear transport by the gravitational field E_Q which is explicit. Note that the characteristic curves of this field are contained in the level sets of $e(x,v) = \frac{|v|^\alpha}{\alpha} + \phi_Q$ and are trapped for $e < 0$, hence no dispersion occurs again, and nontrapped for $e \geq 0$, hence an explicit linear dispersion holds. The part supported on K is proved to grow at most algebraically, thanks to a Gronwall-type argument, and this concludes the proof of Theorem 1.4. □

Remark 1. We have focused in this paper on the polytropic ground states only. Let us recall that the class of ground states solutions is much wider and a large set of convex functions j is known to generate a ground state $Q(j)$; see [16, 22]. If we aim at treating the case of a minimizer obtained from (1.12) for a more general convex function j, a classical difficulty will occur which is the understanding of the kernel of the linearized operator. For $j(f) = f^p$, this kernel is explicit; see Lemma 2.4. A similar statement is unknown for general j. Note that similar issues are in fact addressed in Wan [42].

2. Coercivity of the linearized energy-Casimir functional. This section is devoted to the proof of Theorem 1.3. We shall adapt to our setting the analysis by Weinstein [43]. The proof relies on two main ingredients: the variational characterization of the ground state, as given by Proposition 1.1, and the complete description of the kernel of $\mathcal{M}$. This last fact relies in part on the uniqueness of the ground state Q, which is typically a delicate problem for NLS- type equations (see Weinstein [43], Kwong [21], and Maris [29]), but it is simple in our case, thanks to the scaling invariance of (1.17).

2.1. The linearized problem for the potential. In this subsection, we study the linearized problem around ϕ_Q of the nonlinear elliptic equation (1.17). We will in particular give an explicit description of the kernel of the corresponding Schrödinger operator that implies the explicit description of the kernel of $\mathcal{M}$.

The nonlinear elliptic equation (1.17) linearized around ϕ_Q is

$$\mathcal{A}\phi = 0 \quad \text{with} \quad \mathcal{A} = -\Delta - V_Q$$

and with

$$V_Q(x) = \gamma_{\alpha,p,N}\left(\frac{N}{\alpha} + \frac{1}{p-1}\right)(-1-\phi_Q(x))_+^{\frac{N}{\alpha}-\frac{p-2}{p-1}} = \frac{1}{p-1}\int_{\frac{|v|^\alpha}{\alpha}<-\phi_Q(x)-1} \frac{dv}{Q^{p-2}(x,v)}, \tag{2.1}$$

where $\gamma_{\alpha,p,N}$ was defined by (1.18). Note that $\frac{N}{\alpha} - \frac{p-2}{p-1} > 0$ under (1.2) and thus V_Q is a continuous function with compact support on $\mathbb{R}^N$. Hence, classical operator theory (see, e.g., Reed and Simon [34, Theorems XIII.15 and XIII.12]) gives that the operator $\mathcal{A}$ on $L^2(\mathbb{R}^N)$ with the domain $H^2(\mathbb{R}^N)$ is self-adjoint, and that its spectrum can be written as

$$\sigma(\mathcal{A}) = \{\lambda_i < 0\}_{1\leq i\leq I} \cup [0,+\infty), \tag{2.2}$$

where $\{\lambda_i < 0\}_{1\leq i\leq I}$ is the finite set of nonpositive eigenvalues with finite multiplicity and $[0,+\infty)$ is the essential spectrum of $\mathcal{A}$. 0 may be an eigenvalue. We shall denote by $(\psi_j)_{1\leq j\leq J}$, $|\psi_j|_{L^2} = 1$, the finite set of eigenvectors associated with the nonpositive eigenvalues that are well localized in space from standard argument.

Let $\dot{H}^1$ be the completion of $C_0^\infty(\mathbb{R}^N)$ with respect to the norm $|u|_{\dot{H}^1} = |\nabla u|_{L^2}$, or equivalently,

$$\dot{H}^1 = \left\{\phi \in L^1_{loc}(\mathbb{R}^N) : \quad \frac{\phi}{\sqrt{1+|x|^2}} \in L^2(\mathbb{R}^N) \text{ and } \nabla\phi \in L^2(\mathbb{R}^N)\right\}.$$

We have the following coercivity property.

LEMMA 2.1 (coercivity of the linearized problem close to ϕ_Q). *Let (α, N, p) be as in Theorem* 1.3. *Then the set of functions $\phi \in \dot{H}^1$ such that $\mathcal{A}\phi = 0$ in the distributional sense coincides with the kernel of $\mathcal{A}$ which can be characterized as*

$$\mathrm{Ker}(\mathcal{A}) = \mathrm{span}\{\partial_{x_i}\phi_Q\}_{1\leq i\leq N}. \tag{2.3}$$

Moreover, there exists $c_0 > 0$ such that $\forall \phi \in \dot{H}^1$ with radial symmetry,

$$\int_{\mathbb{R}^N} |\nabla\phi|^2 - \int_{\mathbb{R}^N} V_Q\,|\phi|^2 \geq c_0 \int |\nabla\phi|^2 - \frac{1}{c_0}\sum_{j=1}^{J}(\phi,\psi_j)^2. \tag{2.4}$$

Proof of Lemma 2.1. We follow Weinstein's strategy [43, proof of Proposition 2.8b]; see also Maris [29].

Step 1. Decomposition into spherical harmonics. Let $\phi \in \mathrm{Ker}(\mathcal{A})$. Then

$$-\Delta\phi = V_Q\phi \quad \text{with} \quad \phi \in \dot{H}^1 \hookrightarrow L^{\frac{2N}{N-2}}; \tag{2.5}$$

hence $\phi \in C^2(\mathbb{R}^N)$ from standard elliptic theory. One can thus decompose ϕ into spherical harmonics. More precisely, let $\mathcal{P}_k$ be the space of spherical harmonics of degree k, with dim $\mathcal{P}_k = a_k = C^k_{N+k-1} - C^k_{N+k-3}$, and for each k, let $\{Y_i^k\}_{1\leq i\leq a_i}$ be the L^2 orthonormal basis of $\mathcal{P}_k$. Then ϕ has a unique expansion

$$\phi(x) = \sum_{k=0}^{+\infty}\sum_{i=1}^{a_k} \varphi_{k,i}(|x|)\, Y_i^k\left(\frac{x}{|x|}\right) \tag{2.6}$$

with

$$\varphi_{k,i}(|x|) = \int_{S^{N-1}} \phi(|x|\theta)Y_i^k(\theta)d\theta \ \to 0 \ \text{ as } |x| \to +\infty. \tag{2.7}$$

Since the potential V_Q has radial symmetry, (2.5) implies

$$\mathcal{A}_k\,\varphi_{k,i} = 0 \qquad \text{with } \mathcal{A}_k = -\frac{d^2}{dr^2} - \frac{N-1}{r}\frac{d}{dr} + \frac{k(k+N-2)}{r^2} - V_Q(r). \tag{2.8}$$

Let $r_Q > 0$ be the unique solution to

$$\phi_Q(r_Q) = -1. \tag{2.9}$$

Then the potential V_Q is compactly supported on $[0, r_Q]$. Hence one can solve (2.8) explicitly outside its support with the constraint $\varphi_{k,i} \in L^\infty(\mathbb{R})$ deduced from $\phi \in L^\infty(\mathbb{R}^N)$ and (2.6). We get

$$\forall k \geq 0, \quad \forall i \in \{1, \ldots, a_k\}, \quad \forall r > r_Q, \qquad \varphi_{k,i}(r) = \frac{C_{k,i}}{r^{k+N-2}} \tag{2.10}$$

for some constant $C_{k,i}$.

Observe now that (2.3) is equivalent to

$$\varphi_{k,i} = 0 \;\text{ when }\; k \neq 1 \;\text{ and }\; \varphi_{1,i}(r) = a_{1,i}\, \phi_Q'(r) \;\text{ for }\; 1 \leq i \leq N. \tag{2.11}$$

Step 2. The case $k \geq 1$. Let $k \geq 1$. Observe that (2.10) implies $\phi_{k,i} \in H_r^1$, where H_r^1 denotes the set of H^1 distributions of $\mathbb{R}^N$ with radial symmetry. We now take the derivative of (1.17) with respect to the radial coordinate r and get after direct calculations

$$\mathcal{A}_1\, \phi_Q' = 0.$$

Therefore ϕ_Q' is an eigenfunction of $\mathcal{A}_1$ corresponding to the eigenvalue zero. Observe from (1.17) that ϕ_Q' is nonnegative on $(0, +\infty)$ and it follows from standard spectral analysis [34] that it is the ground state of $\mathcal{A}_1$ on H_r^1. We conclude that

$$\forall w \in H_r^1, \qquad (\mathcal{A}_1 w, w) \geq 0 \qquad \text{and} \qquad (\mathrm{Ker}\mathcal{A}_1)_{H_r^1} = \mathrm{span}(\phi_Q'), \tag{2.12}$$

and the case $k = 1$ of (2.11) follows.

For $k \geq 2$, we have from (2.8),

$$0 = (\mathcal{A}_k \phi_{k,i}, \phi_{k,i}) = (\mathcal{A}_1 \phi_{k,i}, \phi_{k,i}) + (k(k+N-2) - (N-1)) \int_0^{+\infty} |\phi_{k,i}|^2\, r^{N-3}\, dr,$$

which gives $\phi_{k,i} = 0$, thanks to the positivity of $\mathcal{A}_1$ (2.12), and the case $k \geq 2$ in (2.10) is also solved.

Step 3. The case $k = 0$. The remaining case $k = 0$ has to be treated in a different way. The fact that $\varphi_0 = 0$ is a consequence of the scaling structure of the ϕ_Q equation (1.17). Indeed, φ_0 solves

$$\mathcal{A}_0\, \varphi_0 = \left(-\frac{d^2}{dr^2} - \frac{N-1}{r}\frac{d}{dr} - V_Q \right) \varphi_0 = 0 \tag{2.13}$$

and, by (2.10), satisfies $\varphi_0(r) = \frac{C_0}{r^{N-2}}$ for r large enough. In particular,

$$\varphi_0(r) \to 0 \;\text{ as }\; r \to +\infty. \tag{2.14}$$

Note also that from the C^2 regularity of ϕ, $\varphi_0'(0) = 0$.

Now let

$$h = -1 - \phi_Q \;\text{ and }\; \beta = \frac{2\alpha(p-1)}{\alpha + (N-\alpha)(p-1)}. \tag{2.15}$$

We claim that

$$\mathcal{A}_0\, H = 0 \qquad \text{with } H(r) = \beta h(r) + r h'(r). \tag{2.16}$$

Indeed, for $\lambda > 0$, let $h_\lambda(r) = \lambda^\beta h(\lambda r)$. From (1.17) and the above choice of β, h_λ solves

$$\left(-\frac{d^2}{dr^2} - \frac{N-1}{r}\frac{d}{dr}\right) h_\lambda - \gamma_{\alpha,p,N}\,(h_\lambda)_+^{\frac{1}{p-1}+\frac{N}{\alpha}} = 0.$$

Differentiating this expression with respect to λ and evaluating the result at $\lambda = 1$ yield (2.16). We now observe from (2.10), (2.15), and (2.16) that $H(r) \to -\beta \neq 0$ as $r \to +\infty$. From standard ODE analysis, all the solutions to (2.13) with a vanishing derivative at $r = 0$ are proportional. Since $H'(0) = -(\beta+1)\phi_Q'(0) = 0$, both functions φ_0 and H are proportional, which implies by (2.14) that φ_0 is identically zero and concludes the proof of (2.11). The proof of (2.3) is now complete.

Step 4. *Proof of* (2.4). We now conclude the proof of (2.4), which follows from standard variational arguments. We briefly sketch the proof for the sake of completeness. Let the quadratic form

$$(\mathcal{A}\phi, \phi) = \int |\nabla\phi|^2 dx - \int V_Q\,|\phi|^2 dx$$

be continuous on $\dot{H}^1$ since V_Q is compactly supported. Let Λ be the set of $\phi \in \dot{H}^1$ with radial symmetry such that

$$(\phi, \psi_j) = 0 \qquad \text{for } j = 1, \dots, J. \tag{2.17}$$

Note that these L^2 scalar products are well defined since the ψ_k's are well localized in space. The spectral property (2.2) implies

$$\forall \phi \in \Lambda \cap L^2(\mathbb{R}^N), \qquad (\mathcal{A}\phi, \phi) \geq 0. \tag{2.18}$$

From a standard density argument, (2.18) holds also on Λ. We claim that, in fact,

$$\inf_{\phi\in\Lambda,\,|\nabla\phi|_{L^2}=1} (\mathcal{A}\phi, \phi) > 0. \tag{2.19}$$

This, together with the continuity of the quadratic form $(\mathcal{A}\phi, \phi)$ on $\dot{H}^1$, now implies (2.4).

Proof of (2.19). We argue by contradiction and consider a sequence ϕ_n such that

$$\phi_n \in \Lambda, \qquad (\mathcal{A}\phi_n, \phi_n) \to 0 \qquad \text{and} \qquad |\nabla\phi_n|_{L^2} = 1. \tag{2.20}$$

Up to a subsequence, $\phi_n \rightharpoonup \phi$ in $\dot{H}^1_r$. Moreover, since the Sobolev embedding $\dot{H}^1 \hookrightarrow L^{\frac{2N}{N-2}}$ is locally compact, we have $\phi \in \Lambda$ and

$$1 - (\mathcal{A}\phi_n, \phi_n) = \int V_Q\,|\phi_n|^2 \to \int V_Q\,|\phi|^2 = 1 \qquad \text{as } n \to +\infty.$$

By lower semicontinuity, $|\nabla\phi|_{L^2} \leq 1$ and thus $(\mathcal{A}\phi, \phi) \leq 0$. Since $\phi \in \Lambda$, this implies

$$(\mathcal{A}\phi, \phi) = 0, \qquad |\nabla\phi|_{L^2} = 1, \tag{2.21}$$

and the convergence $\phi_n \to \phi$ holds in the strong $\dot{H}^1$ topology. Hence $\inf(\mathcal{A}\phi, \phi)$ is attained and the Euler–Lagrange equation of this constrained variational problem reads

$$-\lambda\Delta\phi - V_Q\phi = \sum_{j=1}^{J} b_j\,\psi_j. \tag{2.22}$$

We take the L^2 inner product of (2.22) with ϕ and get $\lambda|\nabla\phi|^2_{L^2} = \int V_Q\,|\phi|^2 = 1$, thanks to the orthogonality conditions (2.20). Thus $\lambda = 1$ and (2.22) becomes

$$\mathcal{A}\phi = \sum_{j=1}^{J} b_j\,\psi_j. \tag{2.23}$$

Taking the scalar product of (2.23) with ψ_{j_0} now gives

$$b_{j_0} = (\mathcal{A}\phi, \psi_{j_0}) = \lambda_{j_0}(\phi, \psi_{j_0}) = 0,$$

where we also used (2.17) again. Hence $b_{j_0} = 0$ and $\phi \in \mathrm{Ker}(\mathcal{A})$ from (2.23). It remains to remark that (2.3) and the radial symmetry of ϕ imply that $\phi = 0$, which contradicts (2.21). The proof of (2.19) is complete. $\square$

This concludes the proof of Lemma 2.1. $\square$

2.2. Variational estimates and proof of Theorem 1.3. In this subsection, we study the linear operator $\mathcal{M}$, defined by (1.21) on $L^2(K, d\mu)$, and prove Theorem 1.3.

Let us start with the following continuity result.

LEMMA 2.2 (continuity of $\mathcal{M}$ on $L^2(K, d\mu)$). *Let (α, N, p) be as in Theorem 1.3. Then the quadratic form $(\mathcal{M}f, f)$ is continuous and self-adjoint on $L^2(K, d\mu)$. Moreover, let a sequence $f_n \in L^2(K, d\mu)$ be such that*

$$f_n \rightharpoonup f \quad \text{in } L^2(K, d\mu). \tag{2.24}$$

Then

$$E_{f_n} \to E_f \quad \text{in } L^2(\mathbb{R}^N). \tag{2.25}$$

Proof of Lemma 2.2. Let the potential V_Q be as given by (2.1). Then from the Cauchy–Schwarz inequality, $\forall f \in L^2(K, d\mu)$,

$$|\rho_f(x)| = \left|\int_K f(x,v)\,dv\right| \le \left((p-1)\int_K f^2(x,v)\,Q^{p-2}dv\right)^{1/2} (V_Q(x))^{1/2}. \tag{2.26}$$

Observe now that the potential V_Q is a continuous function with compact support on $\mathbb{R}^N$. Thus (2.26) implies $\rho_f \in L^1 \cap L^2(\mathbb{R}^N)$ with $\mathrm{Supp}(\rho_f) \subset \{|x| \le r_Q\}$ (r_Q is defined by (2.9)). Sobolev embeddings now imply $E_f \in L^2(\mathbb{R}^N)$ and the continuity of $\mathcal{M}$ on $L^2(K, d\mu)$ follows. The fact that $\mathcal{M}$ is self-adjoint follows from integration by parts.

Now let a sequence f_n satisfy (2.24). The estimate (2.26) gives an $L^1 \cap L^2$ bound for ρ_{f_n}, and thus E_{f_n} is locally compact in $L^2(\mathbb{R}^N)$ from Sobolev embeddings. Observe now that $|x| > 2r_Q$ and $|x-y| < r_Q$ imply $|y| > |x| - r_Q > \frac{|x|}{2}$, and thus

$$|E_f(x)| \le C\int_{|x-y|\le r_Q} \frac{|\rho_f(x-y)|}{|y|^{N-1}}\,dy \le C\frac{2^{N-1}}{|x|^{N-1}}|\rho_f|_{L^1}.$$

We conclude that E_{f_n} is L^2 compact and (2.25) follows. This concludes the proof of Lemma 2.2. $\square$

We now claim the following positivity property for $\mathcal{M}$ that is a consequence of the variational characterization of Q as given by Proposition 1.1 and is the very heart of the proof of Theorem 1.3.

LEMMA 2.3 (positivity of $\mathcal{M}$ induced by the variational structure of Q). *Let (α, N, p) be as in Theorem* 1.3. *Let $f \in L^2(K, d\mu)$ with*

$$\left(f, \frac{|v|^\alpha}{\alpha} + \phi_Q\right) = 0. \tag{2.27}$$

Then the quadratic form defined by (1.22) *satisfies*

$$(\mathcal{M}f, f) \geq 0. \tag{2.28}$$

Proof of Lemma 2.3. Let $f \in C_0^\infty(K)$. Then for any η small enough, $Q + \eta f \in \mathcal{E}$. Let $J_{\alpha,p,N}$ be the functional defined by (1.14) and denote $J(\eta) = J_{\alpha,p,N}(Q+\eta f)$. The variational characterization of Q given by Proposition 1.1 implies

$$J'(0) = 0, \qquad J''(0) \geq 0. \tag{2.29}$$

Now, from a direct computation using the identities

$$\omega := \frac{\alpha\theta_1}{||v|^\alpha Q|_{L^1}} = \frac{\theta_2}{|Q|_{L^p}^p} = \frac{\theta_3}{|Q|_{L^1}} = \frac{2}{|E_Q|_{L^2}^2}$$

obtained during the construction of Q (see [22]), we get

$$\begin{aligned}(\mathcal{M}f, f) = \frac{J''(0)}{\omega J(0)} + \omega \left\{ \frac{\alpha+2-N}{N-2} \left(f, \frac{|v|^\alpha}{\alpha}\right)^2 + \frac{p}{\theta_2}\left(f, Q^{p-1}\right)^2 + \frac{1}{\theta_3}(f,1)^2 \right\} \\ + \omega \left(f, \frac{|v|^\alpha}{\alpha} + \phi_Q\right)\left(f, \frac{|v|^\alpha}{\alpha} - \phi_Q\right).\end{aligned}$$

Note that $3 \leq N \leq \alpha + 2$ in the range of parameters (1.2), and thus (2.29) and the orthogonality condition (2.27) now imply (2.28). The general case $f \in L^2(K, d\mu)$ follows by density. This concludes the proof of Lemma 2.3. □

The second key to the proof of Theorem 1.3 is the fact that the kernel of $\mathcal{M}$ is explicit and in particular $\mathcal{M}$ is invertible when restricted to radially symmetric distributions. Moreover, some inverses are explicit as a consequence of the action of the large group of symmetries (1.7).

LEMMA 2.4 (explicit description of the kernel of $\mathcal{M}$). *Let (α, N, p) be as in Theorem* 1.3. *Then*

$$\mathrm{Ker}\,(\mathcal{M}) = \left\{f \in L^2(K, d\mu) \text{ with } \mathcal{M}f = 0\right\} = \mathrm{span}\{\partial_{x_i} Q\}_{1\leq i\leq N}. \tag{2.30}$$

Moreover, let

$$S_1 = \frac{N-\alpha}{2}\, x \cdot \nabla_x Q - v \cdot \nabla_v Q\,, \qquad S_2 = -x \cdot \nabla_v Q\,, \qquad S_3 = \mathcal{M}^{-1}\left(\frac{|v|^{2-\alpha}|x|^2}{2}\right). \tag{2.31}$$

Then we have the following identities:

$$\mathcal{M}S_1 = \alpha\left(\frac{|v|^\alpha}{\alpha} + \phi_Q\right)\mathbf{1}_K\,, \qquad \mathcal{M}S_2 = x \cdot v\,|v|^{\alpha-2}\,\mathbf{1}_K\,, \qquad \mathcal{M}S_3 = \frac{|v|^{2-\alpha}|x|^2}{2}\,\mathbf{1}_K. \tag{2.32}$$

Proof of Lemma 2.4.

Step 1. Description of Ker($\mathcal{M}$). We claim that

$$\operatorname{span}\{\partial_{x_i}Q\}_{1\le i\le N}\subset \operatorname{Ker}(\mathcal{M}).$$

Indeed, we rewrite (1.16) of Q as

$$\left(Q^{p-1}+\frac{|v|^\alpha}{\alpha}+\phi_Q+1\right)\mathbf{1}_K=0 \tag{2.33}$$

and take a derivative with respect to $(x_i)_{1\le i\le N}$ to derive

$$\left((p-1)Q^{p-2}\partial_{x_i}Q+\partial_{x_i}\phi_Q\right)\mathbf{1}_K=0. \tag{2.34}$$

In particular,

$$\int_{\mathbb{R}^{2N}}Q^{p-2}\,(\partial_{x_i}Q)^2=\frac{1}{p-1}\int_{\mathbb{R}^N}(\partial_{x_i}\phi_Q)^2\,V_Q<+\infty,$$

where V_Q is the potential defined by (2.1), and thus $\partial_{x_i}Q\in L^2(K,d\mu)$. Using also $\partial_{x_i}\phi_Q=\phi_{\partial_{x_i}Q}$, we deduce from (2.34) that $\partial_{x_i}Q\in\operatorname{Ker}(\mathcal{M})$.

Now let $f\in\operatorname{Ker}(\mathcal{M})$. For $(x,v)\in K$ we have

$$(p-1)Q^{p-2}f(x,v)+\phi_f(x)=0 \tag{2.35}$$

and thus, $\forall |x|\le r_Q$,

$$\Delta\phi_f(x)=\rho_f(x)=-\frac{1}{p-1}\int_K\frac{\phi_f(x)}{Q^{p-2}(x,v)}\,dv=-\phi_f(x)\,V_Q(x). \tag{2.36}$$

For $|x|>r_Q$, $\rho_f(x)=0$, so (2.36) still holds. We conclude that ϕ_f belongs to the kernel of the operator $\mathcal{A}=-\Delta-V_Q$ on $\dot H^1$. By Lemma 2.1, there exists $\{c_i\}_{1\le i\le N}$ such that

$$\phi_f=\sum_{i=1}^N c_i\,\partial_{x_i}\phi_Q.$$

It follows from (2.34) and (2.35) that

$$f=\sum_{i=1}^N c_i\,\partial_{x_i}Q$$

and this concludes the proof of (2.30).

Step 2. Derivation of the algebraic identities (2.32). The first identity in (2.32) is a consequence of the scaling invariance of (2.33). For a parameter $\mu>0$, define

$$Q_\mu(x,v)=\mu^{N-\alpha}Q(x,\mu v).$$

Then the corresponding microscopic energy defined by

$$e(x,v)=\frac{|v|^\alpha}{\alpha}+\phi_Q(x)$$

scales according to $e_\mu(x,v) = \frac{1}{\mu^\alpha}\, e(x,\mu v)$. We thus compute from (2.33)

$$\left(\mu^{-(p-1)(N-\alpha)}\, Q_\mu^{p-1} + \mu^\alpha e_\mu(x,v) + 1\right) \mathbf{1}_{(x,\mu v)\in K} = 0. \tag{2.37}$$

Differentiating this relation with respect to μ and evaluating the result at $\mu = 1$ yield

$$(p-1)Q^{p-2}R - (p-1)(N-\alpha)Q^{p-1} + |v|^\alpha = 0 \quad \text{on} \quad K,$$

where we have denoted

$$R = (N-\alpha)Q + v\cdot\nabla_v Q = \frac{dQ_\mu}{d\mu}_{|\mu=1}.$$

Noting that

$$\phi_R = -\alpha\phi_Q \quad \text{from} \quad \rho_R = \int ((N-\alpha)Q + v\cdot\nabla_v Q)\, dv = -\alpha\rho_Q,$$

we get the following intermediate identity:

$$\mathcal{M}R = (p-1)(N-\alpha)Q^{p-1} - \alpha\left(\frac{|v|^\alpha}{\alpha} + \phi_Q\right). \tag{2.38}$$

We now rescale the x variable in Q and set for $\lambda > 0$:

$$Q_\lambda(x,v) = \frac{1}{\lambda^2}\, Q\left(\frac{x}{\lambda}, v\right).$$

The microscopic energy scales according to $e_\lambda(x,v) = e\left(\frac{x}{\lambda}, v\right)$ and (2.33) becomes

$$\left(\lambda^{2(p-1)}\, Q_\lambda^{p-1} + e_\lambda(x,v) + 1\right) \mathbf{1}_{(x/\lambda,v)\in K} = 0.$$

We differentiate this expression with respect to λ and evaluate the result at $\lambda = 1$ to get

$$-(p-1)Q^{p-2}\widetilde{R} + 2(p-1)Q^{p-1} - x\cdot\nabla_x\phi_Q = 0 \qquad \text{on } K,$$

where we have denoted

$$\widetilde{R} = 2Q + x\cdot\nabla_x Q = -\frac{dQ_\lambda}{d\lambda}_{|\lambda=1}.$$

Now noting that

$$\phi_{\widetilde{R}} = x\cdot\nabla_x\phi_Q,$$

we get the second intermediate inequality

$$\mathcal{M}\widetilde{R} = 2(p-1)Q^{p-1}. \tag{2.39}$$

Multiplying (2.39) by $\frac{N-\alpha}{2}$ and subtracting (2.38) yield the first identity in (2.32).

The second identity in (2.32) can be proved by a direct computation, noting simply that

$$\phi_{S_2} = 0 \qquad \text{from} \quad \rho_{S_2} = -\int x\cdot\nabla_v Q dv = 0$$

and that by (1.16)

$$\nabla_v Q = -\frac{1}{p-1}\,\frac{v\,|v|^{\alpha-2}}{Q^{p-2}}. \tag{2.40}$$

Finally, the last identity in (2.32) is obvious as soon as we are able to define S_3 according to (2.31). To this aim, we recall that since Q is radially symmetric, we have $\int_{\mathbb{R}^{2N}} |v|^{2-\alpha}|x|^2\partial_{x_i} Q = 0$ for $1 \le i \le N$, and (2.30) implies that $\frac{|v|^{2-\alpha}|x|^2}{2} \in \mathrm{Ker}(\mathcal{M})^{\perp}$. Hence Lemma 2.3 and the Lax–Milgram theorem ensure the invertibility of $\mathcal{M}$ on $\mathrm{Ker}(\mathcal{M})^{\perp}$ and (2.31) defines S_3 in $L^2(K, d\mu)$ without ambiguity.

This concludes the proof of Lemma 2.4. ◻

We are now in position to conclude the proof of Theorem 1.3, which follows from standard variational techniques.

Proof of Theorem 1.3. Let $\mathcal{I}$ be the set of $f \in L^2(K, d\mu)$ with

$$\left\{ \begin{array}{ll} \left(f, \dfrac{|v|^\alpha}{\alpha} + \phi_Q\right) = (f, x_i) = 0, \quad 1 \le i \le N, & \text{if } N-\alpha \neq 2, \\ \left(f, \dfrac{|v|^\alpha}{\alpha} + \phi_Q\right) = (f, x_i) = (f, |v|^{2-\alpha}|x|^2) = 0, \quad 1 \le i \le N, & \text{if } N-\alpha = 2. \end{array} \right. \tag{2.41}$$

Note that the L^2 inner products are well defined as, $\forall f \in L^2(K, d\mu)$ and $g \in L^\infty(K)$,

$$\begin{aligned} |(f,g)| &\le |g|_{L^\infty}\left(\int_K f^2\,Q^{p-2}\,dxdv\right)^{1/2}\left(\int_K \frac{1}{Q^{p-2}}\,dxdv\right)^{1/2} \\ &\le C|g|_{L^\infty}\,|f|_{L^2(K,d\mu)}\,|V_Q|_{L^\infty}^{1/2} \end{aligned}$$

with V_Q defined by (2.1). We now claim that

$$I = \inf_{f\in\mathcal{I},\ |f|_{L^2(K,d\mu)}=1} (\mathcal{M}f, f) > 0. \tag{2.42}$$

Since $(\mathcal{M}\cdot,\cdot)$ is a continuous quadratic form, (2.42) and the definition (2.41) of $\mathcal{I}$ imply the coercivity property of Theorem 1.3.

Proof of (2.42). Arguing by contradiction and using the positivity property of Lemma 2.3, we let a sequence f_n be such that

$$f_n \in \mathcal{I}, \qquad \int_K f_n^2\,Q^{p-2} = 1 \quad\text{and}\quad (\mathcal{M}f_n, f_n) \to 0 \quad\text{as}\quad n \to +\infty. \tag{2.43}$$

Up to a subsequence, $f_n \rightharpoonup f$ in $L^2(K, d\mu)$ and thus $f \in \mathcal{I}$ with

$$\int_K f^2\,Q^{p-2}\,dxdv \le 1. \tag{2.44}$$

Now by Lemma 2.2, $E_{f_n} \to E_f$ in $L^2(\mathbb{R}^N)$. Since, by (2.43), we have $|E_{f_n}|_{L^2}^2 \to p-1$, we deduce that $|E_f|_{L^2}^2 = p-1$ and thus $f \neq 0$. Moreover, (2.44) implies $(\mathcal{M}f, f) \le 0$ and thus $(\mathcal{M}f, f) = 0$ from Lemma 2.3 and $f \in \mathcal{I}$. This implies $\int_K f^2\,Q^{p-2} = 1$ and the infimum I defined by (2.42) is attained at f.

We now write down the Euler–Lagrange equation for this constrained minimization problem and get the existence of Lagrange multipliers β, $(\gamma_i)_{1\le i\le N}$, κ, τ such that

$$\mathcal{M}f = \beta\,Q^{p-2}f + \sum_{i=1}^N \gamma_i\,x_i\,\mathbf{1}_K + \kappa\left(\frac{|v|^\alpha}{\alpha} + \phi_Q\right)\mathbf{1}_K \qquad \text{for } N-\alpha \neq 2, \tag{2.45}$$

(2.46)

$$\mathcal{M}f = \beta\, Q^{p-2} f + \sum_{i=1}^{N} \gamma_i\, x_i\, \mathbf{1}_K + \kappa \left(\frac{|v|^\alpha}{\alpha} + \phi_Q \right) \mathbf{1}_K + \tau\, |v|^{2-\alpha} |x|^2\, \mathbf{1}_K \qquad \text{for } N-\alpha = 2.$$

Take the $L^2(K, dxdv)$ inner product of (2.45) or (2.46) with f, use $(\mathcal{M}f, f) = 0$, the orthogonality conditions (2.41), and $\int f^2 Q^{p-2} dxdv = 1$ to obtain $\beta = 0$. Then take the inner product of (2.45) or (2.46) with $\partial_{x_i} Q$ for $1 \le i \le N$ to get

$$-\gamma_i \int Q = (\mathcal{M}f, \partial_{x_i} Q) = (f, \mathcal{M}\partial_{x_i} Q) = 0,$$

where we used (A.2) from the appendix and (2.30). Hence the γ_i's are all zero.

Now let $N-\alpha \neq 2$ and take the inner product of (2.45) with S_1 defined by (2.31). Using (2.32), we get

$$\kappa \left(\frac{|v|^\alpha}{\alpha} + \phi_Q, S_1 \right) = (\mathcal{M}f, S_1) = (f, \mathcal{M}S_1) = \alpha \left(f, \frac{|v|^\alpha}{\alpha} + \phi_Q \right) = 0.$$

Since, by (A.1) given in the appendix, in the subcritical case $N-\alpha \neq 2$ the factor of κ is not zero, we deduce that $\kappa = 0$. Hence $\mathcal{M}f = 0$. From (2.30), we thus have the existence of $(c_i)_{1 \le i \le n}$ such that

$$f = \sum_{i=1}^{N} c_i\, \partial_{x_i} Q.$$

Multiplying this expression by x_i and integrating, we deduce from (2.41) that $c_i = 0$ for $i = 1, \ldots, N$, and thus $f = 0$, which which is absurd.

Let $N-\alpha = 2$. Taking the inner product of (2.46) with S_1 and using (A.1), (2.32), and (2.41), we get

$$0 = (f, \mathcal{M}S_1) = (\mathcal{M}f, S_1) = \tau(|v|^{2-\alpha}|x|^2, S_1) = -\alpha\tau \int |v|^{2-\alpha}|x|^2 Q \quad \text{and thus} \quad \tau = 0.$$

Now take the inner product of (2.46) with S_3 (defined by (2.31)). By the orthogonality condition (2.41) we have

$$\kappa \left(\frac{|v|^\alpha}{\alpha} + \phi_Q, S_3 \right) = (\mathcal{M}f, S_3) = (f, \mathcal{M}S_3) = \left(f, \frac{|v|^{2-\alpha}|x|^2}{2} \right) = 0,$$

so we deduce from (A.5) that $\kappa = 0$ and $\mathcal{M}f = 0$. The end of the proof is then identical to the case $N-\alpha \neq 2$.

This concludes the proof of Theorem 1.3. ∎

3. The linearized Vlasov–Poisson system in dimension 3 or 4. In this section, we fix $\alpha = 2$ and take $N = 3$ or $N = 4$ and prove Theorem 1.4.

3.1. Well-posedness of the linearized Vlasov–Poisson system. We prove in this subsection the well-posedness of the linearized equation (1.23) and some conservation laws associated with this flow.

We start with a technical lemma stating a few useful properties of the $\mathcal{LE}$ space defined by (1.25) that is a natural space for the study of the linearized Vlasov–Poisson problem.

LEMMA 3.1 (embedding of the $\mathcal{E}$ space). *The space $\mathcal{LE}$ is continuously embedded into $L^1(\mathbb{R}_x^N \times \mathbb{R}_v^N) \cap L^{\frac{2N}{N+2}}(\mathbb{R}_x^N, L^1(\mathbb{R}_v^N))$. Moreover, there exists a constant C such that*

$$\forall f \in \mathcal{LE}, \qquad |E_f|_{L^2} \le C|f|_{\mathcal{LE}}. \tag{3.1}$$

Proof of Lemma 3.1. Let $f \in \mathcal{LE}$ and decompose this function into $f = f^i + f^e$ with $f^i = f\mathbf{1}_K \in L^2(K, d\mu)$ and $f^e = f\mathbf{1}_{K^c} \in \mathcal{E}$. By (2.26) we have

$$|f^i|_{L^1} + |f^i|_{L_x^2 L_v^1} \le C|f^i|_{L^2(K,d\mu)},$$

and, by the standard interpolation inequality,

$$|f^e|_{L^1} + |f^e|_{L_x^q L_v^1} \le C|f^e|_{\mathcal{E}}$$

with $q = \frac{(N+2)p-N}{Np-N+2}$. The assumption $p > p_{crit}$ ensures that $q > \frac{2N}{N+2}$, thus, noting also that $2 > \frac{2N}{N+2}$, we get

$$|f|_{L^1} + |f|_{L_x^{\frac{2N}{N+2}} L_v^1} \le C|f|_{\mathcal{LE}}.$$

The estimate (3.1) of the field now follows from the Poisson equation and the generalized Young inequality. This concludes the proof of Lemma 3.1. □

We now state the well-posedness of the linearized Vlasov–Poisson system (1.23) that is the main result of this subsection.

PROPOSITION 3.2 (properties of the linearized flow in $\mathcal{E}$). *Let $(N, \alpha) = (3, 2)$ or $(4, 2)$. Let $f_0 \in L^1(\mathbb{R}^N \times \mathbb{R}^N)$. Then (1.23) admits a unique weak solution $e^{-t\mathcal{L}} f_0 \in C(\mathbb{R}_+, L^1(\mathbb{R}^N \times \mathbb{R}^N))$. Assume, moreover, that $f_0 \in \mathcal{LE}$ and decompose it as follows:*

$$f_0 = f_0^i + f_0^e \quad \text{with} \quad f_0^i = f_0\, \mathbf{1}_K\,, \quad f_0^e = f_0\, \mathbf{1}_{K^c}. \tag{3.2}$$

Let $f^i(t) = e^{-t\mathcal{L}} f_0^i$, $f^e(t) = e^{-t\mathcal{L}} f_0^e$. Then $\forall t \ge 0$, we have $f(t) = e^{-t\mathcal{L}} f_0 = f^i(t) + f^e(t) \in \mathcal{LE}$ and the following conservation laws hold: $\forall t \ge 0$,

$$\text{Supp } f^i(t) \subset K, \quad \left(\mathcal{M}f^i, f^i\right)(t) = \left(\mathcal{M}f_0^i, f_0^i\right), \tag{3.3}$$

$$\int_{K^c} \left(1 + \frac{|v|^2}{2} + \phi_Q(x,v)\right) f^e(t,x,v)\, dxdv = \int_{K^c} \left(1 + \frac{|v|^2}{2} + \phi_Q\right) f_0^e\, dxdv, \tag{3.4}$$

$$\forall q \in [1,p], \qquad |f^e(t)|_{L^q(K^c)} = |f_0^e|_{L^q}, \tag{3.5}$$

$$\left(\mathcal{M}f^e, f^e\right)(t) = -2\int_0^t \int_K f^e(s)\, v \cdot E_{f^e \mathbf{1}_{K^c}}(s) ds. \tag{3.6}$$

Here we extended the operator $\mathcal{M}$ to $\mathcal{LE}$ by

$$\mathcal{M}f = \left((p-1)Q^{p-2} f + \phi_{f\mathbf{1}_K}\right) \mathbf{1}_K = \mathcal{M}(f\mathbf{1}_K). \tag{3.7}$$

Remark 2. From (3.3), $L^2(K, d\mu)$ is invariant under the flow (1.23) and the linearized energy-Casimir functional $(\mathcal{M}f, f)$ is conserved by the flow. This is, however, no longer true for a general initial data and the error to the conservation law is measured by (3.6).

Proof of Proposition 3.2.

Step 1. Transport by E_Q. Let $\mathcal{T}$ be the linear transport operator induced by the field of the ground state Q:

$$\mathcal{T} = v \cdot \nabla_x - E_Q \cdot \nabla_v. \tag{3.8}$$

Thanks to the regularity of the field $E_Q = \nabla_x \phi_Q$, one can define the characteristics associated with $\mathcal{T}$. For (t, x, v), we denote by $X(s; t, x, v)$, $V(s; t, x, v)$ the global solution of the differential system

$$\frac{d}{ds}X = V, \qquad \frac{d}{ds}V = -E_Q(X), \qquad X(s=t) = x, \qquad V(s=t) = v. \tag{3.9}$$

Recall that the energy is an invariant of this system, i.e., $\frac{1}{2}|V(s;t,x,v)|^2 + \phi_Q(X(s;t,x,v))$ is independent of s, and that for any s, t the Jacobian of the Lagrangian change of variable

$$(x, v) \mapsto (X(s; t, x, v), V(s; t, x, v)) \tag{3.10}$$

is equal to 1. An important consequence of the energy invariant and the fact that Q is a function of the microscopic energy is that a characteristic curve cannot cross the boundary of the support Q: K and K^c are both invariant along the flow (3.9).

Step 2. Well-posedness of (1.23) in L^1. It is a simple consequence of the existence of the characteristics curves (3.9) and we briefly sketch the proof for the sake of completeness. Let $T > 0$ and introduce the following mapping on $\mathcal{C}([0,T], L^1(\mathbb{R}^N \times \mathbb{R}^N))$: for f in this space, $G(f)$ is defined as the unique weak solution g of

$$\partial_t g + v \cdot \nabla_x g - E_Q \cdot \nabla_v g = E_f \cdot \nabla_v Q\,, \qquad g(t=0) = f_0, \tag{3.11}$$

given, thanks to the characteristics, by

$$\begin{aligned} G(f)(t,x,v) &= f_0(X(0;t,x,v), V(0;t,x,v)) \\ &\quad + \int_0^t (E_f \cdot \nabla_v Q)\,(s, X(s;t,x,v), V(s;t,x,v))\,ds. \end{aligned} \tag{3.12}$$

It is useful to note that $f \in \mathcal{C}([0,T], L^1(\mathbb{R}^N \times \mathbb{R}^N))$ implies $\rho_f \in \mathcal{C}([0,T], L^1(\mathbb{R}^N))$ and thus $E_f \in \mathcal{C}([0,T], L^{\frac{N}{N-1},\infty})$, where $L^{\frac{N}{N-1},\infty}$ stands for the weak $L^{\frac{N}{N-1}}$ space (or Marcinkiewicz space). Hence $E_f \in L^1_{loc}(\mathbb{R}^N)$. Observe from (2.40) that

$$\int_{\mathbb{R}^N} |\nabla_v Q(x,v)|\,dv \le C V_Q(x)$$

is bounded on $\mathbb{R}^N$ and compactly supported. Thus the right-hand side of (3.11) belongs to $\mathcal{C}([0,T], L^1(\mathbb{R}^N \times \mathbb{R}^N))$ and $\forall t \in [0,T]$,

$$|E_f \cdot \nabla_v Q|_{L^1}(t) \le C\,|f|_{L^1}(t).$$

Integrating (3.12) on $\mathbb{R}^N \times \mathbb{R}^N$ and performing the Lagrangian change of coordinate (3.10), we get for any $f_1, f_2 \in C([0,T], L^1(\mathbb{R}^N \times \mathbb{R}^N))$

$$|G(f_1) - G(f_2)(t)|_{L^1} \le C \int_0^t |f_1 - f_2|_{L^1}(s)\,ds.$$

This is enough to conclude by the Banach fixed point theorem for T small enough. We have proved that (1.23) admits a unique solution that satisfies

(3.13)

$$f(t,x,v)=f_0(X(0;t,x,v),V(0;t,x,v))+\int_0^t (E_f\cdot\nabla_v Q)\,(s,X(s;t,x,v),V(s;t,x,v))\,ds.$$

Step 3. Well-posedness in $\mathcal{LE}$. Now let $f=f^i+f^e\in\mathcal{LE}$. The simple remarks that K and K^c remain invariant under the flow of the characteristics (3.9) and that the source term $E_f\cdot\nabla_v Q$ in (1.23) is supported on K enable us to conclude that Supp $f^i\subset K$ and that $h:=f^e\,\mathbf{1}_{K^c}=f\,\mathbf{1}_{K^c}$ solves in the weak sense the equation

$$\partial_t h+v\cdot\nabla_x h-E_Q\cdot\nabla_v h=0\,,\qquad h(t=0)=f_0^e. \tag{3.14}$$

It is clear then that $f^e\,\mathbf{1}_{K^c}\geq 0$ a.e. on $\mathbb{R}_+\times\mathbb{R}^N\times\mathbb{R}^N$ and that (3.4) and (3.5) hold (recall that $\frac{|v|^2}{2}+\phi_Q$ is invariant along the characteristics). Note that the boundedness of ϕ_Q in L^∞, (3.4), and (3.5) with $q=1$ implies a uniform bound for $|f^e\,\mathbf{1}_{K^c}(t)|_{\mathcal{E}}$.

We now square (3.13), multiply by Q^{p-2}, and integrate over K to get

$$\begin{aligned}|f(t)|^2_{L^2(K,d\mu)}&\leq 2|f_0|^2_{L^2(K,d\mu)}\\&\quad+C(t)\textstyle\int_0^t\int_K\Big(Q^{p-2}\,|E_f\cdot\nabla_v Q|^2\Big)(s,X(s;t,x,v),V(s;t,x,v))\,dsdxdv,\end{aligned}$$

where we used the fact that $Q(x,v)=Q(X(s;t,x,v),V(s;t,x,v))$. Then, performing again the Lagrangian change of variable (3.10) and noting from (2.40) that

$$Q^{p-2}\,|\nabla_v Q|^2=\frac{|v|\,|\nabla_v Q|}{p-1}\in L^\infty(\mathbb{R}^N_x,L^1(\mathbb{R}^N_v)),$$

we obtain

$$\begin{aligned}|f(t)|^2_{L^2(K,d\mu)}\quad&\leq 2|f_0|^2_{L^2(K,d\mu)}+C\int_0^t|E_f|^2_{L^2(\mathbb{R}^N)}(s)\,ds\\&\leq 2|f_0|^2_{L^2(K,d\mu)}+C\int_0^t|f(t)|^2_{\mathcal{LE}}(s)\,ds,\end{aligned}$$

where we used (3.1). Since we already have $|f\,\mathbf{1}_{K^c}|_{\mathcal{E}}\leq C\ \forall\ t$, this is enough to conclude with the Gronwall lemma that the function f belongs to $L^\infty_{loc}(\mathbb{R}_+,\mathcal{LE})$.

Step 4. Derivation of the conservation laws. It remains to prove the conservation laws (3.3) and (3.6). To this aim, let us first define a suitable regularization of f. Let $n\in\mathbb{N}^*$ and let f_0^n be a sequence of $C_0^\infty(\mathbb{R}^N\times\mathbb{R}^N)\cap\mathcal{LE}$ functions which converges to f_0 in the $\mathcal{LE}$ topology as $n\to+\infty$. Consider now a nonnegative $C^\infty(\mathbb{R})$ function θ such that $\theta(u)=1$ for $u\geq 1$ and $\theta(u)=0$ for $u\leq 1/2$ and let

(3.15)

$$\theta^n(x,v)=\theta\left(n\left|1+\frac{|v|^2}{2}+\phi_Q(x)\right|\right),\qquad\chi^n(x,v)=\theta\left(n\left(-1-\frac{|v|^2}{2}-\phi_Q(x)\right)\right).$$

Now, we define f^n as the solution of the following problem, which can be constructed by a fixed point procedure similarly as above:

$$\partial_t f^n+v\cdot\nabla_x f^n-E_Q\cdot\nabla_v f^n=\theta^n\,E_{f^n}\cdot\nabla_v Q\,,\qquad f^n(t=0)=f_0^n. \tag{3.16}$$

Since the function $\theta^n\,\nabla_v Q$ is C^∞, it is readily seen that f^n is a sequence of C^∞ function that converges to f^n in the $\mathcal{LE}$ topology as $n\to+\infty$.

Now, from (3.16) and (2.40), we get

(3.17)
$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^{2N}} Q^{p-2}(f^n)^2 dxdv &= 2\int_K Q^{p-2} f^n\,\theta^n\,E_{f^n}\cdot\nabla_v Q\,dxdv\\ &= -\frac{2}{p-1}\int_K f^n\,\chi^n\,v\cdot E_{f^n}\,dxdv\\ &= -\frac{2}{p-1}\int_K f^n\chi^n\,v\cdot E_{f^n\chi^n}\,dxdv - \frac{2}{p-1}\int_K f^n\chi^n\,v\cdot E_{f^n(1-\chi^n)}\,dxdv\\ &= \frac{2}{p-1}\int_{\mathbb{R}^N}\phi_{f^n\chi^n}\nabla_x\cdot\left(\int_{\mathbb{R}^N} vf^n\,\chi^n\,dv\right)dx - \frac{2}{p-1}\int_K f^n\chi^n\,v\cdot E_{f^n(1-\chi^n)}\,dxdv,\end{aligned}$$

where we used the equation $v\cdot\nabla_x Q - E_Q\cdot\nabla_v Q = 0$ and noted that θ^n and χ^n coincide on K. Multiply now (3.16) by χ^n. Since χ^n is a function of the energy $\frac{|v|^2}{2}+\phi_Q(x)$, we have

$$\partial_t(f^n\chi^n) + v\cdot\nabla_x(f^n\chi^n) - E_Q\cdot\nabla_v(f^n\chi^n) = (\chi^n)^2\,E_{f^n}\cdot\nabla_v Q. \tag{3.18}$$

Besides, for the same reason, the function $(\chi^n)^2\,\nabla_v Q$ is an exact derivative with respect to v. Hence an integration of (3.18) with respect to v yields

$$\partial_t\int_{\mathbb{R}^N}(f^n\chi^n)\,dv + \nabla_x\cdot\left(\int_{\mathbb{R}^N} vf^n\,\chi^n\,dv\right) = 0$$

and, by the Poisson equation, the first integral in the right-hand side of (3.17) can be rewritten as follows:

$$\frac{2}{p-1}\int_{\mathbb{R}^N}\phi_{f^n\chi^n}\nabla_x\cdot\left(\int_{\mathbb{R}^N} vf^n\,\chi^n\,dv\right)dx = \frac{1}{p-1}\frac{d}{dt}\int_K \phi_{f^n\chi^n}\,f^n\chi^n\,dxdv.$$

Finally,

$$\begin{aligned}(\mathcal{M}(f^n\chi^n), f^n\chi^n)(t) &= (\mathcal{M}(f_0^n\chi^n), f_0^n\chi^n)\\ &\qquad -2\int_0^t\int_K f^n(s)\chi^n\,v\cdot E_{f^n(1-\chi^n)}(s)\,dxdvds.\end{aligned}$$

Since, $\forall\ T>0$, $f^n\to f$ in $L^\infty((0,T),\mathcal{L}\mathcal{E})$ as $n\to+\infty$, one can pass to the limit in the various terms of this identity, thanks to Lemma 3.1 (recall that v is bounded on K). Applying this inequality to f^i and f^e leads, respectively, to (3.3) and (3.6).

This concludes the proof of Proposition 3.2. ∎

Let us conclude this section with the following commutation formula that will be useful in the next subsections.

Lemma 3.3 (commutation formula). *Let $f_0\in\mathcal{L}\mathcal{E}$ and let $f(t)=e^{-t\mathcal{L}}f_0$ be the corresponding weak solution of* (1.23). *Let $h\in L^2(K,d\mu)\cap C^1(\mathring{K})$ be such that $\mathcal{L}h\in L^2(K,d\mu)$. Then*

$$(f(t),\mathcal{M}h) = (f_0,\mathcal{M}h) + \int_0^t (f(s),\mathcal{M}\mathcal{L}h)\,ds - \int_0^t\int_K v\cdot\left(hE_{f\mathbf{1}_{K^c}}\right)(s)ds. \tag{3.19}$$

Moreover, for $1\le i\le N$, we have

$$(f(t),x_i\,\mathbf{1}_K) = (f_0,x_i\,\mathbf{1}_K) + \int_0^t (f(s),v_i\,\mathbf{1}_K)\,ds. \tag{3.20}$$

Proof of Lemma 3.3. Let us observe the following algebraic identity, which follows from a direct computation: Let $\mathcal{L}$, $\mathcal{T}$, $\mathcal{M}$ be, respectively, given by (1.24), (3.8), and (1.21). Then

$$\forall f \in L^2(K, d\mu), \quad \mathcal{L} f = \frac{1}{(p-1)Q^{p-2}} \mathcal{T}(\mathcal{M} f) \tag{3.21}$$

in the sense of distributions.

Introduce then the smooth function θ^n defined by (3.15) and let $\widetilde{f_0^n}$ be a C_0^∞ regularization of f_0. Then, setting $f_0^n = \widetilde{f_0^n}\,\theta^n$, we define f^n as the classical solution of (3.16). One can see that $f^n \to f$ in $\mathcal{L}\mathcal{E}$ as $n \to +\infty$. Moreover, since the flow (3.9) preserves the energy and θ^n depends only on the energy, the support of f^n is included in the support of θ^n, where h is C^1. We split $f^n = f_i^n \mathbf{1}_K + f_e^n \mathbf{1}_{K^c} = f_i^n + f_e^n$ with $f_{i,e}^n$ smooth from the support localization of f^n, and thus

$$\frac{d}{dt}(f^n, \mathcal{M}h) = -(\mathcal{L} f^n, \mathcal{M}h) = -(\mathcal{L} f_i^n, \mathcal{M}h) - (\mathcal{L} f_e^n, \mathcal{M}h).$$

The first term is computed using (3.21), the self-adjointness of $\mathcal{M}$, the skew-adjointness of $\mathcal{T}$, and the fact that $\mathcal{T}(F(e)) = 0$ for any function F:

$$\begin{aligned} -(\mathcal{L} f_i^n, \mathcal{M}h) &= -\left(\frac{1}{(p-1)Q^{p-2}} \mathcal{T}(\mathcal{M} f_i^n), \mathcal{M}h\right) = \left(\mathcal{M} f_i^n, \mathcal{T}\left[\frac{1}{(p-1)Q^{p-2}} \mathcal{M}h\right]\right) \\ &= (\mathcal{M} f_i^n, \mathcal{L}h) = (f_i^n, \mathcal{M}\mathcal{L}h) = (f^n, \mathcal{M}\mathcal{L}h). \end{aligned}$$

For the second term, we use that f_e^n and $\mathcal{M}h$ have disjoint support to compute

$$-(\mathcal{L} f_e^n, \mathcal{M}h) = \int_{\mathbb{R}^N} E_{f_e^n} \cdot \nabla_v Q (p-1) Q^{p-2} h = -\int_K hv \cdot E_{f_e^n}$$

from (2.40). We then integrate in time and pass to the limit as $n \to +\infty$, and (3.19) follows. The second identity (3.20) can be proved by a similar regularization procedure and by direct calculations. This concludes the proof of Lemma 3.3. $\square$

Remark 3. If the support of f_0 is in K, then (3.19) becomes simpler:

$$(f(t), \mathcal{M}h) = (f_0, \mathcal{M}h) + \int_0^t (f(s), \mathcal{M}\mathcal{L}h)\, ds. \tag{3.22}$$

3.2. The linearized dynamics on the support of Q. From Proposition 3.2, the solution $e^{-t\mathcal{L}} f_0$ of (1.23) remains supported on K when it has this property at $t = 0$. In this section we estimate the action of the linearized Vlasov–Poisson system on $L^2(K, d\mu)$.

Let us start by introducing the following decomposition of $L^2(K, d\mu)$ whose proof is given in the appendix.

Lemma 3.4 (decomposition of $L^2(K, d\mu)$). *Let $(N, \alpha) \in \{(3,1), (3,2), (4,2)\}$. There holds the decomposition*

$$L^2(K, d\mu) = M \oplus S,$$

where M is defined as the set of $f \in L^2(K, d\mu)$ with

$$\begin{cases} \textit{for } N \neq \alpha + 2, & \left(f, \dfrac{|v|^\alpha}{\alpha} + \phi_Q\right) = (f, x_i) = (f, v_i) = 0, \quad 1 \le i \le N, \\ \textit{for } N = \alpha + 2, & \left(f, \dfrac{|v|^\alpha}{\alpha} + \phi_Q\right) = (f, x \cdot v|v|^{\alpha-2}) = (f, |x|^2 |v|^{2-\alpha}) \\ & \qquad = (f, x_i) = (f, v_i) = 0, \quad 1 \le i \le N, \end{cases}$$

and S is defined, thanks to the functions S_i given by (2.31), *according to*

$$\begin{cases} \textit{if } N \neq \alpha+2, \textit{ then } S = \operatorname{span}\{S_1\ , \partial_{x_i} Q\ ,\ \partial_{v_i} Q\ ,\ 1 \leq i \leq N\}\,, \\ \textit{if } N = \alpha+2, \textit{ then } S = \operatorname{span}\{S_1\ ,\ S_2\ ,\ S_3\ , \partial_{x_i} Q\ , \partial_{v_i} Q\ ,\ 1 \leq i \leq N\}\,. \end{cases}$$

Our main claim is now that S and M are invariant under the linearized flow (1.23). The subspace S is the so-called *flag space* and contains the algebraically growing modes induced by the large set of symmetries (1.7), while the free evolution remains bounded on M in the $\mathcal{LE}$ norm.

PROPOSITION 3.5 (splitting of the motion). *Let $(N,\alpha) \in \{(3,2),(4,2)\}$. Consider the decomposition $L^2(K,d\mu) = M \oplus S$, where the spaces M, S are defined in Lemma* 3.4. *Then M and S are both invariant under the linearized flow* (1.23) *and there holds: $\forall t \in \mathbb{R}_+$,*

$$\forall g_0 \in M, \quad \left|e^{-t\mathcal{L}} g_0\right|_{L^2(K,d\mu)} \leq C\,|g_0|_{L^2(K,d\mu)}\,, \tag{3.23}$$

$$\forall g_0 \in S, \quad \left|e^{-t\mathcal{L}} g_0\right|_{L^2(K,d\mu)} \leq \begin{cases} C\,(1+t)\,|g_0|_{L^2(K,d\mu)} & \textit{for} \quad N=3, \\ C\,(1+t^2)\,|g_0|_{L^2(K,d\mu)} & \textit{for} \quad N=4. \end{cases} \tag{3.24}$$

Proof of Proposition 3.5.

Step 1. The evolution on S. The free evolution is explicit on S. Indeed, letting $N = 3,4$ and $(S_i)_{1\leq i\leq 3}$ be defined by (2.31), we claim that

$$\mathcal{L}S_1 = 0 \qquad \mathcal{L}(\partial_{x_i} Q) = 0, \qquad \mathcal{L}(\partial_{v_i} Q) = -\partial_{x_i} Q, \tag{3.25}$$

and the extra relations for $N = 4$:

$$\mathcal{L}S_2 = S_1\,, \qquad \mathcal{L}S_3 = S_2. \tag{3.26}$$

Proof of (3.25) *and* (3.26). They follow from (2.32) and (3.21). Let us prove (3.26). In the interior of K, $(S_i)_{1\leq i\leq 3}$ are smooth. We then compute using (2.32): $\forall (x,v) \in \overset{\circ}{K}$,

$$\mathcal{L}S_2 = \frac{1}{(p-1)Q^{p-2}}\,\mathcal{T}(\mathcal{M}S_2) = \frac{1}{(p-1)Q^{p-2}}\mathcal{T}(x\cdot v) = \frac{1}{(p-1)Q^{p-2}}(|v|^2 - x\cdot E_Q).$$

We now take the derivative of (2.33) in x and v to get

$$E_Q + (p-1)Q^{p-2}\nabla_x Q = 0, \qquad v + (p-1)Q^{p-2}\nabla_v Q = 0,$$

and thus

$$\mathcal{L}S_2 = \frac{1}{(p-1)Q^{p-2}}(|v|^2 - x\cdot E_Q)\mathbf{1}_K = -v\cdot\nabla_v Q + x\cdot\nabla_x Q = S_1.$$

Similarly,

$$\begin{aligned} \mathcal{L}S_3 &= \frac{1}{(p-1)Q^{p-2}}\,\mathcal{T}(\mathcal{M}S_3) = \frac{1}{(p-1)Q^{p-2}}\,\mathcal{T}\left(\frac{|x|^2}{2}\right) \\ &= \frac{x\cdot v}{(p-1)Q^{p-2}} = -x\cdot\nabla_v Q = S_2, \end{aligned}$$

where we used (2.40). In order to conclude the proof of (3.26), we use the following technical remark. Let $h \in L^2(K, d\mu)$ such that $h \in C^1(\overset{\circ}{K})$ and denote by $(\mathcal{T}h)_K$ the function defined pointwise on $\overset{\circ}{K}$ and continued by zero outside K. Assume that $(\mathcal{T}h)_K \in L^1(K)$. Then, due to the fact that the boundary of K is a level set of the microscopic energy, the distribution $\mathcal{T}h$ defined as the distributional derivative of h by the derivation $\mathcal{T}$ and $(\mathcal{T}h)_K$ coincide in $\mathcal{D}'(\mathbb{R}^N \times \mathbb{R}^N)$. Applying this with $h = S_1$ or S_2 concludes the proof of (3.26). Next (3.25) follows similarly and is left to the reader.

Now let $g_0 \in S$, i.e., according to Lemma 3.4,

$$\text{if } N = 3, \qquad g_0 = \alpha\, S_1 + \sum_{i=1}^{N} \delta_i\, \partial_{x_i} Q + \sum_{i=1}^{N} \epsilon_i\, \partial_{v_i} Q,$$
$$\text{if } N = 4, \qquad g_0 = \alpha\, S_1 + \beta\, S_2 + \gamma\, S_3 + \sum_{i=1}^{N} \delta_i\, \partial_{x_i} Q + \sum_{i=1}^{N} \epsilon_i\, \partial_{v_i} Q.$$

From (3.25) and (3.26), the evolution $e^{-t\mathcal{L}} g_0$ is explicit:

$$\text{if } N = 3, \quad e^{-t\mathcal{L}} g_0 = \alpha\, S_1 + \sum_{i=1}^{N} (\epsilon_i\, t + \delta_i)\, \partial_{x_i} Q + \sum_{i=1}^{N} \epsilon_i\, \partial_{v_i} Q,$$
$$\text{if } N = 4, \quad e^{-t\mathcal{L}} g_0 = \left(\frac{\gamma}{2} t^2 - \beta t + \alpha\right) S_1 + (-\gamma t + \beta)\, S_2 + \gamma\, S_3$$
$$+ \sum_{i=1}^{N} (\epsilon_i\, t + \delta_i)\, \partial_{x_i} Q + \sum_{i=1}^{N} \epsilon_i\, \partial_{v_i} Q,$$

which shows the stability of S and (3.24) is proved.

Step 2. The evolution on M. Now let $g_0 \in M$. From (3.25) and (3.26), we have $\mathcal{M}\mathcal{L}S_1 = 0$ and for $N = 4$, $\mathcal{M}\mathcal{L}S_2 = \mathcal{M}S_1$, $\mathcal{M}\mathcal{L}S_3 = \mathcal{M}S_2$. Hence Lemma 3.3 and Remark 3 give

$$\frac{d}{dt}(e^{-t\mathcal{L}} g_0, \mathcal{M}S_1) = 0, \tag{3.27}$$

and, if $N = 4$,

$$\frac{d}{dt}(e^{-t\mathcal{L}} g_0, \mathcal{M}S_2) = (e^{-t\mathcal{L}} g_0, \mathcal{M}S_1), \qquad \frac{d}{dt}(e^{-t\mathcal{L}} g_0, \mathcal{M}S_3) = (e^{-t\mathcal{L}} g_0, \mathcal{M}S_2). \tag{3.28}$$

Furthermore, noting that $\mathcal{M}\partial_{v_i} Q = -v_i$, we deduce from (3.25), (3.22), (3.20), and (2.30) that for $1 \le i \le N$,

$$\frac{d}{dt}(e^{-t\mathcal{L}} g_0, v_i) = (e^{-t\mathcal{L}} g_0, \mathcal{M}\partial_{x_i} Q) = 0, \qquad \frac{d}{dt}(e^{-t\mathcal{L}} g_0, x_i) = (e^{-t\mathcal{L}} g_0, v_i). \tag{3.29}$$

Recalling from Lemma 3.4 and from (2.32) that M can be characterized as the set of $f \in L^2(K, d\mu)$ with

$$\begin{cases} \text{for } N = 3, \quad (f, \mathcal{M}S_1) = (f, x_i) = (f, v_i) = 0, \quad 1 \le i \le N, \\ \text{for } N = 4, \quad (f, \mathcal{M}S_j) = (f, x_i) = (f, v_i) = 0, \quad 1 \le j \le 3,\ 1 \le i \le N, \end{cases}$$

we infer from (3.27), (3.28), and (3.29) that $g_0 \in M$ implies $e^{-t\mathcal{L}} g_0 \in M\ \forall\ t \ge 0$. The uniform bound (3.23) on $e^{-t\mathcal{L}} g$ in $\mathcal{L}\mathcal{E}$ now follows from the conservation of the linearized energy-Casimir functional (3.3) and the coercivity property of Theorem 1.3 which ensures that the quadratic form $(\mathcal{M}\,\cdot, \cdot)$ is coercive on M. This concludes the proof of Proposition 3.5. $\square$

3.3. Proof of Theorem 1.4. We are now in position to conclude the proof of Theorem 1.4.

Let $f_0 = f_0^i + f_0^e \in \mathcal{L}\mathcal{E}$ according to the decomposition (3.2) and denote by, respectively, $f^i(t) = e^{-t\mathcal{L}} f_0^i$, $f^e(t) = e^{-t\mathcal{L}} f_0^e$ the corresponding solutions to (1.23). The evolution of f^i is already controlled, thanks to Proposition 3.5. It remains to study the evolution of f^e, which is not supported a priori in K^c and may spread onto the whole space.

From Proposition 3.2, we have already

$$|f^e\, \mathbf{1}_{K^c}|_{\mathcal{E}}\,(t) \le C\; |f_0^e|_{\mathcal{E}} = C\; |f_0^e|_{\mathcal{L}\mathcal{E}}.$$

In particular,

$$|E_{f^e \mathbf{1}_{K^c}}|_{L^2} \le C\; |f_0^e|_{\mathcal{L}\mathcal{E}}. \tag{3.30}$$

It remains to bound $f^e \mathbf{1}_K$. Using Lemma 3.3 with $h = S_1$, $h = S_2$, $h = S_3$, and $h = \partial_{v_i} Q$ (the S_i's are defined in (2.31)), we obtain successively

$$\left|\left(f^e, \left(\frac{|v|^2}{2} + \phi_Q\right) \mathbf{1}_K\right)\right|(t) \le C(1+t)\,|f_0^e|_{\mathcal{L}\mathcal{E}},$$

$$|(f^e, v_i \mathbf{1}_K)|\,(t) \le C(1+t)\,|f_0^e|_{\mathcal{L}\mathcal{E}},$$

$$|(f^e, x_i \mathbf{1}_K)|\,(t) \le C(1+t^2)\,|f_0^e|_{\mathcal{L}\mathcal{E}},$$

and, if $N = 4$,

$$|(f^e, x\cdot v \mathbf{1}_K)|\,(t) \le C(1+t^2)\,|f_0^e|_{\mathcal{L}\mathcal{E}},$$

$$\left|\left(f^e, |x|^2 \mathbf{1}_K\right)\right|(t) \le C(1+t^3)\,|f_0^e|_{\mathcal{L}\mathcal{E}},$$

where we applied Lemma 2.4 and also used (3.30) to bound the various terms $\int_K hv \cdot E_{\mathbf{1}_{K^c}}(s) ds$. Therefore, Theorem 1.3 implies that

$$|f^e(t)|^2_{L^2(K,d\mu)} \le C|(\mathcal{M} f^e, f^e)|(t) + C(1+t^{2\alpha})\,|f_0^e|^2_{\mathcal{L}\mathcal{E}}$$

with $\alpha = 2$ if $N = 3$ and $\alpha = 3$ if $N = 4$. Now, one deduces from (3.6) and (3.30) that

$$\begin{aligned} |f^e(t)|^2_{L^2(K,d\mu)} &\le C\,|(\mathcal{M} f^e, f^e)|\,(t) + C(1+t^{2\alpha})\,|f_0^e|^2_{\mathcal{L}\mathcal{E}} \\ &\le C(1+t^{2\alpha})\,|f_0^e|^2_{\mathcal{L}\mathcal{E}} + C\; |f_0^e|_{\mathcal{L}\mathcal{E}} \int_0^t |f^e(s)|_{L^2(K,d\mu)} ds, \end{aligned}$$

and (1.26) follows from a standard sublinear Gronwall lemma. This concludes the proof of Theorem 1.4. ∎

Appendix. This appendix is devoted to the proof of the orthogonal decomposition $L^2(K, d\mu) = M \oplus S$ of Lemma 3.4.

Proof of Lemma 3.4. Explicit computations using the identities

$$||v|^\alpha Q|_{L^1} = \frac{N-2}{2}\,|E_Q|_{L^2}^2 = \int_{\mathbb{R}^{2N}} x\cdot\nabla_x\phi_Q\, Q\,dxdv$$

lead to

$$\left(\frac{|v|^\alpha}{\alpha} + \phi_Q\,, S_1\right) = (N-\alpha-2)\frac{(2\alpha+(\alpha+2)N-N^2)}{4\alpha}\,|E_Q|_{L^2}^2\,. \tag{A.1}$$

Moreover, integrations by parts yield

$$\big(x_i\,,\,\partial_{x_j}Q\big) = \big(v_i\,,\,\partial_{v_j}Q\big) = -\delta_{ij}\int Q\,,\quad 1\le i\le N,\quad 1\le j\le N, \tag{A.2}$$

and the radial symmetry of Q implies that

$$0 = \left(\frac{|v|^\alpha}{\alpha} + \phi_Q\,, S_2\right) = \left(\frac{|v|^\alpha}{\alpha} + \phi_Q\,, \partial_{x_i}Q\right) = \left(\frac{|v|^\alpha}{\alpha} + \phi_Q\,, \partial_{v_i}Q\right),\quad 1\le i\le N, \tag{A.3}$$

$$(x_i\,, S_j) = 0,\quad 1\le i\le N,\quad 1\le j\le 3. \tag{A.4}$$

Let $N\neq\alpha+2$ and $f\in L^2(K,d\mu)$. We look for λ, $(\delta_i)_{1\le i\le N}$, $(\epsilon_i)_{1\le i\le N}$ such that

$$\tilde f = f - \lambda S_1 - \sum_{i=1}^N \delta_i\,\partial_{x_i}Q - \sum_{i=1}^N \epsilon_i\,\partial_{v_i}Q \in M.$$

Taking the inner product of $\tilde f$ with $\frac{|v|^\alpha}{\alpha} + \phi_Q$ and using (A.1) and (A.3), we get

$$\lambda = \left(f, \frac{|v|^\alpha}{\alpha} + \phi_Q\right)\left[(N-\alpha-2)\frac{(2\alpha+(\alpha+2)N-N^2)}{4\alpha}\,|E_Q|_{L^2}^2\right]^{-1}.$$

Taking the inner product of $\tilde f$ with x_i or v_i and using (A.4) and (A.2), we get

$$\delta_i = (f, x_i)\left[-\int Q\right]^{-1},\quad \epsilon_i = (f, v_i)\left[-\int Q\right]^{-1},\qquad 1\le i\le N.$$

Since λ, the δ_i's, and the ϵ_i's are uniquely defined in order to ensure $\tilde f\in M$, the result is proved.

Let $N=\alpha+2$. By explicit computations and using the symmetries of Q, we get

$$\alpha\left(\frac{|v|^\alpha}{\alpha} + \phi_Q\,, S_3\right) = (\mathcal{M}S_1, S_3) = (S_1, \mathcal{M}S_3) = \left(S_1, \frac{|v|^{2-\alpha}|x|^2}{2}\right) = -\,\frac{\alpha}{2}\int |v|^{2-\alpha}|x|^2 Q, \tag{A.5}$$

$$\big(|v|^{2-\alpha}|x|^2, S_2\big) = \big(|v|^{2-\alpha}|x|^2, \partial_{x_i}Q\big) = \big(|v|^{2-\alpha}|x|^2, \partial_{v_i}Q\big) = 0,\quad 1\le i\le N, \tag{A.6}$$

$$\big(x\cdot v|v|^{\alpha-2}, S_1\big) = \big(x\cdot v|v|^{\alpha-2}, \partial_{x_i}Q\big) = \big(x\cdot v|v|^{\alpha-2}, \partial_{v_i}Q\big) = 0,\quad 1\le i\le N, \tag{A.7}$$

$$\big(x \cdot v|v|^{\alpha-2}, S_3\big) = (\mathcal{M}S_2, S_3) = (S_2, \mathcal{M}S_3) = 0, \tag{A.8}$$

$$\big(x \cdot v|v|^{\alpha-2}, S_2\big) = \int |x|^2 Q + (\alpha-2)\int (x\cdot v)^2 |v|^{\alpha-4} Q. \tag{A.9}$$

Let $f \in L^2(K, d\mu)$. We deduce from (A.1)–(A.9) that

$$\tilde{f} = f - \lambda S_1 - \beta S_2 - \gamma S_3 - \sum_{i=1}^{N} \delta_i\, \partial_{x_i} Q - \sum_{i=1}^{N} \epsilon_i\, \partial_{v_i} Q \in M$$

if and only if

$$\gamma = \left(f, \frac{|v|^\alpha}{\alpha} + \phi_Q\right)\left[-\frac{1}{2}\int |v|^{2-\alpha}|x|^2 Q\right]^{-1},$$

$$\lambda = \left[\big(f, |v|^{2-\alpha}|x|^2\big) - \gamma\left(S_3, |v|^{2-\alpha}|x|^2\right)\right]\left[-\alpha\int |v|^{2-\alpha}|x|^2 Q\right]^{-1},$$

$$\beta = \big(f, x\cdot v|v|^{\alpha-2}\big)\left[\int |x|^2 Q + (\alpha-2)\int (x\cdot v)^2|v|^{\alpha-4} Q\right]^{-1},$$

$$\delta_i = (f, x_i)\left[-\int Q\right]^{-1}, \quad \epsilon_i = (f, v_i)\left[-\int Q\right]^{-1}, \quad 1 \le i \le N,$$

where we have successively taken the inner product of $\tilde{f}$ with $\frac{|v|^\alpha}{\alpha} + \phi_Q$, $|v|^{2-\alpha}|x|^2$, $x\cdot v|v|^{\alpha-2}$, x_i, and v_i. This concludes the proof of Lemma 3.4. ∎

Acknowledgment. The authors express their gratitude to the anonymous referee who pointed out a number of important references connected to this work.

REFERENCES

[1] A. Antonov, *Remarks on the problem of stability in stellar dynamics*, Soviet Astronomy AJ, 4 (1961), pp. 859–867.

[2] J. Batt, W. Faltenbacher, and E. Horst, *Stationary spherically symmetric models in stellar dynamics*, Arch. Ration. Mech. Anal. 93 (1986), pp. 159–183.

[3] J. Batt, P. J. Morrison, and G. Rein, *Linear stability of stationary solutions of the Vlasov–Poisson system in three dimensions*, Arch. Ration. Mech. Anal., 130 (1995), pp. 163–182.

[4] J. Binney and S. Tremaine, *Galactic Dynamics*, Princeton University Press, Princeton, NJ, 1987.

[5] J. Bourgain and W. Wang, *Construction of blowup solutions for the nonlinear Schrödinger equation with critical nonlinearity*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 25 (1997), (1998), pp. 197–215.

[6] T. Cazenave and P.-L. Lions, *Orbital stability of standing waves for some nonlinear Schrödinger equations*, Comm. Math. Phys., 85 (1982), pp. 549–561.

[7] R. J. DiPerna and P.-L. Lions, *Global weak solutions of kinetic equations*, Rend. Sem. Mat. Univ. Politec. Torino, 46 (1988), pp. 259–288.

[8] R. J. DiPerna and P.-L. Lions, *Solutions globales d'équations du type Vlasov–Poisson*, C. R. Acad. Sci. Paris Sér. I Math., 307 (1988), pp. 655–658.

[9] A. M. Fridmann and V. L. Polyachenko, *Physics of Gravitating Systems*, Springer-Verlag, New York, 1984.

[10] R. T. Glassey, *The Cauchy Problem in Kinetic Theory*, SIAM, Philadelphia, PA, 1996.
[11] R. T. Glassey and J. Schaeffer, *On symmetric solutions of the relativistic Vlasov–Poisson system*, Comm. Math. Phys., 101 (1985), pp. 459–473.
[12] R. T. Glassey and J. Schaeffer, *On global symmetric solutions to the relativistic Vlasov–Poisson equation in three space dimensions*, Math. Methods Appl. Sci., 24 (2001), pp. 143–157.
[13] Y. Guo, *Variational method for stable polytropic galaxies*, Arch. Ration. Mech. Anal., 130 (1995), pp. 163–182.
[14] Y. Guo and G. Rein, *Stable steady states in stellar dynamics*, Arch. Ration. Mech. Anal., 147 (1999), pp. 225–243.
[15] Y. Guo and G. Rein, *Existence and stability of Camm type steady states in galactic dynamics*, Indiana Univ. Math. J., 48 (1999), pp. 1237–1255.
[16] Y. Guo and G. Rein, *Isotropic steady states in galactic dynamics*, Comm. Math. Phys., 219 (2001), pp. 607–629.
[17] Y. Guo and G. Rein, *A non-variational approach to nonlinear stability in stellar dynamics applied to the King model*, Comm. Math. Phys., 271 (2007), pp. 489–509.
[18] M. Hadzic, *A constraint variational problem arising in stellar dynamics*, Quart. Appl. Math., 65 (2007), pp. 145–153.
[19] E. Horst and R. Hunze, *Weak solutions of the initial value problem for the unmodified nonlinear Vlasov equation*, Math. Methods Appl. Sci., 6 (1984), pp. 262–279.
[20] H. E. Kandrup and J. F. Sygnet, *A simple proof of dynamical stability for a class of spherical clusters*, Astrophys. J., 298 (1985), pp. 27–33.
[21] M. K. Kwong, *Uniqueness of positive solutions of* $\Delta u - u + u^p = 0$ *in* R^n, Arch. Ration. Mech. Anal., 105 (1989), pp. 243–266.
[22] M. Lemou, F. Méhats, and P. Raphaël, *On the orbital stability of the ground states and the singularity formation for the gravitational Vlasov–Poisson system*, Arch. Ration. Mech. Anal., to appear.
[23] M. Lemou, F. Méhats, and P. Raphaël, *Uniqueness of the critical mass blow up solution for the four dimensional gravitational Vlasov–Poisson system*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 24 (2007), pp. 825–833.
[24] M. Lemou, F. Méhats, and P. Raphaël, *Stable self-similar blowup dynamics for the relativistic gravitational Vlasov–Poisson system*, J. Amer. Math. Soc., to appear.
[25] E. H. Lieb and H. T. Yau, *The Chandrasekhar theory of stellar collapse as the limit of quantum mechanics*, Comm. Math. Phys., 112 (1987), pp. 147–174.
[26] P.-L. Lions, *The concentration-compactness principle in the calculus of variations. The locally compact case.* I, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), pp. 109–145.
[27] P.-L. Lions, *The concentration-compactness principle in the calculus of variations. The locally compact case.* II, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), pp. 223–283.
[28] P.-L. Lions and B. Perthame, *Propagation of moments and regularity for the* 3*-dimensional Vlasov–Poisson system*, Invent. Math., 105 (1991), pp. 415–430.
[29] M. Maris, *Existence of nonstationary bubbles in higher dimensions*, J. Math. Pures Appl. (9), 81 (2002), pp. 1207–1239.
[30] Y. Martel, F. Merle, and T. P. Tsai, *Stability and asymptotic stability in the energy space of the sum of* N *solitons for subcritical gKdV equations*, Comm. Math. Phys., 231 (2002), pp. 347–373.
[31] F. Merle and P. Raphaël, *On a sharp lower bound on the blow-up rate for the* L^2 *critical nonlinear Schrödinger equation*, J. Amer. Math. Soc., 19 (2006), pp. 37–90.
[32] J. Perez and J. J. Aly, *Stability of spherical stellar systems*-I. *Analytical results*, Monthly Not. Royal. Astronomical Soc., 280 (1996), pp. 689–699.
[33] K. Pfaffelmoser, *Global classical solutions of the Vlasov–Poisson system in three dimensions for general initial data*, J. Differential Equations, 95 (1992), pp. 281–303.
[34] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Vol. 4, Academic Press, New York, 1979.
[35] G. Rein, *Flat steady states in stellar dynamics—Existence and stability*, Comm. Math. Phys., 205 (1999), pp. 229–247.
[36] G. Rein and M. Hadzic, *Global existence and nonlinear stability for the relativistic Vlasov–Poisson system in the gravitational case*, Indiana Univ. Math. J., to appear.
[37] Ó. Sánchez and J. Soler, *Orbital stability for polytropic galaxies*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 23 (2006), pp. 781–802.
[38] J. Schaeffer, *Global existence of smooth solutions to the Vlasov–Poisson system in three dimensions*, Comm. Partial Differential Equations, 16 (1991), pp. 1313–1335.

[39] J. Schaeffer, *Steady states in galactic dynamics*, Arch. Ration. Mech. Anal., 172 (2004), pp. 1–19.

[40] J. F. Sygnet, G. Des Forets, M. Lachieze-Rey, and R. Pellat, *Stability of gravitational systems and gravothermal catastrophe in astrophysics*, Astrophys. J., 276 (1984), pp. 737–745.

[41] N. G. Van Kampen and B. V. Felderhof, *Theoretical Methods in Plasma Physics*, North–Holland, Amsterdam, 1967.

[42] Y.-H. Wan, *On nonlinear stability of isotropic models in stellar dynamics*, Arch. Ration. Mech. Anal., 147 (1999), pp. 245–268.

[43] M. I. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal., 16 (1985), pp. 472–491.

[44] M. I. Weinstein, *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–67.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 1740–1763

HOMOGENIZATION OF DEGENERATE TWO-PHASE FLOW EQUATIONS WITH OIL TRAPPING*

BEN SCHWEIZER†

Abstract. We consider the one-dimensional degenerate two-phase flow equations as a model for water drive in oil recovery. The effect of oil trapping is observed in strongly heterogeneous materials with large variations in the permeabilities and in the capillary pressure curves. In such materials, a vanishing oil saturation may appear at interior interfaces and inhibit the oil recovery. We introduce a free boundary problem that separates a critical region with locally vanishing permeabilities from a strictly parabolic region and we give a rigorous derivation of the effective conservation law.

Key words. degenerate parabolic equation, effective equations, free boundary problems, two-phase flow

AMS subject classifications. 35B27, 74Q10, 35K65

DOI. 10.1137/060675472

1. Introduction. The equations of two-phase flow describe the motion of two immiscible fluids in a porous medium, e.g., the flow of oil and water in rock. The equations are a challenging subject of modern analysis, in particular due to the nonlinear and degenerate coefficient functions, the permeabilities of the two phases, and the capillary pressure.

Oil trapping is an effect in media with large variations in the coefficients. Well known to experimentalists [13, 14], the mathematical analysis of this effect was initiated in [5, 10, 9]. Let us consider the process of oil recovery from a medium that consists of a mixture of fine and coarse materials. Starting with a high oil saturation u and a high oil pressure p, after some time, the oil pressure falls below the entry pressure of oil in the fine material. From this point on, despite a positive saturation in the coarse material, oil can be trapped in regions that are surrounded by fine material.

In this work we analyze a one-dimensional medium that consists of two materials, distributed periodically with period $\varepsilon > 0$ and with different permeabilities k and different capillary pressures p_c. We denote the saturation function of the corresponding solutions by u^ε. Our aim is to find a macroscopic or effective equation, i.e., an equation that characterizes weak limits u^0 of the family u^ε for $\varepsilon \to 0$. An effective equation allows us to determine, e.g., in a numerical scheme, the averaged profile of the solution u^ε without resolving the scale $\varepsilon > 0$.

With the method of two-scale convergence developed in [1] and measure-theoretic tools from [3] we rigorously derive the macroscopic equation

$$\partial_t u^0 + \operatorname{div} \mathcal{F}(u^0, \partial_x u^0) = 0$$

with a nonlinear function $\mathcal{F}$ that is determined by the coefficient functions through a finite-dimensional nonlinear problem. The effective flux function reflects the effect of oil trapping: it satisfies $\mathcal{F}(u, v) = 0$ for all $v \in \mathbb{R}$ and all $u \leq u^*/2$, where u^* is the residual oil saturation in the coarse material. Our contribution continues the

*Received by the editors November 20, 2006; accepted for publication (in revised form) July 26, 2007; published electronically February 15, 2008.
http://www.siam.org/journals/sima/39-6/67547.html

†Fakultät für Mathematik, TU Dortmund, Vogelpothsweg 87, D-44227 Dortmund, Germany (Ben.Schweizer@tu-dortmund.de).

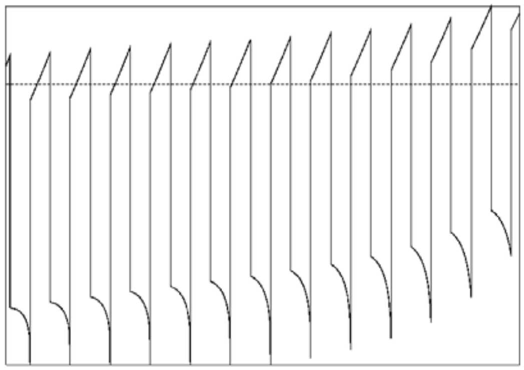


FIG. 1. *Oil saturation u^ε, zoomed view. Due to different capillary pressure curves in fine and coarse material, the saturation has jumps and oil is trapped in the coarse material. The saturation vanishes at some points. At these points, the permeability degenerates, but the infinite slope $\partial_x u^\varepsilon = -\infty$ makes transport of oil still possible. This figure is an unpublished numerical result and courtesy of I. S. Pop.*

analysis of [10, 9], where effective equations were formally derived with an asymptotic expansion. In [10], which contains various scalings and treats our scaling as the "capillary limit," the authors mention the specific difficulties in the homogenization of nonlinear equations and do not attempt a rigorous derivation of the effective equations. In fact, in [9] a different law is derived by starting with another ansatz; our analysis recovers the nonlinear function of [9]. For rigorous homogenization results of nonlinear equations we refer the reader to [7] for a double porosity model, to [6] for a stochastic setting, and to [4, 17] for models of capillary hysteresis. All these results concern the case of nondegenerate coefficients.

We want to highlight two difficulties in the homogenization process. The first concerns the nonlinear structure of the equations: loosely speaking, fluxes are of the form $g(u^\varepsilon)\partial_x u^\varepsilon$. In order to pass to the two-scale limit in such a term, we need a strong convergence of the argument in the nonlinear function. The strong convergence is usually obtained from estimates for first derivatives. This procedure cannot be performed in our case, since u^ε is an oscillatory function with jumps, and certainly not strongly convergent. The key point in the derivation of macroscopic equations in Proposition 1 is the compactness result of (3.14).

The second difficulty regards the degeneracy of the permeabilities. A strictly positive permeability k results in L^2-estimates for spatial derivatives and allows us to use the compactness result. But the effect of oil trapping appears precisely in the case that, in parts of the domain, the saturation vanishes; see Figure 1. In this situation, a vanishing permeability appears and no estimate for gradients is available. Our analysis uses the technique of a free boundary description in order to proceed. We decompose the domain into a "good" region G of strictly positive saturation and a "bad" region B; see Figure 2. We then derive the effective equations separately: in region G we use two-scale convergence (Proposition 1) to find the effective equations in Corollary 1. In region B, instead, the limit equations are trivial and are derived in the form of two-sided a priori estimates. The main point is then the continuity condition across the free boundary shown in Proposition 2. This condition allows us to combine the equations again into a single equation on the whole domain. The method exploits that oscillations of the free boundary do not appear; this is ensured by the boundary and initial conditions which imply a monotonicity of the free boundary.

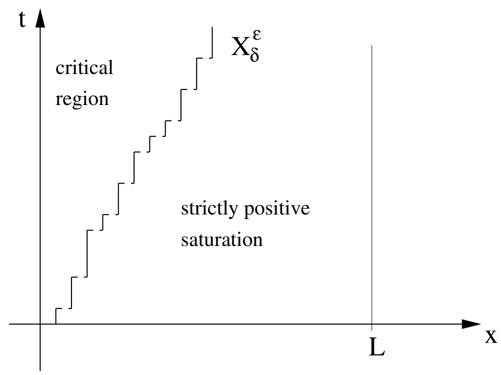


FIG. 2. *A free boundary separating the critical region from a region of uniformly positive saturation and permeability. The graph illustrates that the free boundary is discontinuous for $\varepsilon > 0$.*

As a by-product of our description, we learn more about qualitative features of solutions. We may define an experimentally observable free boundary $X_0^\varepsilon : (0,T) \to \mathbb{R}$ as the smallest function such that in all points (x,t) with $x > X_0^\varepsilon(t)$ the saturation u^ε is strictly positive. We now ask about properties of the limiting function $\tilde{X}_0^0(t) := \lim_{\varepsilon\to 0} X_0^\varepsilon(t)$. We prove in Lemma 6 that the corresponding limit curve $\{(x,t) : x = \tilde{X}_0^0(t)\}$ is contained in the critical domain $B \subset \bar{\Omega}_T$. Thus, the effective solution provides bounds for the experimentally observable free boundary. In particular, if it can be shown that the limit equation allows only solutions u^0 with $u^0 > u^*/2$ on Ω_T, then the experimentally observable free boundary must vanish in the limit $\varepsilon \to 0$.

Oil trapping in one-dimensional domains. We denote pressure and saturation of the oil phase by $p = p_1 = p_{oil}$ and $u = u_{oil}$ and the corresponding quantities of the water phase by $p_2 = p_{water}$ and $u_{water} = 1 - u$. The absolute permeability is denoted by k and the relative permeabilities by $k_{r,1} = k_{rel,oil}$ and $k_{r,2} = k_{rel,water}$. The equations in primary variables are the conservation laws for oil and water combined with the Darcy law for the velocities and the capillary pressure relation:

$$\partial_t u = \nabla \cdot (k(x)k_{r,1}(u)\nabla p_1), \tag{1.1}$$

$$-\partial_t u = \nabla \cdot (k(x)k_{r,2}(u)\nabla p_2), \tag{1.2}$$

$$p_1 - p_2 = p_c(u). \tag{1.3}$$

Summing the conservation laws and inserting the relation between the pressure functions yields, with $K(x,u) = k(x)(k_{r,1}(u) + k_{r,2}(u))$, and writing now p instead of p_1,

$$\nabla \cdot (K(x,u)\nabla p - k(x)k_{r,2}(u)\nabla[p_c(u)]) = 0. \tag{1.4}$$

One may regard this as an elliptic equation for p that defines the relation between p and u. Together with this relation, at least formally, (1.1) is an evolution equation for u.

In this work we study only the one-dimensional case with spatial domain $x \in \Omega = (0,L)$. Equation (1.4) then implies that the expression in parentheses is constant in space. Physically, the constant describes the total flux and we write

$$K(x,u)\partial_x p - k(x)k_{r,2}(u)\partial_x[p_c(x,u)] = -q_0. \tag{1.5}$$

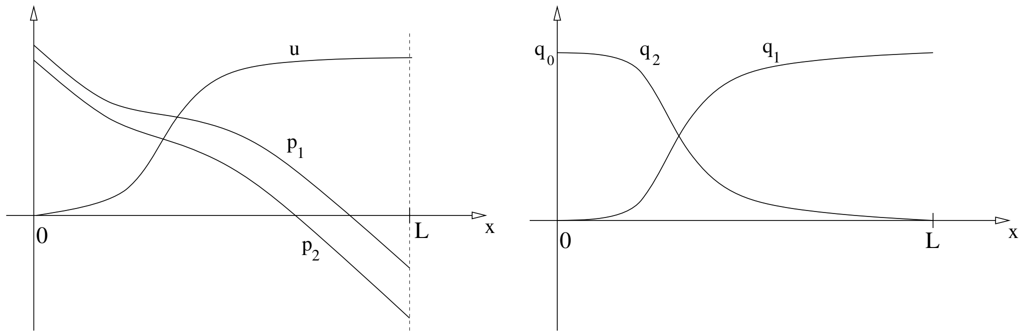


FIG. 3. *Typical solutions for homogeneous materials. Left: The typical shape of oil saturation u, oil pressure p_1, and water pressure p_2. Right: Shape of oil flux $q_1 = -k_1(x,u)\partial_x p_1$ and water flux $q_2 = -k_2(x,u)\partial_x p_2$ for the chosen boundary conditions. The curves illustrate the shape for the standard equations with spatially homogeneous coefficient functions. The solutions of the effective equations in the oil trapping problem look similar, but they exhibit a residual oil saturation $u \geq u^*/2$.*

In order to describe imbibition from the left, we assume $q_0 > 0$. The value of q_0 is given to us by the boundary conditions.[1] In order to find a single evolution equation, we solve (1.5) for $\partial_x p$. With the shorthand $k_i(x,u) = k(x)k_{r,i}(u)$ we find

$$\partial_x p = -q_0 \frac{1}{K(x,u)} + \frac{k_2(x,u)}{K(x,u)} \partial_x [p_c(x,u)]. \tag{1.6}$$

Inserting into (1.1) yields

$$\partial_t u = -\partial_x \left(f(u) - k(x)\lambda(u)\partial_x [p_c(x,u)] \right) \tag{1.7}$$

with

$$f(u) := q_0 \frac{k_{r,1}(u)}{k_{r,1}(u) + k_{r,2}(u)}, \qquad \lambda(u) := \frac{k_{r,1}(u)k_{r,2}(u)}{k_{r,1}(u) + k_{r,2}(u)}.$$

Equation (1.7) is an evolution equation of the form $\partial_t u + \partial_x F = 0$, where F is given by

$$F(x,u) := f(u) - k(x)\lambda(u)\partial_x [p_c(x,u)].$$

The qualitative shape of solutions is shown in Figure 3. We emphasize that the coefficient functions are degenerate,

$$k_{r,1}(s) \to 0, \ f(s) \to 0, \ \lambda(s) \to 0 \ \text{ for } s \to 0.$$

Less critical in this context is an additional degeneracy $\partial_s p_c(s) \to 0$ for $s \to 0$. Regarding high oil saturation we have $k_{r,2}(s) \to 0$ and $\lambda(s) \to 0$ for $s \to 1$. Our interest here is in the degeneracies for $s \to 0$, and we consider a physical situation where the saturation remains bounded away from 1 for all times.

[1]One choice of the boundary conditions is the following. At the inlet, $x = 0$, pure water enters the medium at a given rate; hence we have $u = 0$ at the left boundary and $q_0 > 0$ given. At the right boundary, $x = L$, only the nonwetting fluid oil can exit; hence $k_2(L, u(L))\partial_x p_2(L) = 0$ or, equivalently, $-k_1(L, u(L))\partial_x p_1(L) = q_0$. Notationally simpler is to impose a fixed saturation at the right boundary; we will therefore work with $u(0) = 0$ and $p_c(u(L)) = p_{max}$ in what follows.

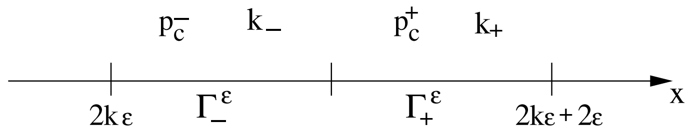


FIG. 4. *Interfaces inside the material.* Γ_-^ε *is a region with a fine material and* Γ_+^ε *is a region with a coarse material. The permeabilities satisfy* $k_+ > k_-$*, the capillary pressure curves* ${p_c}^+(s) < {p_c}^-(s)$ *for all s.*

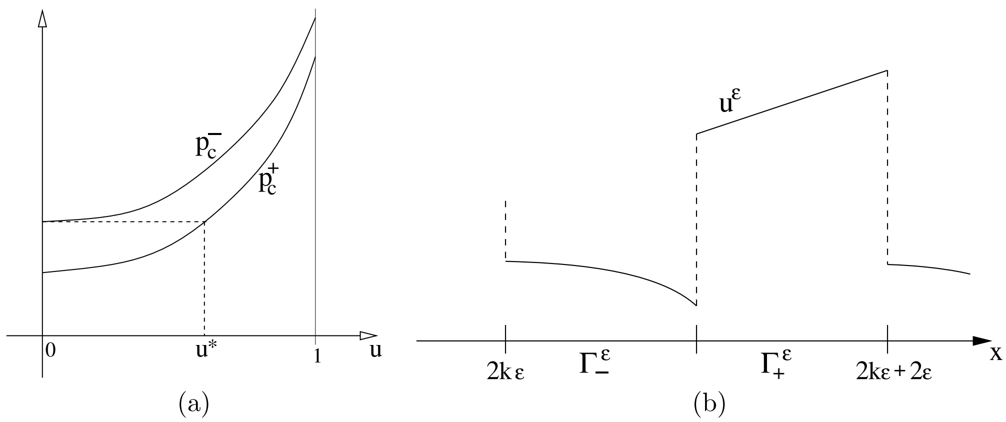


FIG. 5. (a) *The graphs* ${p_c}^+$ *and* ${p_c}^-$*.* (b) *Typical shape of a solution* $u^\varepsilon(.,t)$ *in a small interval* $(2k\varepsilon, 2k\varepsilon + 2\varepsilon)$*. The pressures* p_i *and the fluxes* q_i *are almost constant, and the capillary pressure* ${p_c}^\varepsilon(x, u^\varepsilon(x,t))$ *is continuous in x; hence the saturation* u^ε *jumps from high values on* Γ_+^ε *to low values on* Γ_-^ε*.*

Oscillatory coefficient functions. In this work we are interested in oscillatory coefficients $k_i^\varepsilon = k_i^0(x/\varepsilon, u)$ and ${p_c}^\varepsilon = {p_c}^0(x/\varepsilon, u)$. To simplify, we consider oscillations between two different coefficient functions. We distinguish the subdomains $\Gamma_-^\varepsilon := \varepsilon(2\mathbb{Z} + (0,1)) \cap (0,L)$ and $\Gamma_+^\varepsilon := \varepsilon(2\mathbb{Z} + (1,2)) \cap (0,L)$. For later use we additionally introduce $\Gamma^\varepsilon := \Gamma_+^\varepsilon \cup \Gamma_-^\varepsilon$ for the spatial domain without the interfaces.

We study coefficients as sketched in Figure 4:

$$k^\varepsilon(x) = \begin{cases} k_+ & \text{for } x \in \Gamma_+^\varepsilon, \\ k_- & \text{for } x \in \Gamma_-^\varepsilon, \end{cases} \qquad {p_c}^\varepsilon(x,u) = \begin{cases} {p_c}^+(u) & \text{for } x \in \Gamma_+^\varepsilon, \\ {p_c}^-(u) & \text{for } x \in \Gamma_-^\varepsilon. \end{cases} \tag{1.8}$$

A typical shape of ${p_c}^\pm$ is indicated in Figure 5(a); the corresponding local behavior of solutions is shown in Figure 5(b). The minimal pressure in Γ_-^ε with a positive saturation is $p_{min}^- = \lim_{u \searrow 0} {p_c}^-(u)$. Of importance is the residual oil saturation u^* in the coarse material, i.e., in Γ_+^ε. It is defined by the relation ${p_c}^+(u^*) = p_{min}^-$.

Our aim is to study solutions of (1.7) for this choice of coefficients. Understanding the equations in the distributional sense means to demand at the interfaces $\xi \in \varepsilon\mathbb{Z}$ the continuity of flux and capillary pressure. Since the capillary pressure curves are multivalued in general, we demand for all $\xi \in \varepsilon\mathbb{Z}$

$$F(\xi - 0, u(\xi - 0)) = F(\xi + 0, u(\xi + 0)), \tag{1.9}$$

$$p_c(\xi - 0, u(\xi - 0)) \cap p_c(\xi + 0, u(\xi + 0)) \neq \emptyset. \tag{1.10}$$

Here, we use the notation $h(x \pm 0)$ for $\lim_{\pm\delta \searrow 0} h(x + \delta)$, or, if $h \in H^1$, for the corresponding trace. Relation (1.10) is a compact way to write the standard interface condition for the capillary pressure which was derived in [5]. We use set-valued capillary pressure functions that assign to the saturations $s = 0$ and $s = 1$ an interval, e.g., ${p_c}^-(\xi, 0) := (-\infty, p^-_{min}]$ for $\xi \in \Gamma^\varepsilon_-$. In this way, if the saturation vanishes at one side of the interface, the pressure at the other side must be below the entry pressure p_{min}, but the exact value is not determined. The classical description of (1.10) is that we necessarily are in one of the following situations: (a) at both sides, the saturation is strictly between 0 and 1, and the capillary pressures on both sides coincide; (b) we have $s = 0$ at side A, $s \in (0, 1)$ at side B, and there holds $p_c \le p_{min}$, where p_c is evaluated at side B and p_{min} at side A; (c) we have $s = 0$ at both sides; (d) an analogous case with $s = 1$ on one side.

In the next step we write the equations in a compact and symmetric form. The conservation law (1.7) is recovered in (1.11) with $g^\pm(u) := k_\pm\, \lambda(u)\, \partial_u {p_c}^\pm(u)$ and $f^\pm(u) := f(u)$.

Mathematical description and main result. We assume that the coefficients are x-independent on each set $\Gamma^\varepsilon_\pm$,

$$g^\varepsilon(x,u) := \begin{cases} g^+(u) & \text{for } x \in \Gamma^\varepsilon_+, \\ g^-(u) & \text{for } x \in \Gamma^\varepsilon_-, \end{cases} \qquad f^\varepsilon(x,u) := \begin{cases} f^+(u) & \text{for } x \in \Gamma^\varepsilon_+, \\ f^-(u) & \text{for } x \in \Gamma^\varepsilon_-. \end{cases}$$

We study the conservation law

$$\begin{aligned} &\partial_t u^\varepsilon + \partial_x F^\varepsilon = 0 \text{ on } \Gamma^\varepsilon, \\ &F^\varepsilon = f^\varepsilon(x, u^\varepsilon) - g^\varepsilon(x, u^\varepsilon)\partial_x u^\varepsilon, \\ &F^\varepsilon \text{ and } {p_c}^\varepsilon(x, u^\varepsilon) \text{ are continuous in } \mathbb{Z}\varepsilon. \end{aligned} \tag{1.11}$$

Here, the continuity is understood in the classical sense for F^ε, and in the sense of (1.10) for ${p_c}^\varepsilon$. From now on, we study solution sequences u^ε to this equation, complemented with the initial condition ${p_c}^\varepsilon(u^\varepsilon)|_{t=0} = p_{max}$ on $(0, L)$ for some initial pressure value $p_{max} \in ({p_c}^-(0), {p_c}^+(1))$. As boundary conditions we impose $u^\varepsilon(0, t) = 0$ and ${p_c}^\varepsilon(u^\varepsilon(L, t)) = p_{max}$ for all $t \in (0, T)$. Throughout we assume the following monotonicity and regularity of the coefficient functions.

Assumptions. On flux and diffusivity we assume $0 \le f^\pm \in C^0([0, 1], \mathbb{R})$, $0 \le g^\pm \in C^{0,1}([0, 1], \mathbb{R})$, $f^-(0) = g^-(0) = 0$. Furthermore, $f^\pm \le c g^\pm$ on the interval $[0, ({p_c}^+)^{-1}(p_{max})]$ for some constant $c > 0$, and $f^\pm > 0$ on $(0, 1)$. On the capillary pressure we assume ${p_c}^+ \le {p_c}^-$ with strictly monotone functions ${p_c}^\pm \in C^1([0, 1], \mathbb{R})$.

Our main theorem is the rigorous derivation of effective equations. They are characterized by a nonlinear flux function $\mathcal{F} : [0, 1] \times \mathbb{R} \to \mathbb{R}$, which is constructed in (3.1)–(3.7). For the particular choice of coefficients considered there, our flux function $\mathcal{F}$ coincides with that of [9].

THEOREM 1. *Let $(u^\varepsilon, F^\varepsilon)$ be a family of entropy solutions to* (1.11) *on $\Omega_T = (0, L)\times(0, T)$, as in Definition* 1 *of section* 2*, satisfying the above boundary conditions. Then, for a subsequence $\varepsilon \to 0$ and for appropriate limiting functions, we find*

$$u^\varepsilon \rightharpoonup u^0 \text{ in } L^\infty(\Omega_T) \text{ weak-}\star, \quad F^\varepsilon \rightharpoonup F^0 \text{ in } L^2(\Omega_T) \text{ weakly.}$$

The limits satisfy the conservation law

$$\partial_t u^0 + \partial_x F^0 = 0 \tag{1.12}$$

in the distributional sense on Ω_T. The limit $u^0 \in L^\infty(\Omega_T)$ has the distributional derivative $\partial_x u^0 \in L^1(\Omega_T)$. The flux satisfies the relation

$$F^0 = \mathcal{F}(u^0, \partial_x u^0) \text{ almost everywhere in } \Omega_T \tag{1.13}$$

with the nonlinear function $\mathcal{F}(u, \partial_x u) = \mathcal{F}_0(u) - \mathcal{D}(u)\partial_x u$ defined in (3.1)–(3.7).

The remaining sections are devoted to the proof of Theorem 1 and are organized as follows. In section 2 we construct entropy solutions and derive comparison and monotonicity results. Section 3 is devoted to a two-scale homogenization result for regions with a strictly positive saturation. The homogenization in the general situation with degenerate solutions is performed in section 4. We introduce a description with a free boundary, derive the effective equations in the critical region, and determine the continuity condition across the free boundary.

The analysis of the limit problem (1.12), (1.13) is not the aim of this contribution. Interesting questions concern the existence and uniqueness of solutions to this degenerate problem, and the position of the free boundary $X(t) := \sup\{x : u^0(x,t) = u^*/2\}$. We note that the results of [2] cannot be applied to the equations in the above form, since, for a degenerate function $\mathcal{D}(u)$, the ellipticity assumption (Assumption 1.1(3)) of [2] is not satisfied. But the special structure $\mathcal{F}(u, \partial_x u) = \mathcal{F}_0(u) - \mathcal{D}(u)\partial_x u$ allows us to introduce formally a new variable U with $\partial_x U = \mathcal{D}(u)\partial_x u$ such that with $u = b(U)$ and $a(u, \partial_x U) = -\mathcal{F}_0(u) + \partial_x U$, the results of [2] and [15] may be applicable. The appearance of free boundaries is well known in porous-media-type equations; we refer the reader to [11] for results on the one-dimensional degenerate Cauchy problem $\partial_t u = \partial_x^2[a(u)] + \partial_x[b(u)]$ regarding existence, uniqueness, regularity, and speed of propagation of the free boundary.

Notation. The value of the constant C in estimates may change from one line to the next. For a set Q the function $\mathbf{1}_Q$ denotes the characteristic function $\mathbf{1}_Q(x) = 1$ for $x \in Q$ and $\mathbf{1}_Q(x) = 0$ for $x \notin Q$.

2. Entropy solutions and monotonicity.

2.1. Entropy solutions and regularity. In this section we sketch a solution concept that allows us to derive comparison principles for solutions. For other existence and uniqueness results we refer the reader to [8] and [12], where methods of [2] are extended to two-phase flow. An existence proof that uses a smoothing of the jump condition is performed in [5]. We refer the reader to [16] for a discussion of approximation schemes to degenerate equations that are also used below.

We always assume $p_{max} \in ({p_c}^-(0), {p_c}^+(1))$ and consider only boundary conditions as described above. We use the notion of a *family of regularized equations*: We assume that, for a sequence $\eta \searrow 0$, we have coefficient functions $g_\eta^\pm \in C^1([0,1], \mathbb{R})$, $f_\eta^\pm = f^\pm$, and ${p_c}^{\pm,\eta} \in C^1([0,1], \mathbb{R})$ that are strictly monotone and satisfy the same inequalities as the original coefficients. The equations are regularized in the sense that $g_\eta^\pm \geq \eta$, ${p_c}^{\pm,\eta}$ is single valued, and ${p_c}^{-,\eta}(0) = {p_c}^{+,\eta}(0)$. They approximate the original equation in the sense that ${p_c}^{+,\eta} = {p_c}^+$ on $(0,1)$, $g_\eta^\pm \to g^\pm$ uniformly on $[0,1]$, and ${p_c}^{-,\eta} \to {p_c}^-$ uniformly on compact subsets of $(0,1]$. By an appropriate choice of the regularization we can additionally achieve that the family $g_\eta^\pm \, \partial_u {p_c}^{\pm,\eta}$ is uniformly bounded on $[0,1]$.

In the following definition we use the primitive $G^\varepsilon(x,u)$ of g^ε, i.e., the function with $\partial_u G^\varepsilon(x,u) = g^\varepsilon(x,u)$ and $G^\varepsilon(x,0) = 0$. Since $g^\varepsilon(.,u)$ is piecewise constant, we can interpret the term $g^\varepsilon(u^\varepsilon)\partial_x u^\varepsilon$ as the distribution $\partial_x G^\varepsilon(x,u^\varepsilon)|_{\Gamma^\varepsilon \times (0,T)}$. Since we will work with $\partial_x G^\varepsilon(x,u) \, \mathbf{1}_{\Gamma^\varepsilon}(x) \in L^2(\Omega_T)$, we have well-defined traces $G^\varepsilon(x,u^\varepsilon)|_{\partial\Gamma^\varepsilon \times (0,T)}$. Since $G^\varepsilon(x,.)$ is invertible, this defines also traces of u^ε and gives a precise meaning to the interface conditions.

We write the interface condition (1.10) in a more reader-friendly form, considering ${p_c}^-$ as multivalued and ${p_c}^+$ as a function.

DEFINITION 1. *A saturation-flux pair* $(u^\varepsilon, F^\varepsilon)$ *is a weak solution of* (1.11) *on* $\Omega_T = (0,L)\times(0,T)$ *if* $u^\varepsilon \in L^\infty(\Omega_T,[0,1])$ *and* $F^\varepsilon \in L^2(\Omega_T)$ *satisfy*

$$
\begin{aligned}
\partial_t u^\varepsilon + \partial_x F^\varepsilon = 0 \qquad & \textit{in } \mathcal{D}'(\Omega_T),\\
F^\varepsilon = f^\varepsilon(u^\varepsilon) - \partial_x[G^\varepsilon(u^\varepsilon)] \qquad & \textit{in } \mathcal{D}'(\Gamma^\varepsilon\times(0,T)),\\
{p_c}^+(u^\varepsilon(2k\varepsilon-0)) \in {p_c}^-(u^\varepsilon(2k\varepsilon+0)) \qquad & \forall k\in\mathbb{Z}, 2k\varepsilon\in(0,L),\\
{p_c}^+(u^\varepsilon(2k\varepsilon+\varepsilon+0)) \in {p_c}^-(u^\varepsilon(2k\varepsilon+\varepsilon-0)) \qquad & \forall k\in\mathbb{Z}, 2k\varepsilon\in(0,L).
\end{aligned}
$$

A weak solution $(u^\varepsilon, F^\varepsilon)$ *is called an* entropy solution *if there exists a family of regularized equations and a corresponding family of solutions* $(u^\varepsilon_\eta, F^\varepsilon_\eta)$ *with* $u^\varepsilon_\eta \to u^\varepsilon$ *in* $L^2(\Omega_T)$ *and* $F^\varepsilon_\eta \rightharpoonup F^\varepsilon$ *in* $L^2(\Omega_T)$ *for* $\eta\to 0$.

LEMMA 1 (existence and a priori estimate). *For every* $\varepsilon>0$, *there exists an entropy solution* $(u^\varepsilon, F^\varepsilon)$. *With a constant* C *independent of* ε *there hold* $\|F^\varepsilon\|_{L^2(\Omega_T)} \le C$ *and the following regularity on the domain of positive saturation: For all* $\delta>0$ *there exists* C_δ *independent of* ε *such that*

$$
\int_0^T\int_0^L |\partial_x u^\varepsilon|^2 \mathbf{1}_{\{u^\varepsilon\ge\delta\}}\mathbf{1}_{\Gamma^\varepsilon} \le C_\delta. \tag{2.1}
$$

Proof. The regularized system ($\eta>0$) is a parabolic problem with finitely many transmission points and can be solved by standard methods. The maximum principle implies the bounds $0\le u^\varepsilon_\eta\le 1$. They allow us to select a subsequence $\eta\to 0$ and a weak $L^2(\Omega_T)$-limit u^ε. The monotonicity in t, shown in Lemma 3, implies the boundedness of $\partial_t u^\varepsilon_\eta \in L^1(\Omega_T)$. Estimate (2.1) for $\eta>0$ provides uniform bounds for the positive part $(u^\varepsilon_\eta-\delta)_+ \in L^2((0,T),H^1(\Gamma^\varepsilon))$; hence the sequence $(u^\varepsilon_\eta-\delta)_+$ is precompact in $L^2(\Omega_T)$ for every $\delta>0$. Choosing a diagonal sequence we find a subsequence with $u^\varepsilon_\eta\to u^\varepsilon$ strongly in $L^2(\Omega_T)$.

Below we derive an estimate for the sequence $F^\varepsilon_\eta\in L^2(\Omega_T)$ and thus, by boundedness of f^ε, there is also a uniform bound for $\partial_x G^\varepsilon_\eta(x,u^\varepsilon_\eta)\,\mathbf{1}_{\Gamma^\varepsilon}\in L^2(\Omega_T)$. We can choose a weakly convergent subsequence $F^\varepsilon_\eta \rightharpoonup F^\varepsilon$ in $L^2(\Omega_T)$ and find also $G^\varepsilon_\eta(u^\varepsilon_\eta)\to G^\varepsilon(u^\varepsilon)$ by the strong convergence of u^ε_η. This implies that the pair $(u^\varepsilon,F^\varepsilon)$ solves the conservation law and the characterization of F^ε of the second equation.

The solutions of the regularized problems satisfy the interface inclusions as equalities and we have a weak convergence of $\partial_x G(u^\varepsilon_\eta)$. The trace theorem implies that the limit u^ε satisfies again the interface conditions, and hence it is a weak solution of (1.11).

We now verify the a priori estimates, omitting everywhere the index $\eta>0$. We multiply the conservation law

$$
\partial_t u^\varepsilon + \partial_x[f(u^\varepsilon) - g^\varepsilon(u^\varepsilon)\partial_x u^\varepsilon] = 0 \text{ on } \Gamma^\varepsilon
$$

by the continuous function ${p_c}^\varepsilon(u^\varepsilon) - p_{max}$ and integrate by parts. Interior boundary integrals vanish due to the continuity of the flux. The right boundary integral vanishes because of the boundary condition ${p_c}^\varepsilon(L,u^\varepsilon(L,t)) = p_{max}$. We obtain

$$
\begin{aligned}
&\int_0^T\int_{\Gamma^\varepsilon}[{p_c}^\varepsilon(u^\varepsilon)-p_{max}]\partial_t u^\varepsilon + \int_0^T\int_{\Gamma^\varepsilon} g^\varepsilon(u^\varepsilon)\partial_x u^\varepsilon\,\partial_x[{p_c}^\varepsilon(u^\varepsilon)]\\
&\quad = \int_0^T\int_{\Gamma^\varepsilon} f^\varepsilon(u^\varepsilon)\,\partial_x[{p_c}^\varepsilon(u^\varepsilon)] + \int_0^T [f(u^\varepsilon)-g^\varepsilon(u^\varepsilon)\partial_x u^\varepsilon]_{x=0}\,({p_c}^{-,\eta}(0)-p_{max}).
\end{aligned}
$$

For the flux at the left boundary, which appears in the last integral, we calculate, using $\varphi(x,t) = L - x$,

$$\begin{aligned}
-L\int_0^T [f(u^\varepsilon) - g^\varepsilon(u^\varepsilon)\partial_x u^\varepsilon]_{x=0} &= \int_0^T \int_0^L \partial_x([f(u^\varepsilon) - g^\varepsilon(u^\varepsilon)\partial_x u^\varepsilon]\varphi) \\
&= \int_0^T \int_0^L \partial_x[f(u^\varepsilon) - g^\varepsilon(u^\varepsilon)\partial_x u^\varepsilon]\varphi - \int_0^T \int_0^L [f(u^\varepsilon) - g^\varepsilon(u^\varepsilon)\partial_x u^\varepsilon] \\
&= -\int_0^T \int_0^L \partial_t u^\varepsilon \varphi - \int_0^T \int_0^L [f(u^\varepsilon) - g^\varepsilon(u^\varepsilon)\partial_x u^\varepsilon] \\
&\le L^2 + C + \int_{\Omega_T} g^\varepsilon(u^\varepsilon)|\partial_x u^\varepsilon|.
\end{aligned}$$

We continue the above calculation, exploiting that ${p_c}^\varepsilon(u^\varepsilon)\partial_t u^\varepsilon$ is the time derivative of a bounded function, and use the uniform positivity $\partial_u {p_c}^\varepsilon \ge c_0 > 0$ and the bound $f^\pm \le c g^\pm$:

$$\int_0^T \int_{\Gamma^\varepsilon} g^\varepsilon(u^\varepsilon)\partial_u {p_c}^\varepsilon(u^\varepsilon)|\partial_x u^\varepsilon|^2 \le C + C\int_0^T \int_{\Gamma^\varepsilon} g^\varepsilon(u^\varepsilon)\partial_u {p_c}^\varepsilon(u^\varepsilon)|\partial_x u^\varepsilon|.$$

Application of the Cauchy–Schwarz inequality yields, by the boundedness of $g^\varepsilon\, \partial_u {p_c}^\varepsilon$, a bound for the left-hand side, independent of ε and η. Exploiting once more $\partial_u {p_c}^\varepsilon \ge c_0$ and $g^\pm > 0$ on $(0,1]$, this implies (2.1). It furthermore shows that the family $F^\varepsilon = [f(u^\varepsilon) - g^\varepsilon(x,u^\varepsilon)\partial_x u^\varepsilon]\mathbf{1}_{\Gamma^\varepsilon}$ is uniformly bounded in $L^2(\Omega_T)$. ∎

2.2. Comparison principles and monotonicity. In this subsection we derive results for entropy solutions of (1.11) and assume always the above boundary conditions and the initial condition ${p_c}^\varepsilon(u^\varepsilon(.,0)) = p_{max}$ on $(0,L)$ for $p_{max} \in ({p_c}^-(0), {p_c}^+(1))$.

LEMMA 2 (lower bound for u^ε). *There exist $\delta_0 > 0$ and $\varepsilon_0 > 0$ such that for all $\delta \in (0,\delta_0)$ and $\varepsilon \in (0,\varepsilon_0)$ the following holds. Let $(a,b) \subset (0,L)$ with $a,b \in 2\varepsilon\mathbb{Z} + \varepsilon$, and let u^ε be an entropy solution of* (1.11) *with*

$$u^\varepsilon(a-0,t) \ge \delta,\ u^\varepsilon(b-0,t) \ge \delta \quad \forall t \in (0,t_0).$$

Then there holds $u^\varepsilon \ge \delta$ on $(a,b)\times(0,t_0)$.

Proof. It is sufficient to show the claim for the regularized solutions u^ε_η and to take the limit $\eta \to 0$. We therefore perform all calculations for the more regular solutions u^ε_η, but we omit in what follows the fixed index η.

Our aim is to construct a 2ε-periodic stationary subsolution $U(y) = U^\varepsilon_\delta(y) \ge \delta$ for $y \in [a,b]$. For a parameter $q \in \mathbb{R}_+$ which denotes the constant flux of the subsolution, we define $U = U(.,q) : [0,2\varepsilon] \to \mathbb{R}$, $y \mapsto U(y)$ as the solution of

$$f^\varepsilon(y,U(y)) - g^\varepsilon(y,U(y))\,\partial_y U(y) = q \quad \text{in} \quad (0,\varepsilon)\cup(\varepsilon,2\varepsilon), \tag{2.2}$$

$$U(\varepsilon - 0) = \delta, \quad U(\varepsilon+0) = ({p_c}^+)^{-1}({p_c}^-(\delta)). \tag{2.3}$$

We note that, by the positivity of g^ε, (2.2) is an ordinary differential equation which can be solved locally with the boundary condition (2.3). The smallness of $\varepsilon > 0$ and the fact that $q - f^\varepsilon(y,\delta)$ becomes positive for $\delta \to 0$ imply that solutions can be defined on the whole intervals $(0,\varepsilon)$ and $(\varepsilon, 2\varepsilon)$. The solution operator defines a family of functions $U(.,q)$ and the continuous function $G = G_\delta : \mathbb{R}_+ \to \mathbb{R}$:

$$G(q) := {p_c}^+(U(2\varepsilon-0)) - {p_c}^-(U(0+0)).$$

Our aim is to choose the parameter $q = q^*$ with $G(q^*) = 0$ such that the function $U = U(.,q^*)$ can be extended to a 2ε-periodic stationary solution of (1.11).

For two special values of q we can evaluate the sign of $G(q)$. For $q_1 = f^-(\delta)$, the solution $U(.,q_1) = \delta$ is constant on the interval $(0,\varepsilon)$. On the interval $(\varepsilon, 2\varepsilon)$ the derivative $\partial_y U$ is positive by the positivity of $f^+(u^*)$; here we exploit the continuity of $f^\pm$, $f^-(0) = 0$ and the smallness of δ and ε. We find that

$$G(q_1) \geq {p_c}^+(U(\varepsilon + 0)) - {p_c}^-(\delta) = 0.$$

On the other hand, for $q_2 = f^+(U(\varepsilon+0))$, the solution $U(.,q_2)$ is constant on the interval $(\varepsilon, 2\varepsilon)$ and decreasing on $(0,\varepsilon)$. Monotonicity of ${p_c}^\pm$ implies $G(q_2) \leq 0$. By continuity, there is a critical value $q^* = q^*(\delta,\varepsilon) \in [q_1, q_2]$ with $G(q^*) = 0$, and we use $U = U(.,q^*)$ in what follows.

We now define the subsolution $U^\varepsilon(x,t)$ as the 2ε-periodic continuation of $U^\varepsilon_{\delta e^{-\lambda t}}(x)$, where we replaced δ by $\delta e^{-\lambda t}$ for a small constant $\lambda > 0$. We claim that u^ε can never touch the subsolution U^ε. Because $\partial_t U^\varepsilon < 0$, this is the standard comparison principle for all points x that are not contained in $\varepsilon\mathbb{Z}$. Let us assume that $t > 0$ is the first time instance at which the solutions touch each other and that, for a point $x \in \varepsilon\mathbb{Z}$, we have $u^\varepsilon(x+0,t) = U^\varepsilon(x+0,t)$. Since u^ε and U^ε both satisfy the p_c-jump condition, there also holds $u^\varepsilon(x-0,t) = U^\varepsilon(x-0,t)$. Then, because $u^\varepsilon(.,t) \geq U^\varepsilon(.,t)$, we have $\partial_x u^\varepsilon(x+0,t) \geq \partial_x U^\varepsilon(x+0,t)$ and $\partial_x u^\varepsilon(x-0,t) \leq \partial_x U^\varepsilon(x-0,t)$. The subsolution U^ε has the continuous flux $q^*(\delta e^{-\lambda t}, \varepsilon)$ and u^ε also has a continuous flux; since also $f(u^\varepsilon) = f(U^\varepsilon)$ and the same for g in the point x, the derivatives must coincide, $\partial_x u^\varepsilon(x+0,t) = \partial_x U^\varepsilon(x+0,t)$. As a consequence, also the fluxes of u^ε and U^ε coincide in x. In $(x+0,t)$ it holds that

$$-\partial_x[f^\varepsilon(u^\varepsilon) - g^\varepsilon(u^\varepsilon)\cdot\partial_x u^\varepsilon] = \partial_t u^\varepsilon < 0 = -\partial_x[f^\varepsilon(U^\varepsilon) - g^\varepsilon(U^\varepsilon)\cdot\partial_x U^\varepsilon],$$

and hence $\partial_x^2 u^\varepsilon < \partial_x^2 U^\varepsilon$, which is a contradiction to $u^\varepsilon \geq U^\varepsilon$. ∎

LEMMA 3 (monotonicity of u^ε). *Let u^ε be an entropy solution of* (1.11) *as above. Then the following hold:*

1. *Decay in time. The map $t \mapsto u^\varepsilon(x,t)$ is monotonically nonincreasing for almost every x.*
2. *Monotonicity in space. The map $k \mapsto u^\varepsilon(2k\varepsilon + y, t)$ is monotonically nondecreasing for every $t \in (0,T)$ and every $y \in [0, 2\varepsilon]$.*

Proof. As in the last proof, it suffices to verify the monotonicity for the approximate solutions $u = u^\varepsilon_\eta$. We therefore study, for the strictly positive coefficient $g = g(x,u) \geq \eta$ and the strictly monotone single-valued function $p_c = p_c(x,u)$, solutions u of

$$\begin{aligned} \partial_t u + \partial_x\left(f(x,u) - g(x,u)\partial_x u\right) &= 0,\\ [p_c(x,u)] &= 0,\\ [f(x,u) - g(x,u)\partial_x u(x)] &= 0, \end{aligned}$$

where the last two relations hold in points $x \in \mathbb{Z}\varepsilon$. By regularity theory for strictly parabolic equations we may assume that u is a classical solution of the above system.

Proof of part 1. We claim that $\partial_t u \leq 0$ holds on G. Indeed, the function $v = -\partial_t u$ is nonnegative at $t = 0$ and the equations are, with $\kappa = 0$,

$$\begin{aligned} \partial_t v + \partial_x\left(f_u(x,u)\cdot v - g_u(x,u)\cdot v\partial_x u - g(x,u)\partial_x v\right) &= \kappa,\\ [\partial_u p_c(x,u)\cdot v] &= 0,\\ [f_u(x,u)\cdot v - g_u(x,u)\cdot v\partial_x u(x) - g(x,u)\partial_x v(x)] &= 0. \end{aligned}$$

The boundary conditions are $v(0,t) = v(L,t) = 0$ for all $t \in (0,T)$. In order to show that v remains nonnegative, it suffices to show that the solutions v_κ of the above system with small right-hand side $\kappa \in \mathbb{R}$, $\kappa > 0$, remain nonnegative.

Let $t > 0$ be the first time instance such that $v_\kappa(.,t)$ has a zero in x. By the standard comparison principle, the zero cannot be in $(0,L) \setminus \varepsilon\mathbb{Z}$. Let us therefore assume $x \in \varepsilon\mathbb{Z}$. The first jump condition implies that, with v_κ vanishing on one side of x, it vanishes on both sides. Continuity implies $v_\kappa(.,t) \geq 0$, and hence we have the geometric conditions $\partial_x v_\kappa(x-0,t) \leq 0$ and $\partial_x v_\kappa(x+0,t) \geq 0$. The second jump condition then implies that $\partial_x v_\kappa = 0$ from both sides. The geometric condition $\partial_t v_\kappa(x,t) \leq 0$, together with $\kappa > 0$, implies

$$\partial_x \left(f_u(x,u) \cdot v_\kappa - g_u(x,u) \cdot v_\kappa \partial_x u - g(x,u)\partial_x v_\kappa\right) > 0$$

in the vicinity of x. We conclude that $\partial_x^2 v_\kappa < 0$ and thus a contradiction to $v_\kappa \geq 0$.

Proof of part 2. We claim that the function $v(x,t) = u(x+2\varepsilon,t) - u(x,t)$, defined on $(0, L-2\varepsilon)\times[0,T]$, is nonnegative for all times. Indeed, $v(.,0) = 0$ initially, and there hold $v(0,t) \geq 0$ and $v(L-2\varepsilon,t) \geq 0$ for all $t \in (0,T)$. With $u_+(x,t) := u(x+2\varepsilon,t)$, the equations for $v = u_+ - u$ read

$$\begin{aligned} \partial_t v + \partial_x \left(f(x,u_+) - f(x,u) - g(x,u_+)\partial_x u_+ + g(x,u)\partial_x u\right) = 0,& \\ [p_c(x,u_+) - p_c(u)] = 0,& \\ [f(x,u_+) - g(x,u_+)\partial_x u_+(x) - f(x,u) + g(x,u)\partial_x u(x)] = 0;& \end{aligned}$$

the last two lines indicate again jumps over the interface points. With appropriate evaluation points $\zeta_j(x,t)$ between $u(x,t)$ and $u_+(x,t)$ we may write this as

$$\begin{aligned} &\partial_t v + \partial_x \left(f_u(.,\zeta_1)v - g_u(.,\zeta_2)v\partial_x u_+ - g(.,u)\partial_x v\right) = 0, \\ &\{\partial_u p_c(.,\zeta_3)v\}(k\varepsilon + 0) = \{\partial_u p_c(.,\zeta_4)v\}(k\varepsilon - 0), \\ &\{f_u(.,\zeta_5)v - g_u(.,\zeta_6)v\partial_x u_+ - g(.,u)\partial_x v\}(k\varepsilon + 0) \\ &\qquad = \{f_u(.,\zeta_7)v - g_u(.,\zeta_8)v\partial_x u_+ - g(.,u)\partial_x v\}(k\varepsilon - 0) \end{aligned}$$

for $x = k\varepsilon$ with $k \in \mathbb{Z}$. Starting from this system for v, the nonnegativity of v follows as in part 1. ∎

LEMMA 4 (bounds for averages of u^ε). *Let u^ε be a family of entropy solutions of* (1.11) *as above.*

1. *Lower bound for averages. There exists $c \in \mathbb{R}$ such that for all $\varepsilon > 0$ and all $k \in \mathbb{Z}$ with $(2k\varepsilon - \varepsilon, 2k\varepsilon + \varepsilon) \subset (0,L)$, there holds*

$$\int_{2k\varepsilon-\varepsilon}^{2k\varepsilon+\varepsilon} u^\varepsilon(.,t) \geq \varepsilon(u^* - c\varepsilon). \tag{2.4}$$

2. *Upper bound for averages. For every $\rho > 0$ there exist $\tau > 0$, $\delta > 0$, and $\varepsilon_0 > 0$ such that for all $\varepsilon \in (0,\varepsilon_0)$, $k \in \mathbb{Z}$ with $(2k\varepsilon - \varepsilon, 2k\varepsilon + \varepsilon) \subset (0,L)$, and $t > \tau\varepsilon^2$,*

$$u^\varepsilon(2k\varepsilon + \varepsilon - 0, t - \tau\varepsilon^2) \leq \delta \quad \Rightarrow \quad \int_{2k\varepsilon-\varepsilon}^{2k\varepsilon+\varepsilon} u^\varepsilon(.,t) \leq \varepsilon(u^* + \rho). \tag{2.5}$$

Proof. We consider again approximate solutions u_η^ε from the definition of entropy solutions.

Proof of part 1. We approximate additionally the boundary condition at the left boundary by the artificial condition $u_\eta^\varepsilon(0,t) = \delta$. The subsolutions U^ε of Lemma 2 satisfy $U^\varepsilon \geq u^* - O(\varepsilon)$ on Γ_+^ε, independent of $\delta > 0$, such that $u_\eta^\varepsilon \geq U^\varepsilon$ provides (2.4).

Proof of part 2. We assume the contrary. Then, for some $\rho > 0$, for arbitrary $\tau > 0$, $\delta > 0$, there exists a sequence $\varepsilon_m \to 0$ such that inequality (2.5) fails to hold for some k and t. Now let $\rho > 0$ be such a value. Below we give an explicit choice of τ and δ that leads to a contradiction. We study now sequences of points k_m and time instances $t_m \geq \tau\varepsilon_m^2$ at which (2.5) fails for the sequence u^{ε_m}, which we continue up to time $2T$. We define the rescaled solutions

$$\begin{aligned}\tilde{U}^m &: (-1,1) \times (0, T/\varepsilon_m^2) \to \mathbb{R},\\ \tilde{U}^m(y,s) &:= u^{\varepsilon_m}(2k_m\varepsilon_m + \varepsilon_m y, t_m - \tau\varepsilon_m^2 + \varepsilon_m^2 s)\end{aligned}$$

and recall that we assume the failure of (2.5), that is,

$$\tilde{U}^m(1-0, s=0) \leq \delta, \quad \int_{-1}^{1} \tilde{U}^m(y,\tau)\, dy > u^* + \rho.$$

We now construct a function U^m which serves as an upper bound for $\tilde{U}^m$. We define U^m as the solution of the system

$$\begin{aligned}&\partial_s U^m + \partial_y \left(\varepsilon_m f(U^m) - g(U^m)\partial_y U^m\right) = 0 \text{ on } (-1,1) \setminus \{0\},\\ &U^m(1-0) = \delta, \quad U^m(-1+0) = ({p_c}^+)^{-1}({p_c}^-(\delta)),\\ &{p_c}^+(U^m(0-0)) = {p_c}^-(U^m(0+0)),\\ &[\varepsilon_m f(U^m) - g(U^m)\partial_y U^m](0-0) = [\varepsilon_m f(U^m) - g(U^m)\partial_y U^m](0+0),\end{aligned}$$

but now augmented with the initial condition $p_c(U^m(., s=0)) \equiv p_{max}$. As in the above proofs, exploiting $\tilde{U}^m(1-0,0) \leq \delta$ and the monotonicity of $\tilde{U}^m$, one derives the comparison principle $\tilde{U}^m \leq U^m$. The limit $U^\infty := \lim_{m\to\infty} U^m$ exists and solves the above system with ε_m replaced by 0. The solution U^∞ approaches, as $s \to \infty$, the stationary solution

$$\bar{U}^\infty(y) = \begin{cases}({p_c}^+)^{-1}({p_c}^-(\delta)) & \text{for } y \in (-1,0),\\ \delta & \text{for } y \in (0,1).\end{cases}$$

We can now derive a contradiction. Given $\rho > 0$, we choose $\delta > 0$ such that $\int_{-1}^1 \bar{U}^\infty(y)\, dy < u^* + \rho/3$. We then choose a time instance $\tau > 0$ such that $\int_{-1}^1 (U^\infty(y,\tau) - \bar{U}^\infty(y))\, dy < \rho/3$. With these choices we have

$$u^* + \rho < \int_{-1}^1 \tilde{U}^m(y,\tau)\, dy \leq \int_{-1}^1 U^m(y,\tau)\, dy \to \int_{-1}^1 U^\infty(y,\tau)\, dy \leq u^* + 2\rho/3,$$

a contradiction. ◻

3. Homogenization for a positive saturation. We next define the nonlinear flux function $\mathcal{F}(u^0, v^0)$ that maps an average oil saturation u^0 with an average slope v^0 to the effective flux. The continuity of the capillary pressure imposes a restriction on the values of u^0. Let $U \in [0,1]$ solve ${p_c}^-(U) = {p_c}^+(1)$. Then, with $u^0_{max} := (1+U)/2$, the flux function is a map

$$\mathcal{F} : [0, u^0_{max}] \times \mathbb{R} \mapsto \mathbb{R}, \ (u^0, v^0) \mapsto \mathcal{F}(u^0, v^0).$$

We set $\mathcal{F}(u^0, v^0) := 0$ for all (u^0, v^0) with $u^0 \leq u^*/2$ and construct $\mathcal{F}$ for other values with the help of nonlinear equations. For $(u^0, v^0) \in (u^*/2, u^0_{max}] \times \mathbb{R}$, the following system determines $(u_+, u_-) \in [0,1]^2$, representing typical values of u^ε in $\Gamma^\varepsilon_\pm$:

$$u_+ + u_- = 2u^0, \tag{3.1}$$

$$p_c{}^+(u_+) = p_c{}^-(u_-). \tag{3.2}$$

The monotonicity of $p_c{}^\pm$ assures the unique solvability of (3.1)–(3.2). We introduce auxiliary real numbers $u_{+,x}$ and $u_{-,x}$ that will describe the average slope of u_+ and u_- on a macroscopic scale. They are determined by

$$u_{+,x} + u_{-,x} = 2v^0, \tag{3.3}$$

$$\partial_u p_c{}^+(u_+)\, u_{+,x} = \partial_u p_c{}^-(u_-)\, u_{-,x}. \tag{3.4}$$

This linear system has a unique solution $u_{\pm,x}$ that depends linearly on v_0. We note that, for $v^0 \geq 0$, the average slope satisfies $0 \leq u_{\pm,x} \leq 2v^0$. We next introduce the pair $(v_+, v_-) \in \mathbb{R}^2$ which describes the typical derivatives of u^ε inside a single interval of $\Gamma^\varepsilon_\pm$. They are determined by

$$f^+(u_+) - g^+(u_+)v_+ = f^-(u_-) - g^-(u_-)v_-, \tag{3.5}$$

$$\partial_u p_c{}^+(u_+)v_+ + \partial_u p_c{}^-(u_-)v_- = \partial_u p_c{}^+(u_+)\, u_{+,x} + \partial_u p_c{}^-(u_-)\, u_{-,x}. \tag{3.6}$$

The unique solution $v_\pm$ depends in an affine way on $u_{\pm,x}$. We now define the effective flux function $\mathcal{F}$ as

$$\mathcal{F}(u^0, v^0) := f^+(u_+) - g^+(u_+)v_+, \tag{3.7}$$

where (u_+, v_+) is determined by the system (3.1)–(3.6) of nonlinear equations. For fixed $u \in [0, u^0_{max}]$, the map $\mathcal{F}(u, .)$ is affine. We may therefore also write $\mathcal{F}$ in the form

$$\mathcal{F}(u^0, v^0) = \mathcal{F}_0(u^0) - \mathcal{D}(u^0)v^0.$$

A flux function of this form appears also in [10] and [9]. We note that $\mathcal{F}$ is continuous: For $u^0 = u^*/2$, the solution of system (3.1)–(3.2) is $u_- = 0$ and $u_+ = u^*$, and hence $f^-(u_-) = g^-(u_-) = 0$, and (3.5) yields $\mathcal{F}(u^0, v^0) = 0$.

PROPOSITION 1 (homogenization). *Let $G = (a,b) \times (0, t_0)$ be a subdomain of $\Omega_T = (0,L) \times (0,T)$, $\delta, \varepsilon_0 > 0$ positive real numbers, and u^ε a family of solutions of* (1.11) *with*

$$u^\varepsilon \geq \delta \qquad \text{on } G \quad \forall \varepsilon \leq \varepsilon_0, \tag{3.8}$$

$$u^\varepsilon \rightharpoonup u^0, \quad F^\varepsilon \rightharpoonup F^0 \qquad \text{weakly in } L^2(G). \tag{3.9}$$

Then $u^0 \in L^2(G)$ solves $\partial_t u^0 + \partial_x F^0 = 0$ in the distributional sense on G and has a space derivative $\partial_x u^0 \in L^2(G)$, and, with $\mathcal{F}$ from (3.7)*, the following flux relation holds almost everywhere:*

$$F^0 = \mathcal{F}(u^0, \partial_x u^0). \tag{3.10}$$

Proof. The distributional conservation law is satisfied by the weak convergences. Exploiting the strictly positive diffusivity, (2.1) provides an estimate for the regular part of the derivative, and $\partial_x u^\varepsilon(x,t)\mathbf{1}_{\Gamma^\varepsilon}(x)$ is bounded in $L^2(G)$. In the subsequent proof we will define various two-scale limits for the above functions. For them we derive the relations (3.1)–(3.6), $F^0 = f^+(u_+) - g^+(u_+)v_+$, and $\partial_x u^0 = v^0 \in L^2(G)$. With these verifications, the proof is complete.

Step 1. *Two-scale limits and* (3.1). The uniform $L^2(G)$-bounds allow us to consider the two-scale limits

$$\begin{aligned} u^\varepsilon &\rightharpoonup u_0(x,t,y) \text{ two-scale},\\ \partial_x u^\varepsilon \mathbf{1}_{\Gamma^\varepsilon}(x) &\rightharpoonup v_0(x,t,y) \text{ two-scale}. \end{aligned}$$

The $L^2(G)$-estimate for $\partial_x u^\varepsilon$ immediately implies that u_0 is independent of y on the sets $(0,1)$ and $(1,2)$. Indeed, let $y \mapsto \varphi(y)$ be smooth with support contained in one of the two sets. We find, for $\Phi^\varepsilon(x) = \varepsilon\psi(x)\varphi(x/\varepsilon)$,

$$0 \leftarrow \int_G \partial_x u^\varepsilon \Phi^\varepsilon = \int_G u^\varepsilon \partial_x \Phi^\varepsilon \to \int_G \int_0^2 u_0(x,t,y)\psi(x)\partial_y\varphi(y)\, dy\, dx\, dt.$$

We conclude that u_0 has the special form

$$u_0(x,t,y) = u_-(x,t)\mathbf{1}_{(0,1)}(y) + u_+(x,t)\mathbf{1}_{(1,2)}(y). \tag{3.11}$$

The weak limit u^0 of the sequence u^ε coincides with the y-average of u_0; hence (3.11) implies relation (3.1).

We claim that also v_0 is piecewise constant. To see this, we use the test function $\Phi^\varepsilon(x,t) = \varepsilon\psi(x,t)\varphi(x/\varepsilon)$ with $\psi \in C_0^\infty(G)$ and exploit the equation. We assume here that φ is supported in $(1,2)$. We will verify the limit of the second line (marked with an exclamation mark) in the next step of the proof:

$$\begin{aligned} 0 \leftarrow \int_G \partial_t u^\varepsilon \Phi^\varepsilon &= \int_G f^\varepsilon(u^\varepsilon)\partial_x\Phi^\varepsilon - \int_G g^\varepsilon(u^\varepsilon)\partial_x u^\varepsilon \partial_x \Phi^\varepsilon \\ &\overset{!}{\to} \int_G \int_1^2 f^+(u_+(x,t))\, \psi(x,t)\partial_y\varphi(y)\, dy\, dx\, dt \\ &\quad - \int_G \int_1^2 g^+(u_+(x,t))v_0(x,t,y)\, \psi(x,t)\partial_y\varphi(y)\, dy\, dx\, dt. \end{aligned} \tag{3.12}$$

The first integral vanishes since φ is compactly supported in $(1,2)$ and we conclude that v_0 is independent of y, since u_+ is positive by the lower bound on u^ε. We can perform the same calculations with φ supported in $(0,1)$ to find the same equality with $+$ replaced by $-$,

$$v_0(x,t,y) = v_-(x,t)\mathbf{1}_{(0,1)}(y) + v_+(x,t)\mathbf{1}_{(1,2)}(y). \tag{3.13}$$

In particular, the quantities $u_\pm = u_\pm(x)$ and $v_\pm = v_\pm(x)$ that appear in (3.1)–(3.6) are now defined. For brevity, we will often suppress the dependence on t in the following.

Step 2. *Compactness.* To abbreviate notation we write $I = (a,b)$ for the spatial interval and set $\mathbf{1}_k := \mathbf{1}_{(2k\varepsilon, 2k\varepsilon+2\varepsilon)}$, $\mathbf{1}_k^- := \mathbf{1}_{(2k\varepsilon, 2k\varepsilon+\varepsilon)}$, and $\mathbf{1}_k^+ := \mathbf{1}_{(2k\varepsilon+\varepsilon, 2k\varepsilon+2\varepsilon)}$. We furthermore set $\mathbf{1}_+^\varepsilon := \sum_k \mathbf{1}_k^+$ and $\mathbf{1}_-^\varepsilon := \sum_k \mathbf{1}_k^-$. Our aim in this step of the proof is the following result. Let $h : [0,1] \to \mathbb{R}$ be a continuous function. Then

$$h(u^\varepsilon(x))\mathbf{1}_-^\varepsilon(x) - h(u_-(x))\mathbf{1}_-^\varepsilon(x) \to 0 \text{ strongly in } L^2(G), \tag{3.14}$$

and likewise for $-$ replaced by $+$. We note that this result justifies, with $h = g^+$, the convergence in (3.12). We emphasize that (3.14) is not a consequence of the previous results. For its proof we must control variations of u^ε on points in $2\mathbb{Z}\varepsilon$. Loosely speaking, it must jump down in $2k\varepsilon + 2\varepsilon$ as much as it jumped up in $2k\varepsilon + \varepsilon$.

In order to derive (3.14) we consider the capillary pressure function $P^\varepsilon(x) = {p_c}^\varepsilon(x, u^\varepsilon(x))$. This function has no jumps across interfaces, and hence the spatial derivative has no singular parts. On Γ^ε we have the estimate $|\partial_x P^\varepsilon| \le C|\partial_x u^\varepsilon|$, and therefore a uniform estimate for $P^\varepsilon \in L^2((0,t_0), H^1(I))$.

We have seen in Lemma 3 that $t \mapsto u^\varepsilon(x,t)$ is monotone for almost every $x \in I$. By the monotonicity of ${p_c}^\pm$, this implies the monotonicity of $t \mapsto P^\varepsilon(x,t)$. For the strong solutions of the strictly parabolic equations of the proposition we therefore have $|\partial_t P^\varepsilon| = -\partial_t P^\varepsilon$, and an integration yields $\|\partial_t P^\varepsilon\|_{L^1(G)} \le L\|P^\varepsilon\|_{L^\infty(G)}$, which is uniformly bounded. The spatial and the temporal regularity together provide the boundedness of P^ε in $W^{1,1}(G)$, hence the precompactness of P^ε in $L^1(G)$. Exploiting once more the uniform bound in $L^\infty(G)$, we find a subsequence that converges strongly in $L^2(G)$ and almost everywhere in G to a limit $P^0 \in L^2(G)$.

The convergence almost everywhere can be exploited to conclude the strong convergence of u^ε as claimed in (3.14). Since $(P^\varepsilon - P^0)\mathbf{1}^\varepsilon_- \to 0$ pointwise, also

$$u^\varepsilon \mathbf{1}^\varepsilon_- - ({p_c}^-)^{-1}(P^0)\mathbf{1}^\varepsilon_- \to 0 \tag{3.15}$$

pointwise almost everywhere and, by the uniform boundedness, also strongly in $L^2(G)$. In order to identify the limit function we recall that $u^\varepsilon \mathbf{1}^\varepsilon_- \to u_-(x)\mathbf{1}_{(0,1)}(y)$ in the sense of two-scale convergence. On the other hand, again in the sense of two-scale convergence, $({p_c}^-)^{-1}(P^0(x))\mathbf{1}^\varepsilon_-(x) \to ({p_c}^-)^{-1}(P^0)(x)\mathbf{1}_{(0,1)}(y)$, and hence $({p_c}^-)^{-1}(P^0) = u_-$.

We can now also apply a nonlinear continuous function h to both expressions in (3.15) and find $h(u^\varepsilon)\mathbf{1}^\varepsilon_- - h(u_-)\mathbf{1}^\varepsilon_- \to 0$ pointwise almost everywhere. By the uniform bounds for u^ε, this provides (3.14).

Step 3. *Derivation of the continuity conditions* (3.2) *and* (3.5) *and the flux equality.* With the help of (3.14) it is not difficult to derive the continuity conditions. The strong convergence $P^\varepsilon \to P^0$ in $L^2(G)$, together with $P^0 \in L^2((0,t_0), H^1(I))$, implies

$${p_c}^\varepsilon(u^\varepsilon)\mathbf{1}^\varepsilon_+ - {p_c}^\varepsilon(u^\varepsilon)\mathbf{1}^\varepsilon_- = P^\varepsilon \mathbf{1}^\varepsilon_+ - P^\varepsilon \mathbf{1}^\varepsilon_- \rightharpoonup 0 \text{ in } L^2(G).$$

On the other hand, by (3.14) and the two-scale convergences,

$$\begin{aligned}
{p_c}^\varepsilon(u^\varepsilon)\mathbf{1}^\varepsilon_+ - {p_c}^\varepsilon(u^\varepsilon)\mathbf{1}^\varepsilon_- &= {p_c}^+(u^\varepsilon)\mathbf{1}^\varepsilon_+ - {p_c}^-(u^\varepsilon)\mathbf{1}^\varepsilon_- \\
&= {p_c}^+(u_+)\mathbf{1}^\varepsilon_+ - {p_c}^-(u_-)\mathbf{1}^\varepsilon_- + o(1) \\
&\rightharpoonup \frac{1}{2}({p_c}^+(u_+) - {p_c}^-(u_-)).
\end{aligned}$$

Comparison of the two limits yields (3.2).

For the derivation of (3.5) we consider a test function $\varphi \in C_0^\infty((0,2),\mathbb{R})$, and $\Phi^\varepsilon(x) = \varepsilon\psi(x,t)\varphi(x/\varepsilon)$ as above. Exploiting (3.14) we find

$$\begin{aligned}
0 \leftarrow \int_G \partial_t u^\varepsilon \Phi^\varepsilon &= \int_G f(u^\varepsilon)\partial_x \Phi^\varepsilon - \int_G g^\varepsilon(u^\varepsilon)\partial_x u^\varepsilon \partial_x \Phi^\varepsilon \\
\to \int_G \int_0^1 & f^-(u_-)\psi\, \partial_y \varphi\, dy + \int_G \int_1^2 f^+(u_+)\psi\, \partial_y\varphi\, dy \\
- \int_G \int_0^1 & g^-(u_-)v_-\, \psi\partial_y \varphi\, dy - \int_G \int_1^2 g^+(u_+)v_+\, \psi \partial_y \varphi\, dy \\
= \int_G \psi\, & \left[f^-(u_-) - f^+(u_+)\right]\varphi(1) - \int_G \psi\, \left[g^-(u_-)v_- - g^+(u_+)v_+\right]\varphi(1).
\end{aligned}$$

Since ψ was arbitrary, this yields (3.5). In order to derive the flux equality, we exploit once more (3.14) and calculate

$$\begin{aligned} F^\varepsilon &= f^\varepsilon(u^\varepsilon) - g^\varepsilon(u^\varepsilon)\partial_x u^\varepsilon \\ &= [f^+(u^\varepsilon) - g^+(u^\varepsilon)\partial_x u^\varepsilon]\mathbf{1}_+^\varepsilon + [f^-(u^\varepsilon) - g^-(u^\varepsilon)\partial_x u^\varepsilon]\mathbf{1}_-^\varepsilon \\ &\rightharpoonup \frac{1}{2}[f^+(u_+) - g^+(u_+)v_+] + \frac{1}{2}[f^-(u_-) - g^-(u_-)v_-], \end{aligned}$$

which, because of (3.5), is the result for F^0.

Step 4. *The quantities* $u_{\pm,x}$ *and relation* (3.3). Our aim is to derive $u_{-,x} := \partial_x u_- \in L^2(G)$. Loosely speaking, we need an estimate for the oscillations of u^ε on Γ_-^ε. Such an estimate is the consequence of the corresponding estimate for the capillary pressures P^ε. Our construction serves also as a preparation for Step 5.

We introduce a function $\hat{P}^\varepsilon$ as a piecewise affine approximation of P^ε,

$$\begin{aligned} \hat{P}^\varepsilon(2k\varepsilon) &= \frac{1}{\varepsilon}\int_{2k\varepsilon}^{2k\varepsilon+\varepsilon} P^\varepsilon \qquad \forall k, \\ \hat{P}^\varepsilon \text{ affine on } &(2k\varepsilon, 2k\varepsilon + 2\varepsilon) \quad \forall k. \end{aligned}$$

Exploiting the $L^2((0,t_0),H^1(I))$-regularity of P^ε we find $\hat{P}^\varepsilon \to P^0$. Furthermore, as a projection of P^ε onto the space of piecewise affine functions, the projections $\hat{P}^\varepsilon$ are again bounded in $L^2((0,t_0),H^1(\Omega))$. Choosing a subsequence, we may assume

$$\partial_x \hat{P}^\varepsilon = \sum_k \mathbf{1}_k \frac{1}{2\varepsilon^2}\int_{2k\varepsilon}^{2k\varepsilon+\varepsilon} [P^\varepsilon(.+2\varepsilon) - P^\varepsilon(.)] \rightharpoonup \partial_x P^0 \text{ in } L^2(G).$$

We can now relate the function $\hat{P}^\varepsilon$ with a piecewise linear function $\hat{u}^\varepsilon$ that approximates u_-,

$$\begin{aligned} \hat{u}^\varepsilon(2k\varepsilon) &= \frac{1}{\varepsilon}\int_{2k\varepsilon}^{2k\varepsilon+\varepsilon} u^\varepsilon \qquad \forall k, \\ \hat{u}^\varepsilon \text{ affine on } &(2k\varepsilon, 2k\varepsilon + 2\varepsilon) \quad \forall k \end{aligned}$$

with derivative

$$\partial_x \hat{u}^\varepsilon = \sum_k \mathbf{1}_k \frac{1}{2\varepsilon^2}\int_{2k\varepsilon}^{2k\varepsilon+\varepsilon} [u^\varepsilon(.+2\varepsilon) - u^\varepsilon(.)].$$

We claim that the sequence $\partial_x \hat{u}^\varepsilon$ is uniformly bounded in $L^2(G)$. Our aim is to compare $\partial_x \hat{P}^\varepsilon$ with $\partial_u {p_c}^-(u_-) \cdot \partial_x \hat{u}^\varepsilon$. To this end we write the pressure derivative with the fundamental theorem and the function $\xi(x,\lambda) := \lambda u^\varepsilon(x+2\varepsilon) + (1-\lambda)u^\varepsilon(x)$ as

$$\partial_x \hat{P}^\varepsilon = \sum_k \mathbf{1}_k \frac{1}{2\varepsilon^2}\int_{2k\varepsilon}^{2k\varepsilon+\varepsilon} \left\{ \int_0^1 \partial_u {p_c}^-(\xi(x,\lambda))\, d\lambda \right\} [u^\varepsilon(x+2\varepsilon) - u^\varepsilon(x)]\, dx.$$

The nonnegativity of the integrand provided by Lemma 3 and the lower bound for $\partial_u {p_c}^-$ imply a uniform bound for $\partial_x \hat{u}^\varepsilon \in L^2(G)$. In particular, we may assume that $\hat{u}^\varepsilon$ converges strongly in $L^2(G)$; the limit is easily identified with the weak limit u_-.

Furthermore, choosing a subsequence, we may assume that $\partial_x \hat{u}^\varepsilon$ converges weakly in $L^2(G)$. Denoting the limit function by $u_{-,x}$ we have

$$\partial_x \hat{u}^\varepsilon \rightharpoonup u_{-,x} := \partial_x u_- \text{ in } L^2(G).$$

In a similar way one constructs functions $\tilde{u}^\varepsilon$ that approximate u_+ with $\partial_x \tilde{u}^\varepsilon$ bounded in $L^2(G)$. We may assume that also this sequence converges weakly, $\partial_x \tilde{u}^\varepsilon \rightharpoonup u_{+,x} := \partial_x u_+$ in $L^2(G)$.

The weak convergences $\hat{u}^\varepsilon \rightharpoonup u_-$ and $\tilde{u}^\varepsilon \rightharpoonup u_+$, together with (3.1), imply $\hat{u}^\varepsilon + \tilde{u}^\varepsilon \rightharpoonup 2u^0$. Then the distributional derivatives converge as well, and we conclude (3.3) with $v^0 = \partial_x u^0 \in L^2(G)$.

Step 5. *Derivation of* (3.4) *and* (3.6). We have seen that the capillary pressure functions P^ε are bounded in $L^2((0,t_0), H^1(I))$ and that we may therefore assume $\partial_x P^\varepsilon \rightharpoonup \partial_x P^0$ in $L^2(G)$. In this last step of the proof we calculate the derivative $\partial_x P^0$ in three different ways.

The most direct approach is to calculate with the chain rule, exploiting (3.14) in the second equality,

$$\begin{aligned}
\partial_x P^0 \leftharpoonup \partial_x P^\varepsilon &= \partial_u {p_c}^-(u^\varepsilon)\, \partial_x u^\varepsilon\, \mathbf{1}_-^\varepsilon + \partial_u {p_c}^+(u^\varepsilon)\, \partial_x u^\varepsilon\, \mathbf{1}_+^\varepsilon \\
&= \partial_u {p_c}^-(u_-)\, \partial_x u^\varepsilon\, \mathbf{1}_-^\varepsilon + \partial_u {p_c}^+(u_+)\, \partial_x u^\varepsilon\, \mathbf{1}_+^\varepsilon + o(1) \\
&\rightharpoonup \frac{1}{2}\partial_u {p_c}^-(u_-)\, v_- + \frac{1}{2}\partial_u {p_c}^+(u_+)\, v_+.
\end{aligned}$$

We will now calculate $\partial_x P^0$ in a different way. We introduce the function $P^{*,\varepsilon} := {p_c}^-(\hat{u}^\varepsilon)$. The monotonicity of $\hat{u}^\varepsilon$ in t implies a compactness and allows us to assume the strong and the pointwise almost everywhere convergence $\hat{u}^\varepsilon \to u_-$. We calculate with the chain rule

$$\partial_x P^{*,\varepsilon} = \partial_u {p_c}^-(\hat{u}^\varepsilon) \cdot \partial_x \hat{u}^\varepsilon \rightharpoonup \partial_u {p_c}^-(u_-) \cdot u_{-,x} \text{ in } L^2(G).$$

On the other hand, $P^{*,\varepsilon} = {p_c}^-(\hat{u}^\varepsilon) \to {p_c}^-(u_-) = P^0$, and therefore the distributional limits coincide,

$$\partial_x P^0 = \partial_u {p_c}^-(u_-) \cdot u_{-,x}.$$

The above calculation can also be performed with averages over the set Γ_ε^+ and with the function p_c^+. We find the analogous formula $\partial_x P^0 = \partial_u p_c^+(u_+) \cdot u_{+,x}$ and thus (3.4). In (3.6) we use the symmetric version

$$\partial_x P^0 = \frac{1}{2}\partial_u {p_c}^+(u_+) u_{+,x} + \frac{1}{2}\partial_u {p_c}^-(u_-) u_{-,x}\ .$$

From our first calculation of $\partial_x P^0$ we see that the weighted average of $u_{+,x}$ and $u_{-,x}$ coincides with the weighted average of v_+ and v_-, as claimed in (3.6). ∎

As a preparation for the investigation of the interface condition in the free boundary value problem, we investigate the regularity of solutions in the region of strictly positive saturation.

LEMMA 5 (Hölder's estimate). *We consider a family of entropy solutions u^ε and fix positive numbers C_0 and δ. We assume that $t \in (0,T)$ is a time instance of bounded energy in the sense that*

$$\int_{\Gamma^\varepsilon} |\partial_x u^\varepsilon(.,t)|^2\, \mathbf{1}_{\{u^\varepsilon \geq \delta/2\}} \leq C_0^2. \tag{3.16}$$

Then there exist a constant $C_L = C_L(\delta)$*, independent of* C_0 *and* ε*, and a constant* $\varepsilon_0 = \varepsilon_0(\delta, C_0)$ *such that the following hold:*

1. *Let* $a = 2k\varepsilon + \varepsilon \in (0, L)$ *with* $k \in \mathbb{Z}$ *and* $u^\varepsilon(a - 0, t) \geq \delta$*, and let* $b \in 2\mathbb{Z}\varepsilon + \varepsilon$*,* $b > a$*. Then*

$$|u^\varepsilon(b - 0, t) - u^\varepsilon(a - 0, t)| \leq C_L C_0 \sqrt{|b - a|}. \tag{3.17}$$

2. *For all* $\varepsilon \leq \varepsilon_0(\delta, C_0)$ *there holds*

$$u^\varepsilon(2k\varepsilon + \varepsilon - 0, t) \geq \delta \quad \Rightarrow \quad u^\varepsilon(2k\varepsilon - \varepsilon - 0, t) \geq \delta/2. \tag{3.18}$$

Proof. Cellwise estimate and (3.17). The monotonicity of Lemma 3, together with the lower bound of Lemma 2, implies $u^\varepsilon \geq \delta$ on $(a, L) \times (0, t_0)$, and hence (3.16) provides an $L^2(\Gamma^\varepsilon)$-bound for the spatial derivative. We claim that locally, across a single interval $(2k\varepsilon + \varepsilon, 2k\varepsilon + 3\varepsilon)$, we can control differences of the u^ε values by the integral of the derivative. Indeed, with the variables

$$\begin{aligned} &y_0 := u^\varepsilon(2k\varepsilon + \varepsilon - 0, t), \quad y_1 := u^\varepsilon(2k\varepsilon + 2\varepsilon + 0, t), \quad y_2 := u^\varepsilon(2k\varepsilon + 3\varepsilon - 0, t), \\ &z_0 := u^\varepsilon(2k\varepsilon + \varepsilon + 0, t), \quad z_1 := u^\varepsilon(2k\varepsilon + 2\varepsilon - 0, t), \end{aligned}$$

we have the relations

$$\begin{aligned} y_2 - y_1 &= \int_{2k\varepsilon+2\varepsilon}^{2k\varepsilon+3\varepsilon} \partial_x u^\varepsilon(.,t) =: \Delta_1, \\ z_1 - z_0 &= \int_{2k\varepsilon+\varepsilon}^{2k\varepsilon+2\varepsilon} \partial_x u^\varepsilon(.,t) =: \Delta_2, \\ z_0 &= \Phi(y_0) \text{ and } z_1 = \Phi(y_1) \quad \text{for } \Phi(y) := ({p_c}^+)^{-1}({p_c}^-(y)). \end{aligned}$$

They imply

$$y_2 = y_1 + \Delta_1 = \Phi^{-1}(z_0 + \Delta_2) + \Delta_1 = \Phi^{-1}(\Phi(y_0) + \Delta_2) + \Delta_1.$$

Since Φ and its inverse Φ^{-1} have a bounded derivative on $\{y \geq \delta\}$ we conclude the local estimate

$$|y_2 - y_0| \leq C_L(\delta)\,(|\Delta_1| + |\Delta_2|) \leq C_L \int_{2k\varepsilon+\varepsilon}^{2k\varepsilon+3\varepsilon} |\partial_x u^\varepsilon(.,t)| \mathbf{1}_{\Gamma^\varepsilon}. \tag{3.19}$$

Adding the inequalities (3.19) from $k = (a - \varepsilon)/(2\varepsilon)$ to $k' = (b - \varepsilon)/(2\varepsilon) - 1$, we find

$$\begin{aligned} |u^\varepsilon(b - 0, t) - u^\varepsilon(a - 0, t)| &\leq C_L \int_a^b |\partial_x u^\varepsilon(.,t)| \mathbf{1}_{\Gamma^\varepsilon} \\ &\leq C_L\, |b - a|^{1/2} \left(\int_a^L |\partial_x u^\varepsilon(.,t)|^2 \mathbf{1}_{\Gamma^\varepsilon} \right)^{1/2}. \end{aligned}$$

This is estimate (3.17).

Implication (3.18) *on jumps.* Let C_0 be fixed and let t be a time instance with $\|\partial_x u^\varepsilon(.,t) \mathbf{1}_{\{u^\varepsilon \geq \delta/2\}} \mathbf{1}_{\Gamma^\varepsilon}\|_{L^2}^2 \leq C_0$. As shown in (3.19), we have the estimate

$$\begin{aligned} |u^\varepsilon(2k\varepsilon + \varepsilon - 0) - u^\varepsilon(2k\varepsilon - \varepsilon - 0)| &\leq c \left(\int_{2k\varepsilon-\varepsilon}^{2k\varepsilon+\varepsilon} |\partial_x u^\varepsilon(.,t)|^2 \mathbf{1}_{\Gamma^\varepsilon} \right)^{1/2} \sqrt{\varepsilon} \\ &\leq c C_0 \sqrt{\varepsilon}, \end{aligned}$$

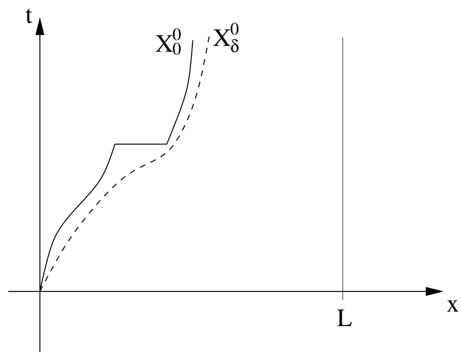


FIG. 6. *Possible shapes of the free boundaries X_δ^0 and X_0^0. We know that they are monotone and that $X_\delta^0 \to X_0^0$ pointwise almost everywhere. With the above graphs we illustrate that X_0^0 may have jumps and that we cannot expect $X_\delta^0(t) \to X_0^0(t)$ for every $t > 0$.*

at least if we can assume (for the last inequality) that the saturation satisfies $u^\varepsilon \geq \delta/2$ in every point of the interval $(2k\varepsilon - \varepsilon, 2k\varepsilon + \varepsilon)$. We choose $\varepsilon_0 = \varepsilon_0(\delta, C_0)$ such that $c\,C_0\sqrt{\varepsilon_0} \leq \delta/4$. The arguments above can be repeated for every point x in the interval $(2k\varepsilon, 2k\varepsilon + \varepsilon)$. Continuity of u^ε inside the interval allows us to conclude that $u^\varepsilon(2k\varepsilon + 0) \geq 3\delta/4$. We repeat the argument on the interval $(2k\varepsilon - \varepsilon, 2k\varepsilon)$ and find the result. ∎

4. The free boundary problem. We study, for $\delta > 0$ and $\varepsilon > 0$, the free boundary separating the region of uniformly positive saturation from the rest:

$$X_\delta^\varepsilon(t) := \inf\left\{x \in (0, L) \cap (2\varepsilon\mathbb{Z} + \varepsilon) : \ u^\varepsilon(x - 0, t) \geq \delta\right\}, \tag{4.1}$$

$$X_0^\varepsilon(t) := \inf_{\delta > 0} X_\delta^\varepsilon(t). \tag{4.2}$$

We set $X_\delta^\varepsilon(t) = L$ if the infimum is taken over the empty set.

LEMMA 6. *There are sequences $\varepsilon_k \searrow 0$ and $\delta_m \searrow 0$ such that, for every $\varepsilon = \varepsilon_k$ and every $\delta = \delta_m$, the following hold:*

1. *The maps $t \mapsto X_\delta^\varepsilon(t)$ and $t \mapsto X_0^\varepsilon(t)$ are monotonically nondecreasing.*
2. *The following limits hold pointwise for almost every t, and the limits are monotone functions:*

$$X_\delta^0(t) = \lim_{k\to\infty} X_\delta^{\varepsilon_k}(t), \qquad X_0^0(t) = \lim_{m\to\infty} X_{\delta_m}^0(t),$$
$$X_0^\varepsilon(t) = \lim_{m\to\infty} X_{\delta_m}^\varepsilon(t), \qquad \tilde{X}_0^0(t) = \lim_{k\to\infty} X_0^{\varepsilon_k}(t).$$

   We can select an upper semicontinuous representative $t \mapsto X(t)$ of the L^1-function $t \mapsto X_0^0(t)$.
3. *There holds $X_\delta^0 \leq X_{\delta'}^0$ for all $\delta \leq \delta'$ and $\tilde{X}_0^0 \leq X_0^0$.*

Proof. Figure 6 indicates possible shapes of X_0^0 and X_δ^0 and recalls the fact that these functions need not be continuous. Lemma 3 provides that the function $t \mapsto u^\varepsilon(x, t)$ is monotonically nonincreasing. This implies the monotonicity of the free boundaries stated in statement 1. The monotonicity of the family of functions $t \mapsto X_\delta^\varepsilon$ implies the uniform boundedness in $BV([0, T], \mathbb{R})$, and hence we can extract subsequences that converge strongly in L^1 and pointwise almost everywhere. Limits of monotone functions are again monotone. Since BV-functions have only countably many jumps, we find an upper semicontinuous representative.

The monotonicity in δ is an immediate consequence of the definition of X^ε_δ. It justifies the infimum of (4.2) and implies $X^\varepsilon_0 \le X^\varepsilon_\delta$, which carries over in the limit $k \to \infty$ as $\tilde{X}^0_0 \le X^0_\delta$. The limit $\delta_m \to 0$ yields statement 3. ∎

With the help of the limiting free boundaries we can transform the results of Proposition 1 into the following statement.

COROLLARY 1 (limit equations in region G). *Let $t \mapsto X(t)$ be as in Lemma 6 and G the open domain:*

$$G := \{(x,t) \in (0,L) \times (0,T) : x > X(t)\}.$$

Let (u^0, F^0) be limits of entropy solutions $(u^\varepsilon, F^\varepsilon)$ as in Theorem 1 and $G' \Subset G$ a subset of G. Then there holds $\partial_x u^0 \in L^2(G')$ and F^0 satisfies on G' the relation $F^0 = \mathcal{F}(u^0, \partial_x u^0)$.

Proof. The function X is monotone and the closure of G' is a compact subset of G; hence G' can be covered by a finite collection of sets $G_0 = (x_0, L) \times (0, t_0)$ with $x_0 > X(t_0)$. It suffices to verify the statements on one such subset G_0. Our aim is to find $\varepsilon_0 > 0$ and $\delta > 0$ such that $u^\varepsilon \ge \delta$ on G_0 for all $\varepsilon \le \varepsilon_0$. Once this is done, the application of Proposition 1 yields the result.

We start by choosing $\eta > 0$ such that $x_0 - \eta > X(t_0 + \eta)$, which is possible, since X is upper semicontinuous, and hence $\limsup_{\eta \searrow 0} X(t_0 + \eta) \le X(t_0) < x_0$. We now choose $\delta > 0$ such that

$$x_0 - \frac{\eta}{2} > X^0_\delta \left(t_0 + \frac{\eta}{2}\right). \tag{4.3}$$

In order to verify that for a small $\delta > 0$ relation (4.3) is satisfied, we exploit that by monotonicity $X(t) < x_0 - \eta$ for all $t \in (t_0 + \eta/2, t_0 + \eta)$. The strong L^2-convergence $X^0_\delta \to X$ and Egoroff's theorem imply that, for $\delta > 0$ small, $|X^0_\delta - X| < \eta/2$ except for a set of t's with measure less than $\eta/2$. For such δ there necessarily exists $s \in (t_0 + \eta/2, t_0 + \eta)$ with $X^0_\delta(s) < x_0 - \eta/2$. By monotonicity of X^0_δ, relation (4.3) holds.

We finally want to choose, in a similar way, a number $\varepsilon_0 > 0$ with $x_0 > X^\varepsilon_\delta(t_0)$. We have $X^0_\delta(t) < x_0 - \eta/2$ for all $t \in (t_0, t_0 + \eta/2)$. By Egoroff's theorem we find $\varepsilon_0 > 0$ such that, for all $\varepsilon \le \varepsilon_0$, we find some $s \in (t_0, t_0 + \eta/2)$ such that $X^\varepsilon_\delta(s) < x_0$. The monotonicity of $X^\varepsilon_\delta(s)$ implies $x_0 > X^\varepsilon_\delta(t_0)$ and thus $G_0 \subset (X^\varepsilon_\delta(t_0), L) \times (0, t_0)$. The definition of X^ε_δ implies the desired lower bound for the sequence u^ε on the left boundary of the domain $(X^\varepsilon_\delta(t_0), L) \times (0, t_0)$. Lemma 2 yields the lower bound on the whole domain. ∎

PROPOSITION 2 (limit equations in region B). *In the domain*

$$B := \{(x,t) \in (0,L) \times (0,T) : x \le X(t)\},$$

there holds that $u^0 \equiv \frac{u^*}{2}$ and $F^0 \equiv 0$ almost everywhere. The function u^0 has no jump across $\partial B \cap \Omega_T$ in the following sense: Let $T_0 \in (0,T]$ be a time instance with $X(T_0) < L$, and let A_r, $r > 0$, be a family of averages of u^0,

$$A_r := \frac{1}{T_0 \cdot r} \int_0^{T_0} \int_0^r u^0(X(t) + s, t)\, ds\, dt. \tag{4.4}$$

Then A_r satisfies

$$A_r \to \frac{u^*}{2} \text{ for } r \to 0.$$

Proof. We select monotone sequences $\delta_j \to 0$ and $\varepsilon_m \to 0$ with the convergences of the free boundaries as in Lemma 6 and with $u^{\varepsilon_m} \to u^0$ weakly in $L^2(\Omega_T)$. Almost all time instances $t \in (0,T)$ are points of continuity of the function $X(.)$ and of all functions $X^0_{\delta_j}(.)$, $j \in \mathbb{N}$. Furthermore, in almost every point $t \in (0,T)$ the convergences of Lemma 6 hold. In the following we consider only time instances t with all these properties.

Step 1. *On B it holds that $u^0 = u^*/2$.* We note that $u^0 \geq u^*/2$ follows immediately from the lower bound in Lemma 4.

For the upper bound let $(x,t) \in B$ be with t as above and with $x < X(t)$. Moreover, let $\rho > 0$ be arbitrary. We choose $\tau > 0$ and an index $j \in \mathbb{N}$ such that, for $\delta = \delta_j$, implication (2.5) of Lemma 4 holds. Since X has no jump in t we find $t' < t$ with $x < X(t')$. By monotonicity in δ, we have $x < X^0_\delta(t')$. We choose $r > 0$ smaller than $\frac{1}{2}(X^0_\delta(t') - x)$; for later use we also demand $r < \frac{1}{2}(t - t')$. We find $m_0 > 0$ such that, for all $m \geq m_0$, additionally $r < X^{\varepsilon_m}_\delta(t') - x$. We may choose m_0 large enough to satisfy additionally $\tau\varepsilon^2_{m_0} < t - t' - r$ and $\varepsilon_{m_0} < \varepsilon_0$ of Lemma 4. The upper bound of Lemma 4 provides, for $\varepsilon = \varepsilon_m$ with $m \geq m_0$,

$$u^\varepsilon(2k\varepsilon + \varepsilon - 0, t') \leq \delta \quad \Rightarrow \quad \int_{2k\varepsilon-\varepsilon}^{2k\varepsilon+\varepsilon} u^\varepsilon(., t' + \tau\varepsilon^2) \leq \varepsilon(u^* + \rho).$$

By construction, the assertion is satisfied for all $k \in \mathbb{Z}$ with $2k\varepsilon + \varepsilon \leq x + r$. The monotone decay of u^ε in t implies

$$\frac{1}{|B_r((x,t))|} \int_{B_r((x,t))} u^{\varepsilon_m} \leq \frac{1}{2}(u^* + \rho). \tag{4.5}$$

This carries over to the weak limit u^0. Since ρ is arbitrary, we have the upper bound in B by the Lebesgue differentiation theorem.

Step 2. *Boundary condition.* We assume that a small number $\rho > 0$ is given; our aim is to choose $r > 0$ small to have $A_r \leq u^*/2 + c\rho$ for some universal constant c. We recall that A_r is defined with an integration over the thin region

$$U_r := \{(x,t) \in \Omega_{T_0} : X(t) < x < X(t) + r\}.$$

We use the numbers $\delta_0 > 0$ and $\tau > 0$ that appear in the upper bound for averages in Lemma 4 and choose $\delta = \delta_j < \delta_0/2$. We consider the ε-dependent set

$$E_0 := \left\{(x,t) \in \Omega_{T_0} : x \in (0,L), \int_0^L |\partial_x u^\varepsilon(.,t)|^2 \, \mathbf{1}_{\{u^\varepsilon \geq \delta/2\}} > C_0^2 \right\},$$

where we denote by $\partial_x u^\varepsilon$ the regular part of the derivative. Choosing C_0 large enough we achieve $|E_0 \cap U_r| \leq \rho r T_0$ for all ε. This is possible since by estimate (2.1) the time integral over the above spatial integral is bounded. We now choose $r > 0$ small enough to satisfy, with $C_L = C_L(\delta/2)$ of Lemma 5, $C_L C_0 (4r)^{1/2} \leq \delta$.

In order to show the upper bound for A_r we may still choose $\varepsilon > 0$ small. We define further ε-dependent exceptional sets as

$$\begin{aligned}
E_1 &:= \left\{(x,t) \in \Omega_{T_0} : x \in (0,L), |X^\varepsilon_\delta(t) - X^0_\delta(t)| \geq r\right\}, \\
E_1' &:= \left\{(x,t) \in \Omega_{T_0} : x \in (0,L), t \geq \tau\varepsilon^2, |X^\varepsilon_\delta(t - \tau\varepsilon^2) - X^0_\delta(t - \tau\varepsilon^2)| \geq r\right\}, \\
E_2 &:= \left\{(x,t) \in \Omega_{T_0} : X^\varepsilon_\delta(t - \tau\varepsilon^2) \leq x \leq X^\varepsilon_\delta(t)\right\}, \\
E_3 &:= \left\{(x,t) \in \Omega_{T_0} : |X(t) - X(t - \tau\varepsilon^2)| \geq r\right\}.
\end{aligned}$$

For the first set we achieve $|E_1| \le \rho r T_0$ for all small ε by the L^1-convergence $X^\varepsilon_\delta \to X^0_\delta$. The set E_1' is obtained from E_1 by a shift, and hence this set also satisfies $|E_1'| \le \rho r T_0$. The set E_2 is contained in a $\tau\varepsilon^2$-neighborhood of the free boundary

$$\Sigma^\varepsilon_\delta := \left\{ (x,t) \in \Omega_{T_0} : \lim_{s \nearrow t} X^\varepsilon_\delta(s) \le x \le \lim_{s \searrow t} X^\varepsilon_\delta(s) \right\}.$$

The set $\Sigma^\varepsilon_\delta$ is a curve of finite length, and hence we achieve $|E_2| \le \rho r T_0$ for $\varepsilon > 0$ small. Finally, $|E_3| \le \rho r T_0$ for $\varepsilon > 0$ small, since X is a BV-function. We may additionally impose on ε that $\varepsilon < r$ and $\varepsilon < \varepsilon_0(\delta, C_0)$; the latter allows us to use the implication of Lemma 5 outside the set E_0,

$$u^\varepsilon(2k\varepsilon + \varepsilon - 0, t) \ge \delta \quad \Rightarrow \quad u^\varepsilon(2k\varepsilon - \varepsilon - 0, t) \ge \delta/2. \tag{4.6}$$

We note that the set

$$E_0' := \left\{ (x,t) \in (0,L) \times (\tau\varepsilon^2, T_0) : \int_0^L |\partial_x u^\varepsilon(.,t-\tau\varepsilon^2)|^2 \,\mathbf{1}_{\{u^\varepsilon(.,t-\tau\varepsilon^2)\ge\delta/2\}} > C_0^2 \right\}$$

also satisfies $|E_0' \cap U_r| \le \rho r T_0$, since it is obtained by a shift of the set E_0.

After these preparations, let us now consider an arbitrary point $(x,t) = (2k\varepsilon + \varepsilon, t) \in U_r \setminus (E_0 \cup E_0' \cup E_1 \cup E_1' \cup E_2 \cup E_3)$. We distinguish two cases.

Case (i). $u^\varepsilon(x,t)$ is small, $x = 2k\varepsilon + \varepsilon < X^\varepsilon_\delta(t)$. Since (x,t) is not contained in E_2, we also have $x < X^\varepsilon_\delta(t - \tau\varepsilon^2)$. Lemma 4 can be applied with the point $(2k\varepsilon + \varepsilon, t - \varepsilon^2\tau)$ and yields

$$\int_{2k\varepsilon-\varepsilon}^{2k\varepsilon+\varepsilon} u^\varepsilon(.,t) \le \varepsilon(u^* + \rho).$$

Case (ii). $u^\varepsilon(x,t)$ is large, $x = 2k\varepsilon + \varepsilon \ge X^\varepsilon_\delta(t)$. We will derive the smallness of $u^\varepsilon(x,t)$ with the help of the Hölder-type estimate of Lemma 5 and conclude again with Lemma 4.

We start by setting $t' := t - \tau\varepsilon^2$ and denote by k' the integer with $2k'\varepsilon + \varepsilon = x' := X^\varepsilon_\delta(t')$. The definition of X^ε_δ implies $u^\varepsilon(x' - 0, t') \ge \delta$, $(x', t') \notin E_0'$ allows us to use (4.6) at the time instance t', and we conclude that $\delta/2 \le u^\varepsilon(2k'\varepsilon - \varepsilon - 0, t') < \delta$. The lower bound allows us to apply the first part of Lemma 5 with $a = 2k'\varepsilon - \varepsilon$ and $b = x = 2k\varepsilon + \varepsilon$. We find

$$u^\varepsilon(x - 0, t') \le u^\varepsilon(2k'\varepsilon - \varepsilon - 0, t') + C_L(\delta/2)\, C_0 (4r)^{1/2} \le \delta + \delta = 2\delta.$$

We used here $(x,t) \notin E_3$ and $(x,t) \notin E_1'$ such that

$$\begin{aligned} x - x' &= (x - X(t)) + (X(t) - X(t')) + (X(t') - X^0_\delta(t')) + (X^0_\delta(t') - X^\varepsilon_\delta(t')) \\ &\le r + r + 0 + r = 3r. \end{aligned}$$

Another application of the upper bound of Lemma 4 yields also in this case

$$\int_{2k\varepsilon-\varepsilon}^{2k\varepsilon+\varepsilon} u^\varepsilon(.,t) \le \varepsilon(u^* + \rho).$$

In both cases we find the same estimate for averages of u^ε. Summation over k and an integration over $t \in (0, T_0)$ yield

$$\frac{1}{T_0 \cdot r} \int_{U_r} u^\varepsilon \le \frac{u^*}{2} + c\rho + O(\varepsilon) \frac{1}{T_0 \cdot r}.$$

The factor c covers the error induced by the exceptional sets, and the error term $O(\varepsilon)$ is induced by the integration over a boundary strip of width 2ε that is necessary to cover U_r with intervals of the form $(2k\varepsilon, 2k\varepsilon + 2\varepsilon)$. We take the limit $\varepsilon \to 0$. Since $\rho > 0$ is arbitrary, we find (4.4).

Step 3. *On B it holds that $F^0 = 0$.* We have shown in Step 1 that u^0 is constant in B, and hence the conservation law implies $\partial_x F^0 = 0$ and we have $F^0(x,t) = F^0(t)$ for almost all $(x,t) \in B$. Our aim is to conclude $F^0(t) = 0$ for almost all t.

Inequality $F^0 \le 0$. For the approximate solutions u^ε, the boundary condition $u^\varepsilon(0,t) = 0$, together with $f(0) = 0$, $g \ge 0$, and $u^\varepsilon \ge 0$, implies $F^\varepsilon(0,t) = \mathrm{trace}\{f(u^\varepsilon) - g(u^\varepsilon)\partial_x u^\varepsilon\} \le 0$. This can be written in a weak form as

$$\int_0^T \int_0^L \{u^\varepsilon \cdot \partial_t \varphi + F^\varepsilon \cdot \partial_x \varphi\} \ge 0 \quad \forall \varphi \in C_0^\infty((0,T) \times [0,L)),\ \varphi \ge 0.$$

We can take the limit $\varepsilon \to 0$ in these integrals and conclude that $F^0 \le 0$ on B from $\partial_t u^0 = 0$.

Inequality $F^0 \ge 0$. This inequality is not a consequence of the boundary conditions but must be concluded with the help of the positivity of the global convection term $f(u)$. We consider an arbitrary rectangle $U \Subset B$ and a number $q > 0$ and show for some m_0 that

$$F^{\varepsilon_m} \ge -q \quad \text{ on } U \text{ for all } m \ge m_0.$$

Once this is shown, we have $F^0 \ge 0$ almost everywhere on B.

We fix the rectangle $U \Subset B$ and the number $q > 0$. We choose $\delta > 0$ small compared to $q \cdot \inf\{x : (x,t) \in U\} > 0$ and refer the reader to the end of the proof for the precise choice. We now select m_0 such that, for all $m \ge m_0$, (i) $x \le X^\varepsilon_{\delta/2}(t)$ for all $(x,t) \in U$, (ii) $u^\varepsilon(x,t) \le \delta$ for all $(x,t) \in U$ with $x \in \Gamma^\varepsilon_-$, (iii) $u^*/2 \le u^\varepsilon(x,t) \le (u^*+1)/2$ for all $(x,t) \in U$ with $x \in \Gamma^\varepsilon_-$. The existence of such an m_0 follows from the BV-convergence $X^\varepsilon_{\delta/2} \to X^0_{\delta/2}$, the argument of (3.18), and the lower bound for averages. We now assume that, for some $(x,t) \in U$ and some $m \ge m_0$,

$$F^{\varepsilon_m}(x,t) = f^{\varepsilon_m}(x, u^{\varepsilon_m}(x,t)) - g^{\varepsilon_m}(x, u^{\varepsilon_m}(x,t))\partial_x u^{\varepsilon_m}(x,t) < -q$$

and derive a contradiction.

Since $\partial_x F^\varepsilon = -\partial_t u^\varepsilon$ is nonnegative, F^ε is monotonically increasing and we have

$$-g^\varepsilon(x, u^{\varepsilon_m}(x',t))\partial_x u^\varepsilon(x',t) = F^\varepsilon(x',t) - f^\varepsilon(x, u^{\varepsilon_m}(x',t)) \le F^\varepsilon(x',t) \le -q$$

for all $x' \in (0,x)$. This implies that $u^\varepsilon(.,t)$ is increasing on $\Gamma^\varepsilon_- \cap (0,x)$, and on $\Gamma^\varepsilon_+ \cap (0,x)$ it is strictly increasing with a lower bound

$$\partial_x u^\varepsilon \ge \frac{q}{g^*} \text{ with } g^* := \sup_{\xi \in (u^*/2, (1+u^*)/2)} g(\xi) > 0.$$

The monotonicity of $p_c{}^\pm$ implies that u^ε is increasing on $(0,x) \cap 2\varepsilon\mathbb{Z}$ with an average slope of at least $q/2g^*$. The boundary condition $u^\varepsilon(0,t) = 0$ leads to

$$u^\varepsilon(x,t) \ge \frac{q}{2g^*} x > \delta$$

if δ was chosen with $\delta < q \inf\{x : (x,t) \in U\}/2g^*$. This is in contradiction with $u^\varepsilon(x,t) \le \delta$ of (ii). The proof of $F^0 = 0$ on B is complete. $\square$

Proof of Theorem 1. A priori estimates for u^ε and F^ε are shown in Lemma 1 of section 2, and we may therefore select weakly convergent subsequences. The weak convergences allow us to take the distributional limit in the conservation law and we find (1.12).

In Lemma 6 we constructed a monotone function $X : [0,T] \to \mathbb{R}$ such that $G = \{(x,t)|x > X(t)\} \subset \Omega_T$ is an open set. In Proposition 2 we derived $u^0 \equiv u^*/2$ and $F^0 = 0$ almost everywhere on $B := \Omega_T \setminus G$; since $\mathcal{F}$ satisfies $\mathcal{F}(u^*/2, \zeta) = 0$ for all $\zeta \in \mathbb{R}$, (1.13) holds pointwise almost everywhere on B. Corollary 1 provides $\partial_x u^0 \in L^2_{loc}(G)$ and (1.13) on G.

We already know that $\partial_x u^0$ is an L^2_{loc}-function on $\Omega_T \setminus \Sigma$ for $\Sigma = \partial G \cap \partial B$. Lemma 3 implies that $\partial_x u^0$ is nonnegative, and the boundedness of u^0 implies that the derivative $\partial_x u^0$ is a nonnegative measure on Ω_T. Each slice $u^0(.,t)$ is a BV-function, and hence the singular part of the measure $\partial_x u^0$ is concentrated on Σ and regular with respect to the one-dimensional Hausdorff measure. Proposition 2 shows that this singular part of the measure $\partial_x u^0$ vanishes. We thus verified the statement $\partial_x u^0 \in L^1(\Omega_T)$. $\square$

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[2] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[3] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of bounded variation and free discontinuity problems*, Oxford Math. Monogr., Clarendon Press, Oxford University Press, New York, 2000.

[4] A. Beliaev, *Homogenization of two-phase flows in porous media with hysteresis in the capillary relation*, European J. Appl. Math., 14 (2003), pp. 61–84.

[5] M. Bertsch, R. Dal Passo, and C. J. van Duijn, *Analysis of oil trapping in porous media flow*, SIAM J. Math. Anal., 35 (2003), pp. 245–267.

[6] A. Bourgeat, S. M. Kozlov, and A. Mikelić, *Effective equations of two-phase flow in random media*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 385–406.

[7] A. Bourgeat, S. Luckhaus, and A. Mikelić, *Convergence of the homogenization process for a double-porosity model of immiscible two-phase flow*, SIAM J. Math. Anal., 27 (1996), pp. 1520–1543.

[8] Z. Chen, *Degenerate two-phase incompressible flow.* I. *Existence, uniqueness and regularity of a weak solution*, J. Differential Equations, 171 (2001), pp. 203–232.

[9] C. J. van Duijn, H. Eichel, R. Helmig, and I. S. Pop, *Effective Equations for Two-Phase Flow in Porous Media: The Effect of Trapping at the Micro Scale*, CASA Report 2005-35, Eindhoven, 2005.

[10] C. J. van Duijn, A. Mikelić, and I. S. Pop, *Effective equations for two-phase flow with trapping on the micro scale*, SIAM J. Appl. Math., 62 (2002), pp. 1531–1568.

[11] B. H. Gilding, *The occurrence of interfaces in nonlinear diffusion-advection processes*, Arch. Ration. Mech. Anal., 100 (1988), pp. 243–263.

[12] D. Kröner and S. Luckhaus, *Flow of oil and water in a porous medium*, J. Differential Equations, 55 (1984), pp. 276–288.

[13] B. H. Kueper and E. O. Frind, *Two-phase flow in heterogeneous porous media:* 1. *Model development*, Water Resources Research, 27 (1991), pp. 1049–1057.

[14] P. van Lingen, *Quantification and Reduction of Capillary Entrapment in Cross-Laminated Oil Reservoirs*, Ph.D. thesis, Delft University of Technology, Delft, 1998.

[15] F. Otto, *L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations*, J. Differential Equations, 131 (1996), pp. 20–38.

[16] B. Schweizer, *Regularization of outflow problems in unsaturated porous media with dry regions*, J. Differential Equations, 237 (2007), pp. 278–306.

[17] B. Schweizer, *Averaging of flows with capillary hysteresis in stochastic porous media*, European J. Appl. Math., 18 (2007), pp. 389–415.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 1764–1787

BOUNDARY HOMOGENIZATION AND REDUCTION OF DIMENSION IN A KIRCHHOFF–LOVE PLATE*

DOMINIQUE BLANCHARD†, ANTONIO GAUDIELLO‡, AND TARAS A. MEL'NYK§

Abstract. We investigate the asymptotic behavior, as ε tends to 0^+, of the transverse displacement of a Kirchhoff–Love plate composed of two domains $\Omega_\varepsilon^+ \cup \Omega_\varepsilon^- \subset \mathbb{R}^2$ depending on ε in the following way. The set Ω_ε^+ is a union of fine teeth, having small cross section of size ε and constant height, ε-periodically distributed on the upper side of a horizontal thin strip with vanishing height h_ε, as ε tends to 0^+. The structure is clamped on the top of the teeth, with a free boundary elsewhere, and subjected to a transverse load. As ε tends to 0^+, we obtain a "continuum" bending model of rods in the limit domain of the comb, while the limit displacement is independent of the vertical variable in the rescaled (with respect to h_ε) strip. We show that the displacement in the strip is equal to the displacement on the base of the teeth if $h_\varepsilon \gg \varepsilon^4$. However, if the strip is thin enough (i.e., $h_\varepsilon \simeq \varepsilon^4$), we show that microscopic oscillations of the displacement in the strip, between the basis of the teeth, may produce a limit average field different from that on the base of the teeth; i.e., a discontinuity in the transmission condition may appear in the limit model.

Key words. Kirchhoff–Love plate, rough boundary, thick junctions, homogenization, dimension reduction

AMS subject classifications. 74K20, 35B27

DOI. 10.1137/070685919

1. Introduction.

Consider a bounded three-dimensional (3D) plate with small thickness and with middle surface $\Omega_\varepsilon^+ \cup \Omega_\varepsilon^- \subset \mathbb{R}^2$ like a comb, where ε is a small positive parameter. Precisely, Ω_ε^+ is a set of fine teeth, with small cross section of size ε and constant height, ε-periodically distributed on the upper side of a horizontal thin strip Ω_ε^- having a small height h_ε (see Figure 1). The structure is clamped on the top of the teeth, with a free boundary elsewhere, and subjected to a transverse load (see Remark 2.3 for other compatible boundary conditions). The transverse displacement is assumed to satisfy the Kirchhoff–Love equation.

The problem under investigation here pertains to the field of stationary problems posed on a domain which has a so-called *rough boundary* or *highly oscillating boundary*. Boundary-value problems involving rough boundaries appear in many fields of physics and engineering sciences such as the scattering of acoustic waves on small periodic obstacles, the free vibrations of strongly nonhomogeneous elastic bodies, and the behavior of fluids over rough walls or of coupled fluid-solid periodic structures. The reader can find a detailed bibliography about this subject in [2], [4], and [16]. More generally, for the study of thin structures and multistructures, we refer the reader to the following monographs: [7], [9], [12], [13], [14], [17], [19], [20], [22], and the references therein.

*Received by the editors March 21, 2007; accepted for publication (in revised form) July 25, 2007; published electronically February 15, 2008.
http://www.siam.org/journals/sima/39-6/68591.html

†Université de Rouen, UMR 6085, F-76821 Mont Saint Aignan Cédex, France; and Laboratoire d'Analyse Numérique, Université P. et M. Curie, Case Courrier 187, 75252 Paris Cédex 05, France (blanchar@ann.jussieu.fr).
‡DAEIMI, Università degli Studi di Cassino, via G. Di Biasio 43, 03043 Cassino (FR), Italy (gaudiell@unina.it).
§Kyiv Nat. Taras Shevchenko Univ., Volodymyrs'ka Str. 64, 01033 Kyiv, Ukraine (melnyk@imath.kiev.ua).

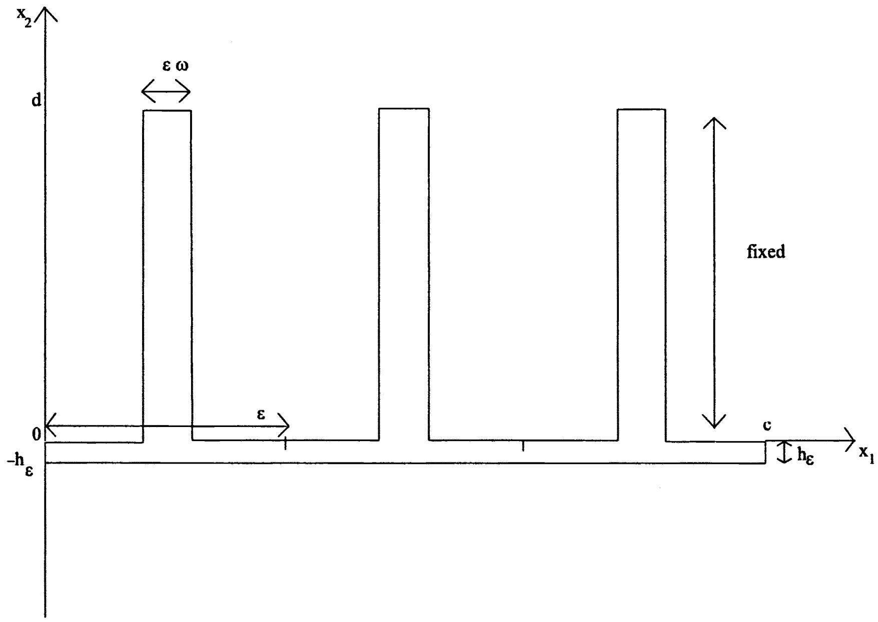


FIG. 1. *The middle surface of our 3D plate.*

A physical example for the specific model and structure considered here appears in MEMS[1] acceleration sensors (see [15]) where the acceleration component in the direction of the motion corresponds to a transverse load applied over the thin strip and the deformation of the teeth is then proportional to the acceleration itself. These sensors are widely used in automotive applications (antilock braking system (ABS), electronic stability program (ESP), airbags, etc.). Our geometrical structure can also model a thin elastic grid used to filter a liquid flow in which case the transverse external forces are due to the action of the fluid on the grid (turn Figure 1 upside down).

It is often impossible to approach these problems directly with numerical methods, because the rough boundary of a comb requires a large number of mesh points in their neighborhood. Thus, the computational cost associated with such a problem grows rapidly when the scale of periodicity gets smaller. Moreover, it can occur that the required discretization step becomes too small for the machine precision. Then, the goal is to construct accurate and numerically implementable asymptotic approximations.

The aim of this paper is to mark a first step in this direction by finding, via an asymptotic analysis as ε tends to zero, a handier limit model which approximates the Kirchhoff–Love equation in the comb. In our case, the asymptotic analysis is still more complicated in consequence of the competition between the parameter of periodicity ε and the parameter of thickness h_ε which vanishes as ε tends to zero.

According to an idea introduced by Ciarlet and Destuynder (see [8]), the terms in the thin strip Ω_ε^- are rescaled in a fixed domain Ω^-. Moreover, in the following, Ω^+ denotes the "limit domain" of the teeth (see Figure 2). In the limit process, under suitable convergence assumptions on the rescaled loads, we obtain, in Ω^+, a

[1]Microelectromechanical systems.

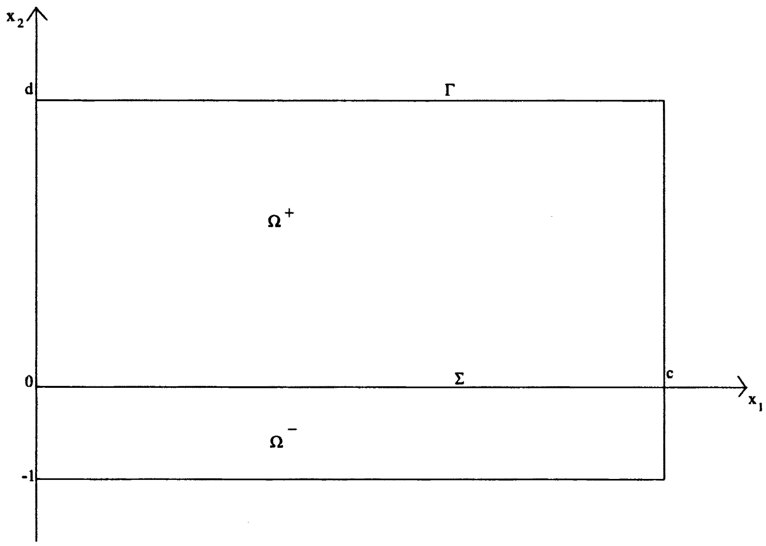


FIG. 2. *The limit middle surface.*

continuum bending model of rods subjected to a force f depending on the limit of the transverse loads on the teeth. The rods are clamped on the upper side Γ and subjected on the lower side Σ to applied forces but without applied momentum. The forces on Σ depend on the limit density g of the transverse loads on the thin strip Ω_ε^- and on the measure of the cross section of the reference tooth.

The limit solution is independent of the vertical variable in Ω^- and meets a Dirichlet transmission condition between Ω^+ and Ω^- if $h_\varepsilon \gg \varepsilon^4$, or if $h_\varepsilon \simeq \varepsilon^4$ and if g is negligible. However, if the strip is thin enough, i.e., $h_\varepsilon \simeq \varepsilon^4$, and the transverse loads on the thin strip are strong enough, a discontinuity in the Dirichlet transmission condition appears. Roughly speaking, this means that microscopic oscillations of the displacement in the strip, between the basis of the teeth of Ω_ε^+, produce a limit average field different from that on the base of the teeth.

Now, we give a short review on problems involving rough boundaries with two small geometrical parameters in competition. As far as we know, these problems have been recently treated in [3], [5], and [6]. Precisely, in [3] and [6] the authors studied the asymptotic behavior of a monotone nonlinear second-order Neumann problem, with growth $p-1$ ($p \in\ ]1,+\infty[$), in a multidomain of $\mathbb{R}^N$ ($N \geq 2$) composed of a "forest" of cylinders with fixed height and small cross section of size ε, distributed with ε-periodicity upon an asymptotically flat part of thickness h_ε. They proved that $h_\varepsilon = \varepsilon^p$ is a critical size for the thickness of the thin domain. In [5], the homogenization process for the junction of a periodic family of elastic rods with a thin elastic plate is studied in the setting of the linearized elasticity. Here different critical sizes appear leading to various limit models: bending-bending model for the rods and the two-dimensional (2D) plate, rigid-bending, or microscopic effects. In [11] the authors considered a thin multidomain of $\mathbb{R}^N$ ($N \geq 2$) consisting (e.g., in a 3D setting) of only one vertical rod upon a horizontal disk. In this thin multidomain they introduced a bulk energy density of the kind $W(D^2U)$, where W is a convex function with growth $p \in\ ]1,+\infty[$. By assuming that the two volumes tend to zero with the same rate, under suitable boundary conditions, they showed that the limit problem is uncoupled if $1 < p \leq \frac{N-1}{2}$, "partially" coupled if $\frac{N-1}{2} < p \leq N-1$, and coupled if $N-1 < p$.

The main difficulty in analyzing the problem considered in the present paper is twofold and it is essentially due to the fourth order of the Kirchhoff–Love model. First, deriving a priori estimates on the displacement for this operator and for an oscillating domain after rescaling is more intricate than for second-order problems. Second, as far as the homogenization process is concerned, the use of the method of oscillating test functions, introduced by Tartar in [21], is also more complicated than for second-order problems since here we have to take into account also the oscillations of the second derivatives of the solution.

Our paper is organized as follows. In section 2 the problem and the main results are stated. In section 3, by making use of the results proved in [3], some a priori norm-estimates are obtained. In section 4, these estimates provide some convergence results in L^2-norm, in the weak topology of L^2, or in the setting of the two-scale convergence method, proposed by Nguetseng in [18] and developed by Allaire in [1]. For convenience of the reader, in this section we also recall the definition and the main properties of the two-scale convergence. Finally, in section 5, the limit problem is derived by making use of the method of oscillating test functions.

2. Statement of the problem and main results. Let $\omega =]a,b[$ with $0 < a < b < 1$, $c, d \in]0,+\infty[$, and let $\{\varepsilon\}$, $\{h_\varepsilon\} \subset]0,1[$ be two sequences converging to zero. For every ε, consider the 3D plate with small thickness $t > 0$ and with middle surface $\Omega_\varepsilon^+ \cup \Omega_\varepsilon^- \subset \mathbb{R}^2$ having the shape of a comb (see Figure 1), where

$$\Omega_\varepsilon^+ = \bigcup_{\{k\in\mathbb{N}:\, \varepsilon b+\varepsilon k<c\}} (\varepsilon\omega + \varepsilon k) \times [0, d[$$

is a set of fine teeth of small cross section $\varepsilon\omega$ and constant height d, ε-periodically distributed on the upper basis of the thin strip:

$$\Omega_\varepsilon^- =]0, c[\times] - h_\varepsilon, 0[,$$

which has a vanishing height h_ε and constant basis. Moreover, denote by Γ_ε the top of the teeth of the middle surface:

$$\Gamma_\varepsilon = \bigcup_{\{k\in\mathbb{N}:\, \varepsilon b+\varepsilon k<c\}} (\varepsilon\omega + \varepsilon k) \times \{d\}.$$

When the plate is clamped on $\Gamma_\varepsilon \times] - \frac{t}{2}, \frac{t}{2}[$, with a free boundary on $(\partial(\Omega_\varepsilon^+ \cup \Omega_\varepsilon^-) \setminus \Gamma_\varepsilon) \times] - \frac{t}{2}, \frac{t}{2}[$, and it is subjected to a transverse load, the Kirchhoff–Love equation satisfied by the transverse displacement U_ε of the middle surface $\Omega_\varepsilon^+ \cup \Omega_\varepsilon^-$ is given by (see pages 205–207 in [10])

$$\text{(2.1)} \quad \begin{cases} \dfrac{Et^3}{12(1-\mu^2)} \Delta^2 U_\varepsilon = F_\varepsilon \text{ in } \Omega_\varepsilon^+ \cup \Omega_\varepsilon^-, \\ U_\varepsilon = \partial_n U_\varepsilon = 0 \text{ on } \Gamma_\varepsilon, \\ \Delta U_\varepsilon + (1-\mu)\left(2n_1 n_2 \partial^2_{x_1 x_2} U_\varepsilon - n_2^2 \partial^2_{x_1} U_\varepsilon - n_1^2 \partial^2_{x_2} U_\varepsilon\right) = 0 \\ \qquad\qquad\qquad\qquad \text{on } \partial(\Omega_\varepsilon^+ \cup \Omega_\varepsilon^-) \setminus \Gamma_\varepsilon, \\ \partial_n \Delta U_\varepsilon + (1-\mu)\partial_\tau \left[n_1 n_2 \left(\partial^2_{x_2} U_\varepsilon - \partial^2_{x_1} U_\varepsilon\right) + (n_1^2 - n_2^2)\partial^2_{x_1 x_2} U_\varepsilon\right] = 0 \\ \qquad\qquad\qquad\qquad \text{on } \partial(\Omega_\varepsilon^+ \cup \Omega_\varepsilon^-) \setminus \Gamma_\varepsilon, \end{cases}$$

where $F_\varepsilon \in L^2(\Omega^+ \cup \Omega_\varepsilon^-)$ represents the transverse load, $\Omega^+ =]0,c[\, \times\,]0,d[$ is the "limit domain" of the comb, $n = (n_1, n_2)$ and τ denote the exterior unit normal and the unit tangent to $\Omega_\varepsilon^+ \cup \Omega_\varepsilon^-$, respectively, $\mu \in]0, \frac{1}{2}[$ is the Poisson ratio, and $E > 0$ is the Young modulus of the plate. In the following, M will denote the flexural rigidity modulus of the plate, i.e.,

$$M = \frac{Et^3}{12(1-\mu^2)}. \tag{2.2}$$

We work on the weak formulation of problem (2.1) (see pages 205–207 in [10]):

(2.3)
$$\begin{cases} U_\varepsilon \in H^2(\Omega_\varepsilon^+ \cup \Omega_\varepsilon^-), \;\; U_\varepsilon = \partial_n U_\varepsilon = 0 \text{ on } \Gamma_\varepsilon, \\[2mm] M \displaystyle\int_{\Omega_\varepsilon^+ \cup \Omega_\varepsilon^-} \Delta U_\varepsilon \Delta V + (1-\mu)\left(2\partial^2_{x_1x_2} U_\varepsilon \, \partial^2_{x_1x_2} V - \partial^2_{x_1} U_\varepsilon \, \partial^2_{x_2} V - \partial^2_{x_2} U_\varepsilon \, \partial^2_{x_1} V\right) dx \\[2mm] \qquad\qquad = \displaystyle\int_{\Omega_\varepsilon^+ \cup \Omega_\varepsilon^-} F_\varepsilon V dx \quad \forall V \in H^2(\Omega_\varepsilon \cup \Omega_\varepsilon^-) : V = \partial_n V = 0 \text{ on } \Gamma_\varepsilon. \end{cases}$$

The goal of our paper is to study the asymptotic behavior of problem (2.3) as ε tends to zero. To this aim, by following an idea of Ciarlet and Destuynder (see [8]), problem (2.3) can be reformulated on a domain independent of h_ε by using the maps

$$(x_1, x_2) \in \Omega^- =]0,c[\, \times\,]-1,0[\longrightarrow (x_1, h_\varepsilon x_2) \in \Omega_\varepsilon^-.$$

Namely, by setting

$$\begin{cases} f_\varepsilon(x) = F_\varepsilon(x) & \text{a.e. } x \in \Omega^+, \\ f_\varepsilon(x) = F_\varepsilon(x_1, h_\varepsilon x_2) & \text{a.e. } x \in \Omega^-, \end{cases} \qquad \begin{cases} u_\varepsilon(x) = U_\varepsilon(x) & \text{a.e. } x \in \Omega_\varepsilon^+, \\ u_\varepsilon(x) = U_\varepsilon(x_1, h_\varepsilon x_2) & \text{a.e. } x \in \Omega^-, \end{cases} \tag{2.4}$$

and $\Omega_\varepsilon = \Omega_\varepsilon^+ \cup \Omega^-$, it turns out that u_ε belongs to the following space:

$$\begin{aligned} V_\varepsilon = \big\{ & v \in H^1(\Omega_\varepsilon) \; : \; v^+ \in H^2(\Omega_\varepsilon^+), \; v^- \in H^2(\Omega^-), \\ & v = 0, \; Dv = 0 \text{ on } \Gamma_\varepsilon, \; \partial_{x_1} v^+ = \partial_{x_1} v^- \text{ on } \Sigma \setminus \partial\Omega_\varepsilon, \\ & h_\varepsilon \partial_{x_2} v^+ = \partial_{x_2} v^- \text{ on } \Sigma \setminus \partial\Omega_\varepsilon \big\}, \end{aligned}$$

where $v^+ = v_{|_{\Omega_\varepsilon^+}}$, $v^- = v_{|_{\Omega^-}}$, $\Sigma =]0,c[\, \times \{0\}$ (note that, since $v = 0$ on Γ_ε, the boundary condition $\partial_n v = 0$ on Γ_ε is equivalent to $Dv = 0$ on Γ_ε). Moreover, u_ε is the unique solution of the following problem:

(2.5)
$$\begin{cases} u_\varepsilon \in V_\varepsilon, \\ M \displaystyle\int_{\Omega_\varepsilon^+} \Delta u_\varepsilon \Delta v + (1-\mu)\left(2\partial^2_{x_1x_2} u_\varepsilon\, \partial^2_{x_1x_2} v - \partial^2_{x_1} u_\varepsilon\, \partial^2_{x_2} v - \partial^2_{x_2} u_\varepsilon\, \partial^2_{x_1} v\right) dx \\ +Mh_\varepsilon \displaystyle\int_{\Omega^-} \left(\partial^2_{x_1} u_\varepsilon + \frac{1}{h_\varepsilon^2}\partial^2_{x_2} u_\varepsilon\right)\left(\partial^2_{x_1} v + \frac{1}{h_\varepsilon^2}\partial^2_{x_2} v\right) dx \\ +M(1-\mu)h_\varepsilon \displaystyle\int_{\Omega^-} 2\frac{1}{h_\varepsilon}\partial^2_{x_1x_2} u_\varepsilon\, \frac{1}{h_\varepsilon}\partial^2_{x_1x_2} v - \partial^2_{x_1} u_\varepsilon \frac{1}{h_\varepsilon^2}\partial^2_{x_2} v - \frac{1}{h_\varepsilon^2}\partial^2_{x_2} u_\varepsilon\, \partial^2_{x_1} v\, dx \\ = \displaystyle\int_{\Omega_\varepsilon^+} f_\varepsilon v dx + h_\varepsilon \int_{\Omega^-} f_\varepsilon v dx \quad \forall v \in V_\varepsilon. \end{cases}$$

The study of the asymptotic behavior of problem (2.5) will be performed under the following assumption:

$$\begin{cases} f_{\varepsilon|_{\Omega^+}} \to f \text{ strongly in } L^2(\Omega^+), \\ h_\varepsilon f_{\varepsilon|_{\Omega^-}} \to g \text{ strongly in } L^2(\Omega^-), \end{cases} \tag{2.6}$$

as ε tends to zero. Moreover, the following spaces will be involved:

(2.7)
$$W^2(\Omega^+) = \left\{ v \in L^2(\Omega^+) \ :\ \partial_{x_2} v \in L^2(\Omega^+),\ \partial^2_{x_2} v \in L^2(\Omega^+),\ v = \partial_{x_2} v = 0 \text{ on } \Gamma \right\},$$

where $\Gamma =]0, c[\, \times \{d\}$ and

$$H^2_{per}(]0,1[) = \left\{ v \in H^2(]0,1[) \ :\ v(0) = v(1),\ v'(0) = v'(1) \right\}$$

with v' denoting the first derivative of v. Note that $H^2_{per}(]0,1[)$ is the closure of $C^\infty_{per}([0,1])$ with respect to the $H^2(]0,1[)$-norm, where $C^\infty_{per}([0,1])$ is the set of functions in $C^\infty(\mathbb{R})$ which are 1-periodic.

In the following, $\widetilde{v}$ denotes the zero-extension to Ω^+ of any function v defined in a subset of Ω^+, and

$$|\omega| = b - a. \tag{2.8}$$

We will show that the limit problem depends on

$$\lim_{\varepsilon \to 0} \frac{\varepsilon^4}{h_\varepsilon} = l \in [0, +\infty[\tag{2.9}$$

and $\int_{-1}^0 g(x_1, x_2) dx_2$. Precisely, the following main result will be proved.

Theorem 2.1. *Let u_ε be the unique solution of problem* (2.5). *Let $W^2(\Omega^+)$ be the space defined in* (2.7). *Assume* (2.6) *and* (2.9). *Then,*

$$\begin{gathered} \widetilde{u_\varepsilon} \rightharpoonup |\omega| u \textit{ weakly in } W^2(\Omega^+), \\ \widetilde{\partial^2_{x_1} u_\varepsilon} \rightharpoonup -\mu|\omega|\partial^2_{x_2} u \textit{ weakly in } L^2(\Omega^+), \\ \widetilde{\partial^2_{x_1x_2} u_\varepsilon} \rightharpoonup 0 \textit{ weakly in } L^2(\Omega^+), \end{gathered}$$

as $\varepsilon \to 0$, where u is the unique solution of the following problem:

$$
(2.10) \qquad \begin{cases} u \in W^2(\Omega^+), \\ |\omega| \dfrac{Et^3}{12} \displaystyle\int_{\Omega^+} \partial^2_{x_2} u \, \partial^2_{x_2} v \, dx = |\omega| \int_{\Omega^+} f v \, dx \\ + \displaystyle\int_0^c \left(\int_{-1}^0 g(x_1, x_2) \, dx_2 \right) v(x_1, 0) \, dx_1 \quad \forall v \in W^2(\Omega^+) \end{cases}
$$

with $|\omega|$ defined in (2.8)*, $\mu \in]0, \frac{1}{2}[$ the Poisson ratio, $E > 0$ the Young modulus, t denoting the small thickness of the* 3*D plate (see problem* (2.1)*), and f and g given by* (2.6)*. Moreover,*

$$
\|\partial_{x_2} u_\varepsilon\|_{L^2(\Omega^-)} \leq C h_\varepsilon^{\frac{3}{4}},
$$

$$
\left\|\partial^2_{x_1 x_2} u_\varepsilon\right\|_{L^2(\Omega^-)} \leq C h_\varepsilon^{\frac{1}{2}}, \quad \left\|\partial^2_{x_2} u_\varepsilon\right\|_{L^2(\Omega^-)} \leq C h_\varepsilon^{\frac{3}{2}}
$$

for every ε, where C is a positive constant independent of ε, and

$$
(2.11) \qquad u_\varepsilon \rightharpoonup u_{|\Sigma} + \int_0^1 v_0 \, dy_1 \ \textit{weakly in } L^2(\Omega^-),
$$

as $\varepsilon \to 0$, where $v_0 = 0$ if $l = 0$ in (2.9)*, while if $l \in]0, +\infty[$, $v_0 \, (= v_0(x_1, y_1))$ is the unique solution of the following problem:*

$$
(2.12) \qquad \begin{cases} v_0 \in L^2(]0, c[, H^2_{per}(]0, 1[)), \\ v_0(x_1, y_1) = 0 \ \textit{in }]0, c[\, \times \, \omega, \\ \dfrac{Et^3}{12} \dfrac{1}{l} \displaystyle\int_{]0,c[\, \times \,]0,1[} \partial^2_{y_1} v_0(x_1, y_1) \partial^2_{y_1} \varphi(x_1, y_1) dx_1 dy_1 \\ = \displaystyle\int_{]0,c[\, \times \,]0,1[} \left(\int_{-1}^0 g(x_1, x_2) dx_2 \right) \varphi\,(x_1, y_1) \, dx_1 dy_1 \\ \qquad \forall \varphi \in L^2(]0, c[, H^2_{per}(]0, 1[)) : \varphi\,(x_1, y_1) = 0 \ \textit{in }]0, c[\, \times \, s\omega \end{cases}
$$

with $u_{|\Sigma}$ denoting the function in $L^2(\Omega^-)$ independent of x_2 and equal, on Σ, to the trace of the solution u of (2.10)*. Furthermore, the convergence of the energies holds:*

$$
\begin{aligned}
\lim_{\varepsilon \to 0} \Bigg\{ & \frac{Et^3}{12(1-\mu^2)} \int_{\Omega^+_\varepsilon} |\Delta u_\varepsilon|^2 + 2(1-\mu) \left(|\partial^2_{x_1 x_2} u_\varepsilon|^2 - \partial^2_{x_1} u_\varepsilon \, \partial^2_{x_2} u_\varepsilon \right) dx \\
& + \frac{Et^3}{12(1-\mu^2)} h_\varepsilon \int_{\Omega^-} \Bigg[\left| \partial^2_{x_1} u_\varepsilon + \frac{1}{h^2_\varepsilon} \partial^2_{x_2} u_\varepsilon \right|^2 \\
& + 2(1-\mu) \left(\left| \frac{1}{h_\varepsilon} \partial^2_{x_1 x_2} u_\varepsilon \right|^2 - \frac{1}{h^2_\varepsilon} \partial^2_{x_2} u_\varepsilon \, \partial^2_{x_1} u_\varepsilon \right) \Bigg] dx \Bigg\} \\
& = \frac{Et^3}{12} \left(|\omega| \int_{\Omega^+} |\partial^2_{x_2} u|^2 dx + \frac{1}{l} \int_{]0,c[\, \times \,]0,1[} \left| \partial^2_{y_1} v_0(x_1, y_1) \right|^2 dx_1 dy_1 \right),
\end{aligned}
$$

where $\infty \cdot 0$ means 0.

Proof. Theorem 2.1 is an immediate consequence of Proposition 4.3, Corollary 4.4, and Propositions 5.1 and 5.2 (see sections 4 and 5). For the sake of clarity, we detail the proof of (2.11) and (2.12). If $l = 0$ in (2.9), convergence (4.6), Corollary 4.4, and (ii) of Proposition 4.2 provide that

$$u_\varepsilon \rightharpoonup \int_0^1 u_0(\cdot, y_1)\, dy_1 = u_{|\Sigma} \text{ weakly in } L^2(\Omega^-).$$

If $l > 0$, by setting $v_0(x_1, y_1) = u_0(x_1, y_1) - u_{|\Sigma}(x_1, 0)$ in $]0, c[\, \times \omega$, convergence (4.6), equality (4.10), and (ii) of Proposition 4.2 provide that

$$u_\varepsilon \rightharpoonup \int_0^1 u_0(\cdot, y_1)\, dy_1 = u_{|\Sigma} + \int_0^1 v_0(\cdot, y_1) dy_1 \text{ weakly in } L^2(\Omega^-),$$

and, by virtue of (5.9) (since $\partial^2_{y_1} u_0 = \partial^2_{y_1} v_0$), v_0 is the unique solution of the following problem:

$$\begin{cases} v_0 \in L^2(]0, c[, H^2_{per}(]0, 1[)), \\ v_0(x_1, y_1) = 0 \text{ in }]0, c[\, \times \omega, \\ M\dfrac{1-\mu^2}{l} \displaystyle\int_{]0,c[\, \times\,]0,1[} \partial^2_{y_1} v_0(x_1, y_1) \partial^2_{y_1} \varphi(x_1, y_1) dx_1 dy_1 \\ = \displaystyle\int_{]0,c[\, \times\,]0,1[} \left(\int_{-1}^0 g(x_1, x_2) dx_2 \right) \varphi\,(x_1, y_1)\, dx_1 dy_1 \\ \qquad\qquad \forall \varphi \in L^2(]0, c[, H^2_{per}(]0, 1[)) : \varphi\,(x_1, y_1) = 0 \text{ in }]0, c[\, \times \omega, \end{cases}$$

which coincides with problem (2.12) by virtue of definition (2.2). Note that, in problem (2.12), $v_0 = 0$ if $\int_{-1}^0 g(x_1, x_2) dx_2 = 0$ a.e. in $]0, c[$.

The convergences of the energies are obtained by passing to the limit, as ε tends to zero, in (2.5) with $v = u_\varepsilon$ and by making use of assumption (2.6), the convergences of $\{\widetilde{u_\varepsilon}\}_\varepsilon$, and the equation satisfied by u and v_0. □

Remark 2.2. Problems (2.10) and (2.12) are the weak formulation of the following problems:

$$\begin{cases} \dfrac{Et^3}{12} \dfrac{\partial^4 u}{\partial x_2^4} = f \text{ in } \Omega^+, \\ u = \dfrac{\partial u}{\partial x_2} = 0 \text{ on } \Gamma, \\ \dfrac{\partial^2 u}{\partial x_2^2} = 0 \text{ on } \Sigma, \\ \dfrac{\partial^3 u}{\partial x_2^3} = \dfrac{12}{|\omega| E t^3} \displaystyle\int_{-1}^0 g(x_1, x_2)\, dx_2 \text{ on } \Sigma, \end{cases} \tag{2.13}$$

and for a.e. $x_1 \in]0, c[$

$$
(2.14) \qquad \begin{cases}
\dfrac{Et^3}{12}\dfrac{1}{l}\dfrac{\partial^4 v_0}{\partial y_1^4}(x_1, y_1) = \displaystyle\int_{-1}^{0} g(x_1, x_2)dx_2 \text{ for } y_1 \in]0, a[, \\[2ex]
\dfrac{Et^3}{12}\dfrac{1}{l}\dfrac{\partial^4 v_0}{\partial y_1^4}(x_1, y_1) = \displaystyle\int_{-1}^{0} g(x_1, x_2)dx_2 \text{ for } y_1 \in]b, 1[, \\[2ex]
v_0(x_1, a) = \dfrac{\partial v_0}{\partial y_1}(x_1, a) = v_0(x_1, b) = \dfrac{\partial v_0}{\partial y_1}(x_1, b) = 0, \\[2ex]
v_0(x_1, 0) = v_0(x_1, 1), \\[2ex]
\dfrac{\partial v_0}{\partial y_1}(x_1, 0) = \dfrac{\partial v_0}{\partial y_1}(x_1, 1), \\[2ex]
\dfrac{\partial^2 v_0}{\partial y_1^2}(x_1, 0) = \dfrac{\partial^2 v_0}{\partial y_1^2}(x_1, 1), \\[2ex]
\dfrac{\partial^3 v_0}{\partial y_1^3}(x_1, 0) = \dfrac{\partial^3 v_0}{\partial y_1^3}(x_1, 1), \\[2ex]
v_0(x_1, y_1) = 0 \text{ for } y_1 \in \omega =]a, b[,
\end{cases}
$$

respectively.

The solution of problem (2.14) can be explicitly computed by solving a linear system of 8 equations with 8 unknowns. Then, for a.e. $x_1 \in]0, c[$, it results that

$$
(2.15) \quad v_0(x_1, y_1) = \begin{cases}
\dfrac{l}{2Et^3}(a - y_1)^2(1 - b + y_1)^2 \displaystyle\int_{-1}^{0} g(x_1, x_2)dx_2 \text{ for } y_1 \in [0, a[, \\[2ex]
0 \text{ for } y_1 \in \omega = [a, b], \\[2ex]
\dfrac{l}{2Et^3}(1 + a - y_1)^2(b - y_1)^2 \displaystyle\int_{-1}^{0} g(x_1, x_2)dx_2 \text{ for } y_1 \in]b, 1],
\end{cases}
$$

and consequently

$$
(2.16) \quad \begin{aligned}
&\int_0^1 v_0(x_1, y_1)\, dy_1 \\
&= \frac{l}{2Et^3}\left\{ \frac{1}{30} + \frac{a}{6} + \frac{a^2}{3} + \frac{a^3}{3} + \frac{a^4}{6} + \frac{a^5}{30} - \frac{b}{6} - \frac{2ab}{3} - a^2 b - \frac{2a^3 b}{3} - \frac{a^4 b}{6} \right. \\
&\quad \left. + \frac{b^2}{3} + ab^2 + a^2 b^2 + \frac{a^3 b^2}{3} - \frac{b^3}{3} - \frac{2ab^3}{3} - \frac{a^2 b^3}{3} + \frac{b^4}{6} + \frac{ab^4}{6} - \frac{b^5}{30} \right\} \int_{-1}^{0} g(x_1, x_2)dx_2.
\end{aligned}
$$

In the limit domain Ω^+ of the comb, we obtain a continuum bending model of rods subjected to a force f, clamped on the upper side Γ, and subjected on the lower side Σ to applied forces but without applied momentum. The forces on Σ depend on the limit density g of the transverse loads on the thin strip Ω_ε^- and on the measure of the cross section ω of the reference tooth. The force f depends on the limit of the transverse loads on the teeth.

The limit solution meets a Dirichlet transmission condition between Ω^+ and the rescaled strip Ω^- if $h_\varepsilon \gg \varepsilon^4$, or if $h_\varepsilon \simeq \varepsilon^4$ and $\int_{-1}^0 g(x_1,x_2)dx_2 = 0$ a.e. in $]0,c[$. However, if the strip is thin enough and the transverse loads on the thin strip are strong enough, i.e., $h_\varepsilon \simeq \varepsilon^4$ and $\int_{-1}^0 g(x_1,x_2)dx_2 \neq 0$ in a subset of $]0,c[$ with positive measure, a discontinuity in the Dirichlet transmission condition appears. Roughly speaking, this means that microscopic oscillations of the displacement in the strip, between the basis of the teeth of Ω_ε^+, produce a limit average field different from that on the base of the teeth. We point out that (2.15) provides that $\int_0^1 v_0(x_1,y_1)\,dy_1 \neq 0$ in a subset of $]0,c[$ with positive measure if and only if $\int_{-1}^0 g(x_1,x_2)dx_2 \neq 0$ in the same subset. Consequently, by taking into account the definition of g in (2.6), for obtaining the additional term in (2.11) when $h_\varepsilon \simeq \varepsilon^4$, it is necessary that the transverse loads in the thin strip Ω_ε^- are strong enough to avoid that $\lim_{\varepsilon\to 0}(h_\varepsilon \int_{\Omega_\varepsilon^-} |F_\varepsilon|^2 dx) = 0$. For instance, if $F_\varepsilon = \varepsilon^{-4\alpha}$ in Ω_ε^-, the additional term in the displacement of the strip intervenes when $\alpha = 1$ and it is given by formula (2.16) with $g = 1$; it does not appear when $\alpha < 1$.

As regards the Laplacian, in [6] the authors proved that $h_\varepsilon \simeq \varepsilon^2$ is a critical size for the thickness of the thin strip. In particular, if $h_\varepsilon \ll \varepsilon^2$, they gave an example in which $g = 0$ and the sequence of the solutions is not even bounded in $L^1(\Omega^-)$. In our paper, as regards the case $h_\varepsilon \ll \varepsilon^4$, we think that a deterministic limit model may hardly be expected, but we have no example to validate it.

Remark 2.3. Let us give a few comments about the boundary condition on $\partial(\Omega_\varepsilon^+ \cup \Omega_\varepsilon^-)$. Indeed, it is physically reasonable to assume that the plate is clamped on a part of its boundary. An alternative to the condition $U_\varepsilon = \partial_n U_\varepsilon = 0$ on Γ_ε could be to impose $U_\varepsilon = \partial_n U_\varepsilon = 0$ on $]0,c[\, \times \{-h_\varepsilon\}$ (i.e., on the lower lateral surface of the 3D plate modeled by the 2D plate Ω_ε^-). A similar and easier analysis than that developed below shows that, in this case, the limit problem (2.13) (also when $\lim_{\varepsilon\to 0}\frac{\varepsilon^4}{h_\varepsilon} = +\infty$) is replaced by the following one:

$$\begin{cases} \dfrac{Et^3}{12}\dfrac{\partial^4 u}{\partial x_2^4} = f \text{ in } \Omega^+, \\[2ex] \dfrac{\partial^2 u}{\partial x_2^2} = \dfrac{\partial^3 u}{\partial x_2^3} = 0 \text{ on } \Gamma, \\[2ex] u = \dfrac{\partial u}{\partial x_2} = 0 \text{ on } \Sigma. \end{cases}$$

Moreover, it results that (see Remark 3.3)

$$u_\varepsilon \to 0 \text{ strongly in } H^1(\Omega^-),$$

and in this case there are no oscillations in the strip.

Concerning the original problem (2.3), the result below immediately follows from Theorem 2.1.

COROLLARY 2.4. *Let U_ε be the solution of problem* (2.3) *under the assumptions of Theorem* 2.1 *with $\{f_\varepsilon\}_\varepsilon$ defined by* (2.4).

Then, it results that

$$
\begin{gathered}
\widetilde{U_\varepsilon} \rightharpoonup |\omega| u \text{ weakly in } W^2(\Omega^+),\\
\widetilde{\partial^2_{x_1} U_\varepsilon} \rightharpoonup -\mu|\omega|\partial^2_{x_2} u \text{ weakly in } L^2(\Omega^+),\\
\widetilde{\partial^2_{x_1x_2} U_\varepsilon} \rightharpoonup 0 \text{ weakly in } L^2(\Omega^+),\\
\lim_{\varepsilon\to 0} \frac{1}{|\Omega^-_\varepsilon|}\int_{\Omega^-_\varepsilon} U_\varepsilon\, dx = \frac{1}{c}\int_0^c \left(u_{|\Sigma} + \int_0^1 v_0 dy_1 \right) dx_1,
\end{gathered}
$$

$$
\begin{aligned}
\lim_{\varepsilon\to 0} \frac{1}{|\Omega^-_\varepsilon|^\alpha}\int_{\Omega^-_\varepsilon} \partial_{x_2} U_\varepsilon dx &= \lim_{\varepsilon\to 0} \frac{1}{|\Omega^-_\varepsilon|^\beta}\int_{\Omega^-_\varepsilon} \partial^2_{x_1x_2} U_\varepsilon dx\\
&= \lim_{\varepsilon\to 0} \frac{1}{|\Omega^-_\varepsilon|^\beta}\int_{\Omega^-_\varepsilon} \partial^2_{x_2} U_\varepsilon dx = 0 \quad \forall\alpha < \frac{3}{4}, \forall\beta < \frac{1}{2},
\end{aligned}
$$

where u is the weak solution of problem (2.13)*, $u_{|\Sigma}$ denotes the trace of u on Σ, and $v_0 = 0$ if $l = 0$ in* (2.9)*, while if $l \in]0,+\infty[$, $v_0\,(= v_0(x_1, y_1))$ is the solution of problem* (2.14)*. Furthermore, the energies converge in the sense that*

$$
\begin{aligned}
&\lim_{\varepsilon\to 0}\left(\frac{Et^3}{12(1-\mu^2)} \int_{\Omega^+_\varepsilon \bigcup \Omega^-_\varepsilon} |\Delta U_\varepsilon|^2 + 2(1-\mu)\left(|\partial^2_{x_1x_2}U_\varepsilon|^2 - \partial^2_{x_1}U_\varepsilon\, \partial^2_{x_2}U_\varepsilon\right) dx \right)\\
&\quad = \frac{Et^3}{12}\left(|\omega| \int_{\Omega^+} |\partial^2_{x_2} u|^2 dx + \frac{1}{l}\int_{]0,c[\ \times\ ]0,1[} \left|\partial^2_{y_1} v_0(x_1,y_1)\right|^2 dx_1 dy_1 \right),
\end{aligned}
$$

where $\infty \cdot 0$ means 0.

3. A priori norm-estimates. Define

$$
D^2(v) = \begin{pmatrix} \partial^2_{x_1} v & \partial^2_{x_1x_2} v \\ \partial^2_{x_1x_2} v & \partial^2_{x_2} v \end{pmatrix}, \qquad v \in H^2(\Omega^+_\varepsilon);
$$

$$
D^2_\varepsilon(v) = \begin{pmatrix} \partial^2_{x_1} v & \dfrac{1}{h_\varepsilon}\partial^2_{x_1x_2} v \\ \dfrac{1}{h_\varepsilon}\partial^2_{x_1x_2} v & \dfrac{1}{h^2_\varepsilon}\partial^2_{x_2} v \end{pmatrix}, \qquad v \in H^2(\Omega^-),
$$

for every ε. This section is devoted to proving the following a priori norm-estimates.

PROPOSITION 3.1. *Let u_ε be the solution of problem* (2.5)*. Assume* (2.9) *and* (2.6)*. Then, there exists a positive constant C, independent of ε, such that*

$$
\|u_\varepsilon\|_{H^2(\Omega^+_\varepsilon)} \le C, \tag{3.1}
$$

$$
\left\| h_\varepsilon^{\frac{1}{2}} D^2_\varepsilon(u_\varepsilon) \right\|_{(L^2(\Omega^-))^4} \le C \tag{3.2}
$$

for every ε.

To prove Proposition 3.1, the following result is required.

LEMMA 3.2. *There exists a positive constant C, independent of ε, such that*

$$\|v\|^2_{L^2(\Omega^-)} \le C\left(\|v\|^2_{L^2(\Sigma\setminus\partial\Omega_\varepsilon)} + \varepsilon^2\|\partial_{x_1}v\|^2_{L^2(\Omega^-)} + \|\partial_{x_2}v\|^2_{L^2(\Omega^-)}\right) \quad \forall v \in H^1(\Omega^-); \tag{3.3}$$

$$\|v\|^2_{L^2(\Sigma\setminus\partial\Omega_\varepsilon)} \le C\left(\|v\|^2_{L^2(\Omega^+_\varepsilon)} + \|\partial_{x_2}v\|^2_{L^2(\Omega^+_\varepsilon)}\right) \quad \forall v \in H^1(\Omega^+_\varepsilon); \tag{3.4}$$

$$\|v\|^2_{H^2(\Omega^+_\varepsilon)} \le C\left\|D^2v\right\|^2_{\left(L^2(\Omega^+_\varepsilon)\right)^4} \quad \forall v \in \left\{v \in H^2(\Omega^+_\varepsilon)\ :\ v=0,\ Dv=0\ on\ \Gamma_\varepsilon\right\} \tag{3.5}$$

for every ε.

Proof. The proof of inequality (3.3) is performed in the proof of Proposition 3.3 in [3]. Standard arguments using Poincaré inequality and Sobolev trace theorem lead to inequalities (3.4) and (3.5). Obviously, boundary conditions $v = 0$ and $Dv = 0$ on Γ_ε intervene for obtaining estimate (3.5). In the case of other boundary conditions, see Remark 3.3. ∎

Proof of Proposition 3.1. In the following, C denotes any positive constant independent of ε.

By choosing $v = u_\varepsilon$ in (2.5), it results that

$$\begin{aligned}
&M\int_{\Omega^+_\varepsilon} \left|\partial^2_{x_1}u_\varepsilon\right|^2 + \left|\partial^2_{x_2}u_\varepsilon\right|^2 + 2\mu\partial^2_{x_1}u_\varepsilon\,\partial^2_{x_2}u_\varepsilon + 2(1-\mu)\left|\partial^2_{x_1x_2}u_\varepsilon\right|^2 dx\\
&+Mh_\varepsilon\int_{\Omega^-} \left|\partial^2_{x_1}u_\varepsilon\right|^2 + \left|\frac{1}{h^2_\varepsilon}\partial^2_{x_2}u_\varepsilon\right|^2 + 2\mu\partial^2_{x_1}u_\varepsilon\,\frac{1}{h^2_\varepsilon}\partial^2_{x_2}u_\varepsilon + 2(1-\mu)\left|\frac{1}{h_\varepsilon}\partial^2_{x_1x_2}u_\varepsilon\right|^2 dx\\
&= \int_{\Omega^+_\varepsilon} f_\varepsilon u_\varepsilon dx + h_\varepsilon\int_{\Omega^-} f_\varepsilon u_\varepsilon dx
\end{aligned}$$

for every ε. Consequently, by taking into account that $-\alpha^2-\beta^2 \le 2\alpha\beta$, for $\alpha,\beta\in\mathbb{R}$, and by making use of assumption (2.6), one obtains that

$$\begin{aligned}
&\int_{\Omega^+_\varepsilon} \left|\partial^2_{x_1}u_\varepsilon\right|^2 + \left|\partial^2_{x_2}u_\varepsilon\right|^2 - \mu\left|\partial^2_{x_1}u_\varepsilon\right|^2 - \mu\left|\partial^2_{x_2}u_\varepsilon\right|^2 + 2(1-\mu)\left|\partial^2_{x_1x_2}u_\varepsilon\right|^2 dx\\
&+h_\varepsilon\int_{\Omega^-} \left|\partial^2_{x_1}u_\varepsilon\right|^2 + \left|\frac{1}{h^2_\varepsilon}\partial^2_{x_2}u_\varepsilon\right|^2 - \mu\left|\partial^2_{x_1}u_\varepsilon\right|^2 - \mu\left|\frac{1}{h^2_\varepsilon}\partial^2_{x_2}u_\varepsilon\right|^2 + 2(1-\mu)\left|\frac{1}{h_\varepsilon}\partial^2_{x_1x_2}u_\varepsilon\right|^2 dx\\
&\le C\left(\|u_\varepsilon\|_{L^2(\Omega^+_\varepsilon)} + \|u_\varepsilon\|_{L^2(\Omega^-)}\right)
\end{aligned}$$

for every ε, that is,

$$\left\|D^2u_\varepsilon\right\|^2_{\left(L^2(\Omega^+_\varepsilon)\right)^4} + h_\varepsilon\left\|D^2_\varepsilon u_\varepsilon\right\|^2_{(L^2(\Omega^-))^4} \le C\left(\|u_\varepsilon\|_{L^2(\Omega^+_\varepsilon)} + \|u_\varepsilon\|_{L^2(\Omega^-)}\right) \tag{3.6}$$

for every ε.

On the other hand, by applying (3.3) three times and by recalling that $\partial_{x_2} u_\varepsilon^- = h_\varepsilon \partial_{x_2} u_\varepsilon^+$ on $\Sigma \setminus \partial\Omega_\varepsilon$, one obtains that

$$
\begin{aligned}
&\|u_\varepsilon\|^2_{L^2(\Omega^-)} \leq C \left(\|u_\varepsilon\|^2_{L^2(\Sigma\setminus\partial\Omega_\varepsilon)} + \varepsilon^2 \|\partial_{x_1} u_\varepsilon\|^2_{L^2(\Omega^-)} + \|\partial_{x_2} u_\varepsilon\|^2_{L^2(\Omega^-)} \right) \\
&\leq C \|u_\varepsilon\|^2_{L^2(\Sigma\setminus\partial\Omega_\varepsilon)} \\
&\quad + C\varepsilon^2 \left(\|\partial_{x_1} u_\varepsilon\|^2_{L^2(\Sigma\setminus\partial\Omega_\varepsilon)} + \varepsilon^2 \left\|\partial^2_{x_1} u_\varepsilon\right\|^2_{L^2(\Omega^-)} + \left\|\partial^2_{x_1x_2} u_\varepsilon\right\|^2_{L^2(\Omega^-)} \right) \\
&\quad + C \left(\left\|\partial_{x_2} u_\varepsilon^-\right\|^2_{L^2(\Sigma\setminus\partial\Omega_\varepsilon)} + \varepsilon^2 \left\|\partial^2_{x_1x_2} u_\varepsilon\right\|^2_{L^2(\Omega^-)} + \left\|\partial^2_{x_2} u_\varepsilon\right\|^2_{L^2(\Omega^-)} \right) \\
&= C \left(\|u_\varepsilon\|^2_{L^2(\Sigma\setminus\partial\Omega_\varepsilon)} + \varepsilon^2 \|\partial_{x_1} u_\varepsilon\|^2_{L^2(\Sigma\setminus\partial\Omega_\varepsilon)} + \left\|h_\varepsilon \partial_{x_2} u_\varepsilon^+\right\|^2_{L^2(\Sigma\setminus\partial\Omega_\varepsilon)} \right) \\
&\quad + C \left(\varepsilon^4 \left\|\partial^2_{x_1} u_\varepsilon\right\|^2_{L^2(\Omega^-)} + \varepsilon^2 \left\|\partial^2_{x_1x_2} u_\varepsilon\right\|^2_{L^2(\Omega^-)} + \left\|\partial^2_{x_2} u_\varepsilon\right\|^2_{L^2(\Omega^-)} \right)
\end{aligned}
$$

for every ε, from which, by taking into account (3.4) and that $\varepsilon < 1$ and $h_\varepsilon < 1$, it follows that

$$
\begin{aligned}
&\|u_\varepsilon\|^2_{L^2(\Omega^-)} \\
&\leq C \left(\|u_\varepsilon\|^2_{H^2(\Omega_\varepsilon^+)} + \varepsilon^4 \left\|\partial^2_{x_1} u_\varepsilon\right\|^2_{L^2(\Omega^-)} + \left\|\partial^2_{x_1x_2} u_\varepsilon\right\|^2_{L^2(\Omega^-)} + \left\|\partial^2_{x_2} u_\varepsilon\right\|^2_{L^2(\Omega^-)} \right) \\
&= C \Bigg[\|u_\varepsilon\|^2_{H^2(\Omega_\varepsilon^+)} \\
&\quad + h_\varepsilon \left(\frac{\varepsilon^4}{h_\varepsilon} \left\|\partial^2_{x_1} u_\varepsilon\right\|^2_{L^2(\Omega^-)} + h_\varepsilon \left\| \frac{1}{h_\varepsilon} \partial^2_{x_1x_2} u_\varepsilon \right\|^2_{L^2(\Omega^-)} + h_\varepsilon^3 \left\| \frac{1}{h_\varepsilon^2} \partial^2_{x_2} u_\varepsilon \right\|^2_{L^2(\Omega^-)} \right) \Bigg] \\
&\leq C \Bigg[\|u_\varepsilon\|^2_{H^2(\Omega_\varepsilon^+)} \\
&\quad + h_\varepsilon \left(\frac{\varepsilon^4}{h_\varepsilon} \left\|\partial^2_{x_1} u_\varepsilon\right\|^2_{L^2(\Omega^-)} + \left\| \frac{1}{h_\varepsilon} \partial^2_{x_1x_2} u_\varepsilon \right\|^2_{L^2(\Omega^-)} + \left\| \frac{1}{h_\varepsilon^2} \partial^2_{x_2} u_\varepsilon \right\|^2_{L^2(\Omega^-)} \right) \Bigg]
\end{aligned} \tag{3.7}
$$

for every ε.

By combining (3.6) with (3.7), making use of (3.5), and assuming that the limit (2.9) is finite, one has that

$$
\|u_\varepsilon\|^2_{H^2(\Omega_\varepsilon^+)} + h_\varepsilon \left\|D^2_\varepsilon u_\varepsilon\right\|^2_{(L^2(\Omega^-))^4} \leq C \left(\|u_\varepsilon\|^2_{H^2(\Omega_\varepsilon^+)} + h_\varepsilon \left\|D^2_\varepsilon u_\varepsilon\right\|^2_{(L^2(\Omega^-))^4} \right)^{\frac{1}{2}}
$$

for every ε, which provides estimates (3.1) and (3.2). $\square$

Remark 3.3. Let us remark that in the case of the alternative boundary conditions mentioned in Remark 2.3, deriving of estimates (3.1) and (3.2) is easier (and they hold true also when $\lim_{\varepsilon\to 0} \frac{\varepsilon^4}{h_\varepsilon} = +\infty$). Indeed, they follow easily from (3.6) by using

$\partial_{x_2} u_\varepsilon^- = h_\varepsilon \partial_{x_2} u_\varepsilon^+$ on $\Sigma \setminus \partial\Omega_\varepsilon$ and the Poincaré inequality in Ω^- (which is licit since one has that $u_\varepsilon = \partial_{x_2} u_\varepsilon = 0$ on $]0, c[\, \times \{-1\}$). We point out that now inequality (3.2) provides that $u_\varepsilon \to 0$ strongly in $H^1(\Omega^-)$.

Corollary 3.4. *Let u_ε be the solution of problem* (2.5). *Assume* (2.9) *and* (2.6). *Then, there exists a positive constant C, independent of ε, such that*

$$\left\|\varepsilon^2 \partial^2_{x_1} u_\varepsilon\right\|_{L^2(\Omega^-)} \leq C, \tag{3.8}$$

$$\left\|\frac{1}{h_\varepsilon^{\frac{1}{2}}} \partial^2_{x_1 x_2} u_\varepsilon\right\|_{L^2(\Omega^-)} \leq C, \tag{3.9}$$

$$\left\|\frac{1}{h_\varepsilon^{\frac{3}{2}}} \partial^2_{x_2} u_\varepsilon\right\|_{L^2(\Omega^-)} \leq C, \tag{3.10}$$

$$\left\|\frac{1}{h_\varepsilon^{\frac{3}{4}}} \partial_{x_2} u_\varepsilon\right\|_{L^2(\Omega^-)} \leq C, \tag{3.11}$$

$$\left\|\varepsilon \partial_{x_1} u_\varepsilon\right\|_{L^2(\Omega^-)} \leq C, \tag{3.12}$$

$$\left\|u_\varepsilon\right\|_{L^2(\Omega^-)} \leq C \tag{3.13}$$

for every ε.

Proof. Estimates (3.8), (3.9), and (3.10) follow immediately from estimate (3.2). By combining estimate (3.2) with inequalities (3.3) and (3.4) and by recalling that $\partial_{x_2} u_\varepsilon^- = h_\varepsilon \partial_{x_2} u_\varepsilon^+$ on $\Sigma \setminus \partial\Omega_\varepsilon$, it is easy to obtain (3.11), (3.12), and (3.13). ∎

4. Convergence results. For convenience of the reader, we recall the definition and the main properties of the two-scale convergence. We refer the reader to [1] and [18] for the proofs.

Definition 4.1. *Let $\mathcal{O} \subset \mathbb{R}^N$ and $Y =]0,1[^N$. A sequence $\{v_\varepsilon\}_\varepsilon \subset L^2(\mathcal{O})$ is said to two-scale converge to a limit $v \in L^2(\mathcal{O} \times Y)$ if, for any function ϕ in $C_0^\infty\left(\mathcal{O}, C^\infty_{per}(Y)\right)$, it results that*

$$\lim_{\varepsilon \to 0} \int_{\mathcal{O}} v_\varepsilon(x) \phi\left(x, \frac{x}{\varepsilon}\right) dx = \int_{\mathcal{O} \times Y} v(x,y)\phi(x,y)\, dxdy.$$

Proposition 4.2. *Let $\mathcal{O} \subset \mathbb{R}^N$ and $Y =]0,1[^N$.*

(i) *Let $\{v_\varepsilon\}_\varepsilon \subset L^2(\mathcal{O})$ be a sequence converging to v strongly in $L^2(\mathcal{O})$. Then, $\{v_\varepsilon\}_\varepsilon$ two-scale converges to the same limit v.*

(ii) *Let $\{v_\varepsilon\}_\varepsilon \subset L^2(\mathcal{O})$ be a bounded sequence, two-scale converging to $v \in L^2(\mathcal{O} \times Y)$. Then, $\{v_\varepsilon\}$ converges to $\int_Y v(\cdot, y) dy$ weakly in $L^2(\mathcal{O})$.*

(iii) *Let $\{v_\varepsilon\}_\varepsilon$ be a bounded sequence in $L^2(\mathcal{O})$. Then, there exist a subsequence of $\{\varepsilon\}$, still denoted by $\{\varepsilon\}$, and a function $v \in L^2(\mathcal{O} \times Y)$ such that $\{v_\varepsilon\}_\varepsilon$ two-scale converges to v.*

(iv) *Let $\{v_\varepsilon\}_\varepsilon \subset W^{1,2}(\mathcal{O})$ be a sequence such that $\{v_\varepsilon\}_\varepsilon$ and $\{\varepsilon D v_\varepsilon\}_\varepsilon$ are bounded in $L^2(\mathcal{O})$ and $(L^2(\mathcal{O}))^N$, respectively. Then, there exist a subsequence of $\{\varepsilon\}$, still* *denoted by* $\{\varepsilon\}$, *and a function* $v \in L^2(\mathcal{O}, W^{1,2}_{per}(Y))$ *such that* $\{v_\varepsilon\}_\varepsilon$ *and* $\{\varepsilon D v_\varepsilon\}_\varepsilon$ *two-scale converge to* v *and* $D_y v$, *respectively.*

The a priori norm-estimates of the solution u_ε of problem (2.5), obtained in section 3, provide the following convergence result.

PROPOSITION 4.3. *Let* u_ε *be the solution of problem* (2.5). *Let* $W^2(\Omega^+)$ *be the space defined in* (2.7). *Assume* (2.6) *and* (2.9). *Then,*

$$\|\partial_{x_2} u_\varepsilon\|_{L^2(\Omega^-)} \leq C h_\varepsilon^{\frac{3}{4}}, \tag{4.1}$$

$$\left\|\partial^2_{x_1 x_2} u_\varepsilon\right\|_{L^2(\Omega^-)} \leq C h_\varepsilon^{\frac{1}{2}}, \quad \left\|\partial^2_{x_2} u_\varepsilon\right\|_{L^2(\Omega^-)} \leq C h_\varepsilon^{\frac{3}{2}} \tag{4.2}$$

for every ε, *where* C *is a positive constant independent of* ε. *Moreover, there exist a subsequence of* $\{\varepsilon\}$, *still denoted by* $\{\varepsilon\}$, $u \in W^2(\Omega^+)$, $\eta, \zeta \in L^2(\Omega^+)$, $u_0(= u_0(x_1, y_1)) \in L^2(]0, c[, H^2_{per}(]0,1[))$, *and* $\xi(= \xi((x_1,x_2),(y_1,y_2))) \in L^2(\Omega^- \times]0,1[^2)$ *such that*

$$\widetilde{u_\varepsilon} \rightharpoonup |\omega| u \text{ weakly in } W^2(\Omega^+), \tag{4.3}$$

$$\widetilde{\partial^2_{x_1} u_\varepsilon} \rightharpoonup \eta \text{ weakly in } L^2(\Omega^+), \tag{4.4}$$

$$\widetilde{\partial^2_{x_1 x_2} u_\varepsilon} \rightharpoonup \zeta \text{ weakly in } L^2(\Omega^+), \tag{4.5}$$

$$\{u_\varepsilon^-\}_\varepsilon \text{ two-scale converges to } u_0, \tag{4.6}$$

$$\{\varepsilon \partial_{x_1} u_\varepsilon^-\}_\varepsilon \text{ two-scale converges to } \partial_{y_1} u_0, \tag{4.7}$$

$$\{\varepsilon^2 \partial^2_{x_1} u_\varepsilon^-\}_\varepsilon \text{ two-scale converges to } \partial^2_{y_1} u_0, \tag{4.8}$$

$$\left\{\frac{1}{h_\varepsilon^{\frac{3}{2}}} \partial^2_{x_2} u_\varepsilon^-\right\}_\varepsilon \text{ two-scale converges to } \xi, \tag{4.9}$$

as $\varepsilon \to 0$, *and*

$$u_0(x_1, y_1) = u_{|\Sigma}(x_1, 0) \text{ in }]0, c[\, \times \omega. \tag{4.10}$$

Proof. Estimates (4.1) and (4.2) follow from estimates (3.11), (3.9), and (3.10). Convergences (4.3), (4.4), and (4.5) are a consequence of estimate (3.1). Estimates (3.13), (3.12), (3.8), (4.1), and (4.2) provide convergences (4.6), (4.7), and (4.8) with $u_0 \in L^2(\Omega^-, H^2_{per}(]0,1[))$. Moreover, u_0 is independent of x_2, too. In fact, it results that

$$0 = \lim_{\varepsilon \to 0} \int_{\Omega^-} \partial_{x_2} u_\varepsilon \varphi\left(x, \frac{x_1}{\varepsilon}\right) dx = - \lim_{\varepsilon \to 0} \int_{\Omega^-} u_\varepsilon \partial_{x_2} \varphi\left(x, \frac{x_1}{\varepsilon}\right) dx$$

$$= - \int_{\Omega^- \times]0,1[} u_0(x, y_1) \partial_{x_2} \varphi(x, y_1)\, dx dy_1 \quad \forall \varphi \in C_0^\infty\left(\Omega^- \times\,]0,1[\right).$$

Convergence (4.9) springs from estimate (3.10).

Finally, statement (4.10) can be obtained by arguing as in the proof of Proposition 6.4 in [6]. More precisely, let χ_ω be the $]0,1[$-periodic extension to $\mathbb{R}$ of the characteristic function of $\omega =]a,b[$ with respect to $]0,1[$. Then, by proceeding as in the proof of (6.7) in [6], i.e., by using mainly (4.1) and (4.6), one obtains that

$$\left\{u_\varepsilon^-(\cdot,0)\chi_\omega\left(\frac{\cdot}{\varepsilon}\right)\right\}_\varepsilon \text{ two-scale converges to } u_0\chi_\omega, \tag{4.11}$$

as $\varepsilon \to 0$. On the other hand, by proceeding as in the proof of (6.13) in [6], i.e., by using mainly (3.1), one proves that

$$\{\widetilde{u_\varepsilon^+}(\cdot,0)\}_\varepsilon \text{ two-scale converges to } u(\cdot,0)\chi_\omega, \tag{4.12}$$

as $\varepsilon \to 0$. Finally, in order to obtain (4.10) it is enough to pass to the two-scale limit in

$$u_\varepsilon^-(x_1,0)\chi_\omega\left(\frac{x_1}{\varepsilon}\right) = \widetilde{u_\varepsilon^+}(x_1,0) \text{ for a.e. } x_1 \in]0,c[,$$

as $\varepsilon \to 0$, and to make use of (4.11) and (4.12). $\square$

If $l = 0$ in (2.9), then u_0 can be completely identified in terms of u.

COROLLARY 4.4. *Let u_ε be the solution of problem* (2.5). *Assume* (2.9)*, with $l = 0$, and* (2.6). *Let $u \in W^2(\Omega^+)$ and $u_0 \in L^2(]0,c[,H^2_{per}(]0,1[))$ be satisfying Proposition* 4.3. *Then,*

$$u_0(x_1,y_1) = u_{|\Sigma}(x_1,0) \textit{ in }]0,c[\times]0,1[. \tag{4.13}$$

Proof. Assumption (2.9) with $l = 0$ and estimate (3.2) ensure that

$$\varepsilon^2\partial^2_{x_1}u_\varepsilon \to 0 \text{ strongly in } L^2(\Omega^-),$$

as $\varepsilon \to 0$. Consequently, by virtue of (4.8), it results that

$$\partial^2_{y_1}u_0 = 0 \text{ in }]0,c[\times]0,1[. \tag{4.14}$$

By combining (4.10) with (4.14), one obtains (4.13). $\square$

5. The limit problem. The following proposition is devoted to identify the limit problem in Ω^+.

PROPOSITION 5.1. *Let u_ε be the solution of problem* (2.5). *Assume* (2.9) *and* (2.6). *Let $u \in W^2(\Omega^+)$ and η, $\zeta \in L^2(\Omega^+)$ satisfy Proposition* 4.3. *Then,*

$$\eta = -\mu|\omega|\partial^2_{x_2}u \textit{ a.e. in } \Omega^+, \tag{5.1}$$

$$\zeta = 0 \textit{ a.e. in } \Omega^+, \tag{5.2}$$

and $u \in W^2(\Omega^+)$ is the unique solution of

$$\begin{aligned} &M|\omega|(1-\mu^2)\int_{\Omega^+}\partial^2_{x_2}u\partial^2_{x_2}v\,dx \\ &= |\omega|\int_{\Omega^+}fv\,dx + \int_0^c\left(\int_{-1}^0 g(x_1,x_2)\,dx_2\right)v(x_1,0)\,dx_1 \quad \forall v \in W^2(\Omega^+), \end{aligned} \tag{5.3}$$

where $\mu \in]0, \frac{1}{2}[$ is the Poisson ratio, $M > 0$ represents the flexural rigidity modulus of the plate (see problem (2.5)), and f, $g \in L^2(\Omega^-)$ are given by (2.6).

Proof. At first, claim (5.1) will be proved. To this aim, choose $v(x_1, x_2) = \varepsilon^2 \psi_1(\frac{x_1}{\varepsilon}) \varphi(x_1, x_2)$ as test function in (2.5), where ψ_1 is the 1-periodic function defined by $\psi_1(y_1) = \frac{1}{2} y_1(y_1 - 1)$ in $[0, 1]$ and $\varphi \in C_0^\infty(\Omega^+)$ (point out that $v \in C^\infty(\overline{\Omega_\varepsilon^+}) \bigcap C_0(\Omega^+) \subset V_\varepsilon$). Then, it results that

$$
\begin{aligned}
&M \int_{\Omega_\varepsilon^+} \Delta u_\varepsilon \left(\varphi + 2\varepsilon \psi_1' \left(\frac{x_1}{\varepsilon}\right) \partial_{x_1} \varphi + \varepsilon^2 \psi_1 \left(\frac{x_1}{\varepsilon}\right) \partial_{x_1}^2 \varphi + \varepsilon^2 \psi_1 \left(\frac{x_1}{\varepsilon}\right) \partial_{x_2}^2 \varphi \right) dx \\
&+M(1-\mu) \int_{\Omega_\varepsilon^+} 2\partial_{x_1 x_2}^2 u_\varepsilon \left(\varepsilon \psi_1' \left(\frac{x_1}{\varepsilon}\right) \partial_{x_2} \varphi + \varepsilon^2 \psi_1 \left(\frac{x_1}{\varepsilon}\right) \partial_{x_1 x_2}^2 \varphi \right) dx \\
&-M(1-\mu) \int_{\Omega_\varepsilon^+} \partial_{x_1}^2 u_\varepsilon \varepsilon^2 \psi_1 \left(\frac{x_1}{\varepsilon}\right) \partial_{x_2}^2 \varphi \, dx \\
&-M(1-\mu) \int_{\Omega_\varepsilon^+} \partial_{x_2}^2 u_\varepsilon \left(\varphi + 2\varepsilon \psi_1' \left(\frac{x_1}{\varepsilon}\right) \partial_{x_1} \varphi + \varepsilon^2 \psi_1 \left(\frac{x_1}{\varepsilon}\right) \partial_{x_1}^2 \varphi \right) dx \\
&= \int_{\Omega_\varepsilon^+} f_\varepsilon \varepsilon^2 \psi_1 \left(\frac{x_1}{\varepsilon}\right) \varphi \, dx
\end{aligned} \tag{5.4}
$$

for every ε. By passing to the limit, as $\varepsilon \to 0$, in (5.4) and by making use of (4.3), (4.4), (4.5), and (2.6), it is easy to see that

$$
\int_{\Omega^+} \eta \varphi + \mu |\omega| \partial_{x_2}^2 u \varphi \, dx = 0 \quad \forall \varphi \in C_0^\infty(\Omega^+),
$$

which provides (5.1).

In the next step, it will be proved that the function $\zeta \in L^2(\Omega^+)$ is independent of x_2. To this aim, choose $v(x_1, x_2) = \varepsilon \psi_2(\frac{x_1}{\varepsilon}) \varphi(x_1, x_2)$ as test function in (2.5), where ψ_2 is the 1-periodic function defined by $\psi_2(y_1) = -y_1 + \frac{1}{2}$ in $[0, 1[$ and $\varphi \in C_0^\infty(\Omega^+)$ (point out that $v \in C^\infty(\overline{\Omega_\varepsilon^+})$ and supp $v \subset \Omega^+$, consequently $v \in V_\varepsilon$). Then, it results that

$$
\begin{aligned}
&M \int_{\Omega_\varepsilon^+} \Delta u_\varepsilon \left(-2\partial_{x_1} \varphi + \varepsilon \psi_2 \left(\frac{x_1}{\varepsilon}\right) \partial_{x_1}^2 \varphi + \varepsilon \psi_2 \left(\frac{x_1}{\varepsilon}\right) \partial_{x_2}^2 \varphi \right) dx \\
&+M(1-\mu) \int_{\Omega_\varepsilon^+} 2\partial_{x_1 x_2}^2 u_\varepsilon \left(-\partial_{x_2} \varphi + \varepsilon \psi_2 \left(\frac{x_1}{\varepsilon}\right) \partial_{x_1 x_2}^2 \varphi \right) dx \\
&-M(1-\mu) \int_{\Omega_\varepsilon^+} \partial_{x_1}^2 u_\varepsilon \varepsilon \psi_2 \left(\frac{x_1}{\varepsilon}\right) \partial_{x_2}^2 \varphi \, dx \\
&-M(1-\mu) \int_{\Omega_\varepsilon^+} \partial_{x_2}^2 u_\varepsilon \left(-2\partial_{x_1} \varphi + \varepsilon \psi_2 \left(\frac{x_1}{\varepsilon}\right) \partial_{x_1}^2 \varphi \right) dx \\
&= \int_{\Omega_\varepsilon^+} f_\varepsilon \varepsilon \psi_2 \left(\frac{x_1}{\varepsilon}\right) \varphi \, dx
\end{aligned} \tag{5.5}
$$

for every ε. By passing to the limit, as $\varepsilon \to 0$, in (5.5) and by making use of (4.3), (4.4), (4.5), (2.6), and (5.1), it is easy to see that

$$\int_{\Omega^+} 2\mu|\omega|\partial^2_{x_2} u \partial_{x_1}\varphi - 2\mu|\omega|\partial^2_{x_2} u \partial_{x_1}\varphi - 2(1-\mu)\zeta\partial_{x_2}\varphi \, dx = 0 \quad \forall \varphi \in C_0^\infty(\Omega^+),$$

that is,

$$\int_{\Omega^+} \zeta \partial_{x_2}\varphi \, dx = 0 \quad \forall \varphi \in C_0^\infty(\Omega^+),$$

which provides that ζ is independent of x_2.

In the third step, claim (5.2) will be proved. To this aim, choose

$$v(x_1, x_2) = \begin{cases} \varepsilon\psi_2\left(\dfrac{x_1}{\varepsilon}\right)\phi(x_2)\varphi(x_1) \text{ in } \Omega^+_\varepsilon, \\ \\ \varepsilon\psi_2\left(\dfrac{x_1}{\varepsilon}\right)\varphi(x_1) \text{ in } \Omega^- \end{cases}$$

as test function in (2.5), where ψ_2 is defined as above, $\phi \in C^\infty([0,d])$ is such that $\phi = 1$ in $[0, \frac{d}{4}]$, $\phi = 0$ in $[\frac{3d}{4}, d]$, and $\varphi \in C_0^\infty(]0,c[)$ (it is evident that $v \in V_\varepsilon$). Then, it results that

$$\begin{aligned}
& M\int_{\Omega^+_\varepsilon} \Delta u_\varepsilon \left(-2\phi\partial_{x_1}\varphi + \varepsilon\psi_2\left(\frac{x_1}{\varepsilon}\right)\phi\partial^2_{x_1}\varphi + \varepsilon\psi_2\left(\frac{x_1}{\varepsilon}\right)\varphi\partial^2_{x_2}\phi\right) dx \\
& + M(1-\mu)\int_{\Omega^+_\varepsilon} 2\partial^2_{x_1x_2} u_\varepsilon \left(-\varphi\partial_{x_2}\phi + \varepsilon\psi_2\left(\frac{x_1}{\varepsilon}\right)\partial_{x_1}\varphi\partial_{x_2}\phi\right) dx \\
& - M(1-\mu)\int_{\Omega^+_\varepsilon} \partial^2_{x_1} u_\varepsilon \varepsilon\psi_2\left(\frac{x_1}{\varepsilon}\right)\varphi\partial^2_{x_2}\phi \, dx \\
& - M(1-\mu)\int_{\Omega^+_\varepsilon} \partial^2_{x_2} u_\varepsilon \left(-2\phi\partial_{x_1}\varphi + \varepsilon\psi_2\left(\frac{x_1}{\varepsilon}\right)\phi\partial^2_{x_1}\varphi\right) dx \\
& + Mh_\varepsilon^{\frac{1}{2}}\int_{\Omega^-} h_\varepsilon^{\frac{1}{2}}\partial^2_{x_1} u_\varepsilon \left(-2\partial_{x_1}\varphi + \varepsilon\psi_2\left(\frac{x_1}{\varepsilon}\right)\partial^2_{x_1}\varphi\right) dx \\
& + Mh_\varepsilon^{\frac{1}{2}}\int_{\Omega^-} \mu\frac{1}{h_\varepsilon^{\frac{3}{2}}}\partial^2_{x_2} u_\varepsilon \left(-2\partial_{x_1}\varphi + \varepsilon\psi_2\left(\frac{x_1}{\varepsilon}\right)\partial^2_{x_1}\varphi\right) dx \\
& = \int_{\Omega^+_\varepsilon} f_\varepsilon\varepsilon\psi_2\left(\frac{x_1}{\varepsilon}\right)\phi\varphi \, dx + \int_{\Omega^-} h_\varepsilon f_\varepsilon\varepsilon\psi_2\left(\frac{x_1}{\varepsilon}\right)\varphi \, dx
\end{aligned} \tag{5.6}$$

for every ε. By passing to the limit, as $\varepsilon \to 0$, in (5.6) and by making use of (4.3), (4.4), (4.5), (5.1), (3.2), and (2.6), it is easy to see that

$$\int_{\Omega^+} \zeta\varphi\partial_{x_2}\phi \, dx = 0 \quad \forall \varphi \in C_0^\infty(]0,c[),$$

from which, by recalling the assumptions on ϕ and that ζ is independent of x_2, it follows that

$$\int_0^c \zeta(x_1)\varphi(x_1) \, dx = 0 \quad \forall \varphi \in C_0^\infty(]0,c[),$$

that is, (5.2).

Now, the limit problem satisfied by u will be identified. To this aim, choose

$$v(x_1,x_2)=\begin{cases}\varphi(x_1,x_2) \text{ in } \Omega_\varepsilon^+,\\ \varphi(x_1,0)+h_\varepsilon x_2\left(\partial_{x_2}\varphi\right)(x_1,0) \text{ in } \Omega^-\end{cases}$$

as test function in (2.5), where $\varphi\in C^\infty(\overline{\Omega^+})$ and $\varphi=0$, $D\varphi=0$ on Γ. Then, it results that

(5.7)

$$\begin{aligned}
&M\int_{\Omega_\varepsilon^+}\Delta u_\varepsilon\Delta\varphi+(1-\mu)\left(2\partial^2_{x_1x_2}u_\varepsilon\,\partial^2_{x_1x_2}\varphi-\partial^2_{x_1}u_\varepsilon\,\partial^2_{x_2}\varphi-\partial^2_{x_2}u_\varepsilon\,\partial^2_{x_1}\varphi\right)dx\\
&+Mh_\varepsilon^{\frac12}\int_{\Omega^-}\left(h_\varepsilon^{\frac12}\partial^2_{x_1}u_\varepsilon+\frac{1}{h_\varepsilon^{\frac32}}\partial^2_{x_2}u_\varepsilon\right)\left(\left(\partial^2_{x_1}\varphi\right)(x_1,0)+h_\varepsilon x_2\left(\partial^3_{x_1^2x_2}\varphi\right)(x_1,0)\right)dx\\
&+M(1-\mu)\int_{\Omega^-}2\partial^2_{x_1x_2}u_\varepsilon\left(\partial^2_{x_2x_1}\varphi\right)(x_1,0)\,dx\\
&-M(1-\mu)h_\varepsilon^{\frac12}\int_{\Omega^-}\frac{1}{h_\varepsilon^{\frac32}}\partial^2_{x_2}u_\varepsilon\left(\left(\partial^2_{x_1}\varphi\right)(x_1,0)+h_\varepsilon x_2\left(\partial^3_{x_1^2x_2}\varphi\right)(x_1,0)\right)\,dx\\
&=\int_{\Omega_\varepsilon^+}f_\varepsilon\varphi dx+h_\varepsilon\int_{\Omega^-}f_\varepsilon\left(\varphi(x_1,0)+h_\varepsilon x_2\left(\partial_{x_2}\varphi\right)(x',0)\right)dx
\end{aligned}$$

for every ε. By passing to the limit, as $\varepsilon\to 0$, in (5.7), making use of (4.3), (4.4), (4.5), (5.1), (5.2), (3.2), and (2.6), and recalling that

$$\chi_{\Omega_\varepsilon^+}\rightharpoonup|\omega| \text{ weakly in } L^2(\Omega^+),$$

it is easy to see that

(5.8)

$$\begin{aligned}
&M|\omega|(1-\mu^2)\int_{\Omega^+}\partial^2_{x_2}u\partial^2_{x_2}\varphi\,dx=|\omega|\int_{\Omega^+}f\varphi\,dx\\
&+\int_0^c\left(\int_{-1}^0 g(x_1,x_2)\,dx_2\right)\varphi(x_1,0)\,dx_1\quad\forall\varphi\in C^\infty(\overline{\Omega^+})\ :\ \varphi=0,\ D\varphi=0 \text{ on } \Gamma,
\end{aligned}$$

which, by density arguments, provides that $u\in W^2(\Omega^+)$ is the unique solution of (5.3). ☐

Now, the limit problem in Ω^- will be identified. Point out that the two-scale convergence method will be explicitly used in Proposition 5.2, while it is not necessary in Proposition 5.1 for identifying the limit problem in Ω^+.

Proposition 5.2. *Let u_ε be the solution of problem* (2.5). *Assume* (2.9), *with*

$l \in]0,+\infty[$, and (2.6). Let $u_0 \in L^2(]0,c[, H^2_{per}(]0,1[))$ satisfy Proposition 4.3. Then,

$$M\frac{1-\mu^2}{l}\int_{]0,c[\,\times\,]0,1[}\partial^2_{y_1}u_0(x_1,y_1)\partial^2_{y_1}\varphi(x_1,y_1)dx_1dy_1$$

(5.9)
$$=\int_{]0,c[\,\times\,]0,1[}\left(\int_{-1}^0 g(x_1,x_2)dx_2\right)\varphi\left(x_1,y_1\right)dx_1dy_1$$

$$\forall\varphi \in L^2(]0,c[, H^2_{per}(]0,1[)) : \varphi\left(x_1,y_1\right) = 0 \text{ in }]0,c[\,\times\,\omega,$$

where $\mu \in]0,\frac{1}{2}[$ is the Poisson ratio, $M > 0$ represents the flexural rigidity modulus of the plate (see problem (2.5)), and $g \in L^2(\Omega^-)$ is given by (2.6).

Proof. In the following, ε takes values in a subsequence satisfying Proposition 4.3.

The proof of (5.9) will be performed in two steps.

At first, it will be proved that

(5.10)
$$\int_{-1}^0\int_0^1 \xi((x_1,x_2),(y_1,y_2))dx_2dy_2 = -\mu l^{-\frac{1}{2}}\partial^2_{y_1}u_0(x_1,y_1) \text{ a.e. in }]0,c[\,\times\,(]0,1[),$$

where $\xi \in L^2(\Omega^- \times]0,1[^2)$ satisfies Proposition 4.3. To this aim, choose

$$v(x_1,x_2) = \begin{cases} 0 & \text{in } \Omega^+_\varepsilon, \\ h_\varepsilon^{\frac{3}{2}}x_2^2\varphi\left(x_1,\dfrac{x_1}{\varepsilon}\right) & \text{in } \Omega^- \end{cases}$$

as test function in (2.5), where $\varphi\ (= \varphi(x_1,y_1)) \in C_0^\infty(]0,c[\,\times\,(]0,1[))$ (point out that $\varphi(x_1,\cdot)$ admits an intrinsic 1-periodic extension on $\mathbb{R}$). Then, it results that

(5.11)
$$\begin{aligned}
&Mh_\varepsilon\int_{\Omega^-}\partial^2_{x_1}u_\varepsilon\partial^2_{x_1}\left(h_\varepsilon^{\frac{3}{2}}x_2^2\varphi\left(x_1,\frac{x_1}{\varepsilon}\right)\right)dx\\
&+Mh_\varepsilon\int_{\Omega^-}\partial^2_{x_1}u_\varepsilon\frac{1}{h_\varepsilon^2}\partial^2_{x_2}\left(h_\varepsilon^{\frac{3}{2}}x_2^2\varphi\left(x_1,\frac{x_1}{\varepsilon}\right)\right)dx\\
&+Mh_\varepsilon\int_{\Omega^-}\frac{1}{h_\varepsilon^2}\partial^2_{x_2}u_\varepsilon\ \partial^2_{x_1}\left(h_\varepsilon^{\frac{3}{2}}x_2^2\varphi\left(x_1,\frac{x_1}{\varepsilon}\right)\right)dx\\
&+Mh_\varepsilon\int_{\Omega^-}\frac{1}{h_\varepsilon^2}\partial^2_{x_2}u_\varepsilon\frac{1}{h_\varepsilon^2}\partial^2_{x_2}\left(h_\varepsilon^{\frac{3}{2}}x_2^2\varphi\left(x_1,\frac{x_1}{\varepsilon}\right)\right)dx\\
&+M(1-\mu)h_\varepsilon\int_{\Omega^-}2\frac{1}{h_\varepsilon}\partial^2_{x_1x_2}u_\varepsilon\ \frac{1}{h_\varepsilon}\partial^2_{x_1x_2}\left(h_\varepsilon^{\frac{3}{2}}x_2^2\varphi\left(x_1,\frac{x_1}{\varepsilon}\right)\right)\\
&-M(1-\mu)h_\varepsilon\int_{\Omega^-}\partial^2_{x_1}u_\varepsilon\frac{1}{h_\varepsilon^2}\partial^2_{x_2}\left(h_\varepsilon^{\frac{3}{2}}x_2^2\varphi\left(x_1,\frac{x_1}{\varepsilon}\right)\right)\\
&-M(1-\mu)h_\varepsilon\int_{\Omega^-}\frac{1}{h_\varepsilon^2}\partial^2_{x_2}u_\varepsilon\,\partial^2_{x_1}\left(h_\varepsilon^{\frac{3}{2}}x_2^2\varphi\left(x_1,\frac{x_1}{\varepsilon}\right)\right)\,dx\\
&=h_\varepsilon\int_{\Omega^-}f_\varepsilon h_\varepsilon^{\frac{3}{2}}x_2^2\varphi\left(x_1,\frac{x_1}{\varepsilon}\right)dx
\end{aligned}$$

for every ε.

Now, pass to the limit, as $\varepsilon \to 0$, in each term of (5.11), after having simplified the second term with the sixth one, as well as the third with the seventh. At first, Note that the assumption $l \neq 0$ in (2.9) gives $\lim_{\varepsilon\to 0} \frac{h_\varepsilon^2}{\varepsilon} = \lim_{\varepsilon\to 0} \frac{h_\varepsilon^2}{\varepsilon^2} = \lim_{\varepsilon\to 0} \frac{h_\varepsilon^{\frac{1}{2}}}{\varepsilon} = 0$ and $\lim_{\varepsilon\to 0} \frac{h_\varepsilon^{\frac{1}{2}}}{\varepsilon^2} = l^{-\frac{1}{2}}$.

From (2.9), with $l \in]0, +\infty[$, and estimate (3.2) it follows that

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\left(h_\varepsilon \int_{\Omega^-} \partial^2_{x_1} u_\varepsilon \partial^2_{x_1}\left(h_\varepsilon^{\frac{3}{2}} x_2^2 \varphi\left(x_1, \frac{x_1}{\varepsilon}\right)\right) dx\right) \\
&= \lim_{\varepsilon\to 0} \int_{\Omega^-} h_\varepsilon^{\frac{1}{2}} \partial^2_{x_1} u_\varepsilon x_2^2 \left(h_\varepsilon^2 \partial^2_{x_1}\varphi + 2\frac{h_\varepsilon^2}{\varepsilon}\partial^2_{x_1 y_1}\varphi + \frac{h_\varepsilon^2}{\varepsilon^2}\partial^2_{y_1}\varphi\right)\left(x_1, \frac{x_1}{\varepsilon}\right) dx = 0.
\end{aligned} \tag{5.12}$$

From (2.9), with $l \in]0, +\infty[$, and (4.8) it follows that

$$\begin{aligned}
&\lim_{\varepsilon\to 0} \int_{\Omega^-} h_\varepsilon \partial^2_{x_1} u_\varepsilon \frac{1}{h_\varepsilon^2}\partial^2_{x_2}\left(h_\varepsilon^{\frac{3}{2}} x_2^2 \varphi\left(x_1, \frac{x_1}{\varepsilon}\right)\right) dx = \lim_{\varepsilon\to 0}\int_{\Omega^-} h_\varepsilon^{\frac{1}{2}} \partial^2_{x_1} u_\varepsilon 2\varphi\left(x_1, \frac{x_1}{\varepsilon}\right) dx \\
&= \lim_{\varepsilon\to 0}\int_{\Omega^-} 2\frac{h_\varepsilon^{\frac{1}{2}}}{\varepsilon^2}\varepsilon^2 \partial^2_{x_1} u_\varepsilon \varphi\left(x_1, \frac{x_1}{\varepsilon}\right) dx = 2l^{-\frac{1}{2}} \int_{]0,c[\ \times\ ]0,1[} \partial^2_{y_1} u_0(x_1, y_1)\varphi(x_1, y_1) dx_1 dy_1.
\end{aligned} \tag{5.13}$$

From (2.9), with $l \in]0, +\infty[$, and the second estimate in (4.2) (which involves $\lim_{\varepsilon\to 0} \|\partial^2_{x_2} u_\varepsilon\|_{L^2(\Omega^-)} = 0$) it follows that

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\int_{\Omega^-} h_\varepsilon \frac{1}{h_\varepsilon^2}\partial^2_{x_2} u_\varepsilon\ \partial^2_{x_1}\left(h_\varepsilon^{\frac{3}{2}} x_2^2 \varphi\left(x_1, \frac{x_1}{\varepsilon}\right)\right) dx \\
&= \lim_{\varepsilon\to 0}\int_{\Omega^-} \partial^2_{x_2} u_\varepsilon x_2^2 \left(h_\varepsilon^{\frac{1}{2}} \partial^2_{x_1}\varphi + 2\frac{h_\varepsilon^{\frac{1}{2}}}{\varepsilon}\partial^2_{x_1 y_1}\varphi + \frac{h_\varepsilon^{\frac{1}{2}}}{\varepsilon^2}\partial^2_{y_1}\varphi\right)\left(x_1, \frac{x_1}{\varepsilon}\right) dx = 0.
\end{aligned} \tag{5.14}$$

From (4.9) it follows that

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\int_{\Omega^-} h_\varepsilon \frac{1}{h_\varepsilon^2}\partial^2_{x_2} u_\varepsilon \frac{1}{h_\varepsilon^2}\partial^2_{x_2}\left(h_\varepsilon^{\frac{3}{2}} x_2^2 \varphi\left(x_1, \frac{x_1}{\varepsilon}\right)\right) dx = \lim_{\varepsilon\to 0}\int_{\Omega^-} \frac{1}{h_\varepsilon^{\frac{3}{2}}}\partial^2_{x_2} u_\varepsilon 2\varphi\left(x_1, \frac{x_1}{\varepsilon}\right) dx \\
&= 2\int_{\Omega^-\times]0,1[^2} \xi((x_1, x_2), (y_1, y_2))\varphi(x_1, y_1)\, d(x_1, x_2) d(y_1, y_2).
\end{aligned} \tag{5.15}$$

From (2.9), with $l \in]0, +\infty[$, and the first estimate in (4.2) it follows that

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\int_{\Omega^-} h_\varepsilon 2\frac{1}{h_\varepsilon}\partial^2_{x_1 x_2} u_\varepsilon\ \frac{1}{h_\varepsilon}\partial^2_{x_1 x_2}\left(h_\varepsilon^{\frac{3}{2}} x_2^2 \varphi\left(x_1, \frac{x_1}{\varepsilon}\right)\right) dx \\
&= \lim_{\varepsilon\to 0}\int_{\Omega^-} 4\partial^2_{x_1 x_2} u_\varepsilon\, x_2 \left(h_\varepsilon^{\frac{1}{2}} \partial_{x_1}\varphi + \frac{h_\varepsilon^{\frac{1}{2}}}{\varepsilon}\partial_{y_1}\varphi\right)\left(x_1, \frac{x_1}{\varepsilon}\right) dx = 0.
\end{aligned} \tag{5.16}$$

From (2.6) it follows that

$$\lim_{\varepsilon\to 0}\int_{\Omega^-} h_\varepsilon f_\varepsilon h_\varepsilon^{\frac{3}{2}} x_2^2 \varphi\left(x_1, \frac{x_1}{\varepsilon}\right) dx = 0. \tag{5.17}$$

Then, by passing to the limit, as $\varepsilon \to 0$, in (5.11) and by making use of (5.12)–(5.17), one obtains that

$$
\begin{aligned}
&2l^{-\frac{1}{2}} \int_{]0,c[\ \times\ ]0,1[} \partial^2_{y_1} u_0(x_1, y_1)\varphi(x_1, y_1) dx_1 dy_1 \\
&+2 \int_{\Omega^- \times]0,1[^2} \xi((x_1, x_2), (y_1, y_2))\varphi\,(x_1, y_1)\, d(x_1, x_2) d(y_1, y_2) \\
&-(1-\mu) 2l^{-\frac{1}{2}} \int_{]0,c[\ \times\ ]0,1[} \partial^2_{y_1} u_0(x_1, y_1)\varphi(x_1, y_1) dx_1 dy_1 = 0 \quad \forall \varphi \in C_0^\infty(]0, c[\ \times (]0, 1[)),
\end{aligned}
$$

that is, (5.10).

Now, to prove (5.9), choose $v(x_1, x_2) = \varphi(x_1, \frac{x_1}{\varepsilon})$ as test function in (2.5), where $\varphi\,(= \varphi\,(x_1, y_1)) \in C^\infty\left([0, c], C^\infty_{per}\,([0, 1])\right)$ such that $\varphi\,(x_1, y_1) = 0$ in $[0, c] \times\ \overline{\omega}$. Then, it results that

(5.18)

$$
\begin{aligned}
&M h_\varepsilon \int_{\Omega^-} \partial^2_{x_1} u_\varepsilon \partial^2_{x_1} \left(\varphi\left(x_1, \frac{x_1}{\varepsilon}\right)\right) dx + M h_\varepsilon \int_{\Omega^-} \frac{1}{h_\varepsilon^2} \partial^2_{x_2} u_\varepsilon\ \partial^2_{x_1} \left(\varphi\left(x_1, \frac{x_1}{\varepsilon}\right)\right) dx \\
&- M(1-\mu) h_\varepsilon \int_{\Omega^-} \frac{1}{h_\varepsilon^2} \partial^2_{x_2} u_\varepsilon\ \partial^2_{x_1} \left(\varphi\left(x_1, \frac{x_1}{\varepsilon}\right)\right)\, dx = h_\varepsilon \int_{\Omega^-} f_\varepsilon \varphi\left(x_1, \frac{x_1}{\varepsilon}\right) dx
\end{aligned}
$$

for every ε.

Pass to the limit, as $\varepsilon \to 0$, in each term of (5.18).

From (2.9), with $l \in]0, +\infty[$, and (4.8) it follows that

$$
\begin{aligned}
&\lim_{\varepsilon \to 0} \int_{\Omega^-} h_\varepsilon \partial^2_{x_1} u_\varepsilon \partial^2_{x_1} \left(\varphi\left(x_1, \frac{x_1}{\varepsilon}\right)\right) dx \\
&= \lim_{\varepsilon \to 0} \int_{\Omega^-} \frac{h_\varepsilon^{\frac{1}{2}}}{\varepsilon^2} \varepsilon^2 \partial^2_{x_1} u_\varepsilon \left(h_\varepsilon^{\frac{1}{2}} \partial^2_{x_1} \varphi + 2 \frac{h_\varepsilon^{\frac{1}{2}}}{\varepsilon} \partial^2_{x_1 y_1} \varphi + \frac{h_\varepsilon^{\frac{1}{2}}}{\varepsilon^2} \partial^2_{y_1} \varphi \right) \left(x_1, \frac{x_1}{\varepsilon}\right) dx \qquad (5.19) \\
&= \frac{1}{l} \int_{]0,c[\ \times\ ]0,1[} \partial^2_{y_1} u_0(x_1, y_1) \partial^2_{y_1} \varphi(x_1, y_1) dx_1 dy_1.
\end{aligned}
$$

From (2.9), with $l \in]0, +\infty[$, (4.9), and (5.10) it follows that

$$
\begin{aligned}
&\lim_{\varepsilon \to 0} \int_{\Omega^-} h_\varepsilon \frac{1}{h_\varepsilon^2} \partial^2_{x_2} u_\varepsilon\ \partial^2_{x_1} \left(\varphi\left(x_1, \frac{x_1}{\varepsilon}\right)\right) dx \\
&= \lim_{\varepsilon \to 0} \int_{\Omega^-} \frac{1}{h_\varepsilon^{\frac{3}{2}}} \partial^2_{x_2} u_\varepsilon \left(h_\varepsilon^{\frac{1}{2}} \partial^2_{x_1} \varphi + 2 \frac{h_\varepsilon^{\frac{1}{2}}}{\varepsilon} \partial^2_{x_1 y_1} \varphi + \frac{h_\varepsilon^{\frac{1}{2}}}{\varepsilon^2} \partial^2_{y_1} \varphi \right) \left(x_1, \frac{x_1}{\varepsilon}\right) dx \\
&= \int_{\Omega^- \times]0,1[^2} \xi((x_1, x_2), (y_1, y_2)) \frac{1}{l^{\frac{1}{2}}} \partial^2_{y_1} \varphi\,(x_1, y_1)\, d(x_1, x_2) d(y_1, y_2) \qquad (5.20) \\
&= -\frac{\mu}{l} \int_{]0,c[\ \times\ ]0,1[} \partial^2_{y_1} u_0(x_1, y_1) \partial^2_{y_1} \varphi(x_1, y_1) dx_1 dy_1.
\end{aligned}
$$

From (2.6) it follows that

$$\lim_{\varepsilon \to 0} \int_{\Omega^-} h_\varepsilon f_\varepsilon \varphi \left(x_1, \frac{x_1}{\varepsilon} \right) dx = \int_{\Omega^- \times]0,1[} g(x_1, x_2) \varphi (x_1, y_1) \, d(x_1, x_2) dy_1. \tag{5.21}$$

Then, by passing to the limit, as $\varepsilon \to 0$, in (5.18) and by making use of (5.19)–(5.21), one obtains that

$$M \frac{1-\mu^2}{l} \int_{]0,c[\ \times \ ]0,1[} \partial^2_{y_1} u_0(x_1, y_1) \partial^2_{y_1} \varphi(x_1, y_1) dx_1 dy_1$$

$$= \int_{]0,c[\ \times \ ]0,1[} \left(\int_{-1}^{0} g(x_1, x_2) dx_2 \right) \varphi (x_1, y_1) \, dx_1 dy_1$$

$$\forall \varphi \in C^\infty \left([0, c], C^\infty_{per} ([0, 1]) \right) \text{ such that } \varphi (x_1, y_1) = 0 \text{ in } [0, c] \times \overline{\omega},$$

which provides (5.9) by density arguments. $\square$

Acknowledgment. The authors wish to thank Professor Giovanni Busatto for the helpful discussions they had on MEMS.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[2] Y. Amirat, O. Bodart, U. De Maio, and A. Gaudiello, *Asymptotic approximation of the Laplace equation in a domain with highly oscillating boundary*, SIAM J. Math. Anal., 35 (2004), pp. 1598–1616.

[3] D. Blanchard and A. Gaudiello, *Homogenization of highly oscillating boundaries and reduction of dimension for a monotone problem*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 449–460.

[4] D. Blanchard, A. Gaudiello, and G. Griso, *Junction of a periodic family of elastic rods with a 3d plate. Part* I, J. Math. Pures Appl. (9), 88 (2007), pp. 1–33.

[5] D. Blanchard, A. Gaudiello, and G. Griso, *Junction of a periodic family of elastic rods with a thin plate. Part* II, J. Math. Pures Appl. (9), 88 (2007), pp. 149–190.

[6] D. Blanchard, A. Gaudiello, and J. Mossino, *Highly oscillating boundaries and reduction of dimension: The critical case*, Anal. Appl. (Singap.), 5 (2007), pp. 137–163.

[7] P. G. Ciarlet, *Plates and Junctions in Elastic Multistructures: An Asymptotic Analysis,* Rech. Math. Appl. 14, Masson, Paris, 1990.

[8] P. G. Ciarlet and P. Destuynder, *A justification of the two-dimensional linear plate model*, J. Mécanique, 18 (1979), pp. 315–344.

[9] D. Cioranescu and J. Saint Jean Paulin, *Homogenization of Reticulated Structures*, Appl. Math. Sci. 139, Springer-Verlag, New York, 1999.

[10] D. Duvaut and J. L. Lions, *Les inéquations en mécanique et en physique*, Dunod, Paris, 1972.

[11] A. Gaudiello and E. Zappale, *Junction in a thin multidomain for a fourth order problem*, Math. Models Methods Appl. Sci., 16 (2006), pp. 1887–1918.

[12] A. G. Kolpakov, *Stressed Composite Structures: Homogenized Models for Thin-Walled Nonhomogeneous Structures with Initial Stresses*, Vol. 12, Springer-Verlag, Berlin, New York, 2004.

[13] V. A. Kozlov, V. G. Ma'zya, and A. B. Movchan, *Asymptotic Analysis of Fields in a Multi-Structure*, Oxford Math. Monogr., The Clarendon Press, Oxford University Press, New York, 1999.

[14] H. Le Dret, *Problèmes variationnels dans les multi-domaines: Modélisation des jonctions et applications*, Rech. Math. Appl. 19, Masson, Paris, 1991.

[15] S. E. Lyshevski, *Mems and Nems: Systems, Devices, and Structures*, CRC Press, Boca Raton, FL, 2002.

[16] T. A. Mel'nyk, *Homogenization of the Poisson equations in a thick periodic junction*, Z. Anal. Anwendungen, 18 (1999), pp. 953–975.

[17] S. A. Nazarov, *Asymptotic Theory of Thin Plates and Rods. Vol.* 1. *Dimension Reduction and Integral Estimates*, Nauchnaya Kniga, Novosibirsk, Russia, 2001.
[18] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.
[19] G. Panasenko, *Asymptotic Analysis of Rod Structures*, Kluwer, Norwell, MA, 2003.
[20] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Lecture Notes in Comput. Phys., Springer-Verlag, New York, 1984.
[21] L. Tartar, *Cours Peccot, Collège de France* (*March* 1977). *Partially written in F. Murat, H-Convergence, Séminaire d'analyse fonctionnelle et numérique de l'Université d'Alger* (1977–78). English translation in Mathematical Modelling of Composite Materials, Progr. Nonlinear Differential Equations Appl., A. Cherkaev and R. V. Kohn, eds., Birkhäuser-Verlag, 1997, pp. 21–44.
[22] L. Trabucho and J. M. Viano, *Mathematical Modelling of Rods*, Handb. Numer. Anal. 4, North–Holland, Amsterdam, 1996.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 1788–1820

ON THE DYNAMICS OF LIQUID-VAPOR PHASE TRANSITION*

KONSTANTINA TRIVISA†

Abstract. We consider a multidimensional model for the dynamics of liquid-vapor phase transitions. In the present context, liquid and vapor are treated as different species with different volume fractions and different molecular weights. The model presented here is a prototype of a "binary fluid mixture" and is formulated by a system that generalizes the Navier–Stokes(–Fourier) equations in Eulerian coordinates. This system takes now a new form due to the choice of rather complex constitutive relations that can accommodate appropriately the physical context. The setting of the problem presented in this work is motivated by physical considerations. The transport fluxes satisfy rather general constitutive laws, the viscosity and heat conductivity depend on the temperature, and the pressure law is a nonlinear function of the temperature depending on the mass density fraction of the vapor (liquid) in the fluid as well as the molecular weights of the individual species. The existence of globally defined weak solutions of the relevant system of partial differential equations that generalizes the Navier–Stokes(–Fourier) equations for compressible fluids is established by using weak convergence methods, and compactness and interpolation arguments in the spirit of Feireisl [*Dynamics of Viscous Compressible Fluids*, Oxford University Press, Oxford, 2004] and Lions [*Mathematical Topics in Fluid Mechanics, Vol.* 2, The Clarendon Press, Oxford University Press, New York, 1998].

Key words. Navier–Stokes system, compressible fluids, liquid-vapor phase transitions, binary fluid mixtures, oscillation and concentration phenomena

AMS subject classifications. 35L65, 76N10, 35B45

DOI. 10.1137/060652397

1. Introduction.

A multidimensional model is presented for the liquid-vapor phase transition formulated by the Navier–Stokes equations in Euler coordinates. In the present context, these equations express the conservation of mass, the balance of momentum and energy, and the balance of species density.

Consider a pure fluid which exhibits liquid-vapor phase changes. In the macroscopic description adopted here $\rho = \rho(t,x)$ denotes the *density* of the *mixture*, $\mathbf{u} = \mathbf{u}(t,x)$ its average *velocity*, and $\theta = \theta(t,x)$ the *temperature*, while $\rho_1 = \rho f_1$ is the density of 1-species (vapor) with $f_1 = f_1(x,t)$ denoting the *mass density fraction* of vapor in the fluid at time $t \in \mathbb{R}$ and at the spatial position $x \in \Omega \subset \mathbb{R}^N,\ N = 3$. These *macroscopic variables* provide a precise characterization of the state of the mixture, which in the present context consists of the species vapor and liquid. The balance of the species density is given by

$$\textit{balance of species density:} \qquad \partial_t(\rho f_1) + \mathrm{div}(\rho f_1 \mathbf{u}) + \mathrm{div}\mathcal{F}_1 = w_1.$$

Here and in what follows,

- $f_i(x,t)$ is the volume fraction of the i species. Vapor and liquid fill up the space, namely,

$$f_1 + f_2 = 1. \tag{1.1}$$

*Received by the editors February 17, 2006; accepted for publication (in revised form) July 30, 2007; published electronically February 15, 2008. This research was supported in part by the National Science Foundation under the Presidential Early Career Award for Scientists and Engineers PECASE DMS 0239063 and an Alfred P. Sloan Foundation Research Fellowship.
http://www.siam.org/journals/sima/39-6/65239.html

†Department of Mathematics, University of Maryland, College Park, MD 20742-4015 (trivisa@math.umd.edu).

- $\rho_i(x,t)$ is the density of the i species

$$\rho_i = \rho f_i, \quad \rho = \rho_1 + \rho_2. \tag{1.2}$$

- w_i is the reaction rate function denoting the mass of the i species produced per unit volume per time unit. In accordance with the *conservation of mass*:

$$\partial_t \rho + \mathrm{div}(\rho \mathbf{u}) = 0,$$

 we have

$$w_1 + w_2 = 0. \tag{1.3}$$

The *momentum conservation equation* can be written as

$$\textit{balance of momentum:} \qquad \partial_t(\rho \mathbf{u}) + \mathrm{div}(\rho \mathbf{u} \otimes \mathbf{u}) + \nabla p = \mathrm{div}\mathbb{S},$$

where ρ is the *total mass density*, $\mathbf{u} \otimes \mathbf{u}$ is the tensor product of *velocity* vectors, p denotes the *pressure*, while $\mathbb{S}$ is the *viscous stress tensor*.

Finally, the *energy conservation equation* reads

$$\textit{balance of energy:} \qquad \partial_t(\rho e) + \mathrm{div}(\rho \mathbf{u} e) + \mathrm{div}\, \mathbf{q} = \mathbb{S} : \nabla u - p\mathrm{div}\mathbf{u}.$$

The physical properties of the mixture are expressed through *constitutive relations*, which specify the relation of the *viscous stress tensor* $\mathbb{S}$, the *heat flux* $\mathbf{q}$, the pressure p, and the internal energy e to the *macroscopic variables*. The hypotheses on the constitutive laws and on the special features of the model are the following.

1.1. Constitutive relations.

- The viscous stress tensor $\mathbb{S}$ is given by *Newton's viscosity formula*

$$\mathbb{S} = \mu(\theta)\left(\nabla \mathbf{u} + \nabla \mathbf{u}^T - \frac{2}{3}\mathrm{div}\mathbf{u}\, \mathbb{I}\right) + \zeta(\theta)\, \mathrm{div}\mathbf{u}\, \mathbb{I}, \tag{1.4}$$

  where the *shear viscosity* μ and the *bulk viscosity* ζ are supposed to be nonnegative and continuously differentiable functions of the absolute temperature.

- The pressure in the portion occupied by the ith component $p_i = p(\rho_i, \theta)$ depends in a crucial way on the volume fraction f_i. In the case where the fluid is in *pure* phase, either (pure) vapor or (pure) liquid, the pressure can be determined quite accurately through experiments. Let us denote by $p_1 = p(\rho_1, \theta)$ and $p_2 = p(\rho_2, \theta)$ the pressure of pure vapor phase and liquid phase, respectively, with ρ_i the density of the ith component and θ the temperature of the mixture (see Figure 1.1). The density at which the pressures of the individual species coincide, namely, $p_1(m, \theta) = p_2(M, \theta) = p_0$, is known as the *Maxwell equilibrium density* and corresponds to $\rho = m(\theta), \rho = M(\theta)$ in Figure 1.1. At this density, both species, liquid and vapor, coexist and are (as we say) *in equilibrium*, and in this case the pressure can be measured in experiments. On the other hand, the region that corresponds to rapid *phase transition,* namely, the area where the material instantly decomposes into liquid, vapor, or their mixtures ($\rho \in (a, b)$ in Figure 1.1), is highly unstable. As a consequence this part of the so-called *van der Waals* pressure curve (the dashed line in Figure 1.1) cannot be measured in experiments and it is

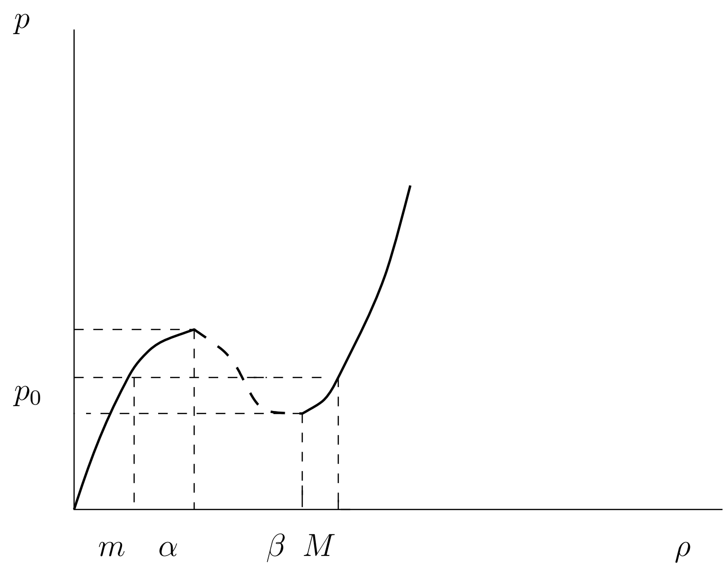


FIG. 1.1. *Pressure.*

regarded as artificial (see also the discussion in Fan [18], Fan and Slemrod [19], Slemrod [29], and the references therein).

As in Fan [18], we proceed considering only the physically relevant part (the part observed in experiments) and we extend each component of the function $p(\rho, \theta)$ continuously as shown in Figure 1.2. This extension may appear at first a bit artificial; however, this part of the pressure does not really affect the outcome, as indicated by physical experiments. In other words, we treat liquid and vapor as different species and specify the transition of the phases by the reaction rate equation for w_1.

Taking the above discussion into consideration, we assume the *Dalton-type law* for the pressure of the mixture, namely,

$$p = p_R + p_B + p_e, \tag{1.5}$$

where the term p_B satisfies *Boyle's law* for the individual species

$$p_B = \frac{\rho_1 R\theta}{m_1} + \frac{\rho_2 R\theta}{m_2}, \tag{1.6}$$

where m_i denotes the molecular weight of the ith component, $p_e = p_e(\rho)$ is the so-called *elastic pressure* whose properties will be discussed in what follows, while p_R accounts for *radiation effects.* We remark that the *elastic pressure* may include higher order terms as in the *Beattie–Bridgman model,* where the state equation for the pressure includes an *elastic* component of the form

$$p_e(\rho) = \beta_1\rho^2 + \beta_2\rho^3 + \beta_3\rho^4$$

for appropriate constants β_i (see [1], [20]).

The point of view presented here takes into consideration the classical as well as the *quantum* aspects associated with the fluid. In the quantum case, the

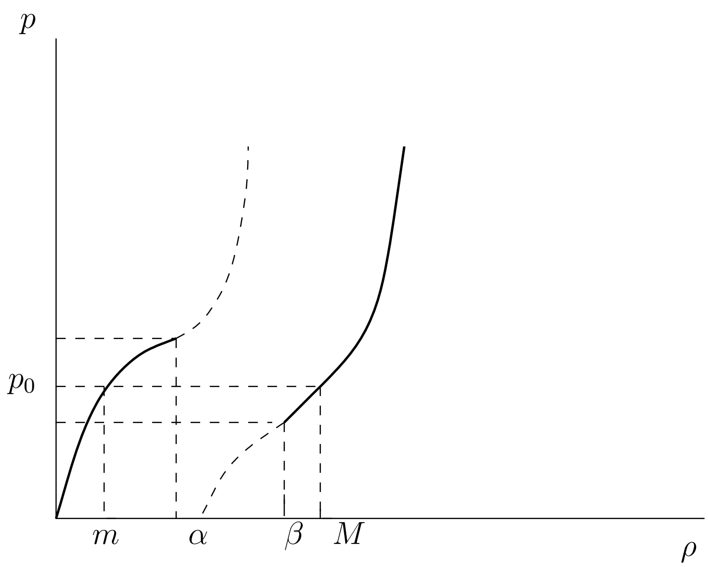


FIG. 1.2. *Modified pressure.*

presence of the photons affects the total pressure p in the fluid. As a result the pressure function now includes an additional (*radiation*) component p_R, which is related to the absolute temperature θ through the *Stefan–Boltzmann law*

$$p_R = \frac{a}{3}\theta^4 \quad \text{with} \quad a > 0 \quad \textit{a constant}.$$

The underlying assumption here (cf. [14], [24], [28]) is that the *high temperature radiation* is at thermal equilibrium with the fluid. Analogously, standard thermodynamic relations require that the specific internal energy of the fluid be also augmented by the term

$$e_R = e_R(\rho, \theta) = \frac{a}{\rho}\theta^4.$$

We remark that *radiation effects* are of particular interest in *astrophysical plasma models* [4].

The pressure therefore satisfies, in the present context, the general law

$$p = \frac{a}{3}\theta^4 + \frac{\rho_1 R\theta}{m_1} + \frac{\rho_2 R\theta}{m_2} + p_e(\rho).$$

Taking into consideration (1.1), (1.2) the pressure law now takes the form

$$p = \frac{a}{3}\theta^4 + \frac{\rho}{m_2}R\theta + R\left(1 - \frac{m_1}{m_2}\right)\frac{\rho_1}{m_1}\theta + p_e(\rho).$$

The molecular weight of the vapor is significantly less than that of the liquid, namely, $m_1 \ll m_2$, which implies that the constant

$$L = \left(1 - \frac{m_1}{m_2}\right)\frac{1}{m_1} > 0, \tag{1.7}$$

and so the pressure law becomes

$$p = \frac{a}{3}\theta^4 + \frac{R}{m_2}\rho\theta + LR\rho_1\theta + p_e(\rho).$$

Motivated by the above discussion, here and in what follows, the pressure law has the form

$$p(\rho, \theta, \rho_1) = p_{radiation} + p_{thermal} + p_{elastic},$$

$$p(\rho, \theta, \rho_1) = \frac{a}{3}\theta^4 + \theta\left[\frac{R}{m_2}p_\theta(\rho) + LRp_\theta(\rho_1)\right] + [p_e(\rho) + p_e(\rho_1)] \tag{1.8}$$

with p_e the *elastic pressure* and p_θ the *thermal pressure* which are functions satisfying certain structural properties to be specified in section 2.3.

- The *internal energy* e satisfies a general constitutive law $e = e(\rho, \theta, \rho_1)$, namely,

$$\rho e = \rho e(\rho, \theta, \rho_1) = a\theta^4 + \frac{c_v}{m_2}\rho\theta + Lc_v\rho_1\theta + \rho P_e(\rho) + \rho_1 P_e(\rho_1) \tag{1.9}$$

with c_v the *specific heat* and

$$P_e(\rho) = \int_1^\rho \frac{p_e(z)}{z^2}\, dz \tag{1.10}$$

denoting the so-called *elastic potential*.

- We treat vapor and liquid as different species, each having its own density ρ_1 and ρ_2, pressure p_1 and p_2, molecular weight m_1 and m_2, but with both components having the same temperature at each point of the mixture.
- The *diffusion flux* $\mathcal{F}_i$ is determined by the law

$$\mathcal{F}_i = -D\rho_i\nabla\log(\rho_i\theta) \tag{1.11}$$

with D denoting the *species diffusion coefficient*, which in accordance with the kinetic theory of gases is a function of the form

$$D(\rho_1) \approx \frac{d}{\rho_1},\ d = \text{constant}.$$

Taking the above into consideration, here and in what follows, the diffusion flux $\mathcal{F}_i$ is of the form

$$\mathcal{F}_i = -d\nabla\log(\rho_i\theta). \tag{1.12}$$

- The *heat flux* $\mathbf{q} = \mathbf{q}(\rho, \rho_i, \theta, \nabla\theta)$ is given by a very general law and consists in the present context of two parts:

$$\mathbf{q} = \mathbf{q_F} + \mathbf{q_{f_i}}. \tag{1.13}$$

The first term $\mathbf{q_F}$ is determined by the *Fourier law,* while the second term $\mathbf{q_{f_i}}$ accounts for the effects of *enthalpy*, which is carried across the surface by the individual species, namely,

$$\begin{cases} \mathbf{q_F} = -\kappa_F(\rho, \theta)\nabla\theta, \\ \mathbf{q_{f_i}} = -\frac{9\gamma-5}{4}Ld\theta\nabla\log(\rho_i\theta), \end{cases} \tag{1.14}$$

where L is the constant introduced in (1.7). The term $\mathbf{q_{f_i}}$ constitutes an additional contribution to $\mathbf{q}$ in *binary* and *multicomponent* systems.

- The function w_1 is the so-called reaction rate function which governs the growth of vapor in the fluid. In general, w_1 consists of two parts, w_{growth} and $w_{nucleation}$. The former accounts for the creation of nuclei of new phase, while the latter describes the subsequent growth of these nuclei,

$$w_1 = w_{growth} + w_{nucleation}.$$

The species production rates w_i are continuous functions of the absolute temperature θ and the species densities ρ_1, ρ_2. For the sake of simplicity, we shall assume that

$$-\underline{c}_1 \le w_1(\theta, \rho_1, \rho_2) \le \bar{c}_1 \quad \text{for all } \theta \ge 0,\ 0 \le \rho_1, \rho_2 \le \rho.$$

Typical examples are presented in [12], [13] in the context of combustion theory, where the rate function is typically given by the *Arrhenius law*, as well as in Fan [18] and Fan and Slemrod [19] in the physical setting of *liquid-vapor* phase transition. In that context, w_1 is the probability of collision between the particles of liquid and vapor multiplied by a parameter known as the *rate parameter*.

Multiplying the *conservation of mass* equation by $(\rho P_e(\rho))'$ we obtain

$$\partial_t(\rho P_e(\rho)) + \mathrm{div}(\rho P_e(\rho)\mathbf{u}) + p_e(\rho)\mathrm{div}\mathbf{u} = 0. \tag{1.15}$$

Analogously,

$$\partial_t(\rho_1 P_e(\rho_1)) + \mathrm{div}(\rho_1 P_e(\rho_1)\mathbf{u}) + p_e(\rho_1)\mathrm{div}\mathbf{u} = (\rho_1 P_e(\rho_1))'\mathrm{div}[\nabla \log(\rho_1\theta)],$$

and so the *conservation of energy equation* yields the *thermal equation*

(1.16)
$$\partial_t\left[a\theta^4 + \frac{c_v}{m_2}\rho\theta + Lc_v\rho_1\theta + \rho_1 P_e(\rho_1)\right] + \mathrm{div}\left[\left(a\theta^4 + \frac{c_v}{m_2}\rho\theta + Lc_v\rho_1\theta + \rho_1 P_e(\rho_1)\right)\mathbf{u}\right]$$
$$+\,\mathrm{div}\mathbf{q} = \mathbb{S} : \nabla\mathbf{u} - \left[\frac{a}{3}\theta^4 + \frac{R}{m_2}\theta p_\theta(\rho) + LR\theta p_\theta(\rho_1) + p_e(\rho_1)\right]\mathrm{div}\mathbf{u} - \sum_{i=1}^{2} h_k w_k$$

with h_k denoting the *formation enthalpies.*

The *internal energy* is related to the *specific entropy* s through a rather general *thermodynamic relation*

$$\theta\mathbf{D}s = \mathbf{D}e + p\mathbf{D}\left(\frac{1}{\rho}\right) - \sum_{i=1}^{2} g_i\mathbf{D}\rho_i, \tag{1.17}$$

where $\mathbf{D}$ denotes the total differential and g_i a quantity that depends on the specific *physical context* and is typically a function of the temperature of the mixture, the formation enthalpies h_k and the formation entropies [25].

If the motion is smooth, starting from the *energy balance*, and in accordance with (1.17), we derive now the *entropy* equation, which now reads

$$\partial_t(\rho s) + \mathrm{div}(\rho s\mathbf{u}) + \mathrm{div}\left[-\frac{k(\theta)}{\theta}\nabla\theta - (\rho s_{f_1})Ld\nabla\log(\rho_1\theta)\right] = \mathbf{r}. \tag{1.18}$$

Here $\mathbf{r}$ denotes the *entropy production*, which is expressed by

$$\mathbf{r} = \frac{1}{\theta}\left(\mathbb{S} : \nabla u + k(\theta)\frac{|\nabla\theta|^2}{\theta} + L\theta d|\nabla \log(\rho_1\theta)|^2 - \sum_{i=1}^{2} g_i w_i\right). \tag{1.19}$$

In accordance with the *second law of thermodynamics* we assume here that

$$\sum_{i=1}^{2} g_i w_i \leq 0,$$

so that the *entropy production rate* (1.19) is nonnegative.

The *specific entropy* s is given by

$$s = s_F + s_{f_1}, \tag{1.20}$$

where

$$\begin{cases} s_F = \frac{4a}{3}\frac{\theta^3}{\rho} + \frac{c_v}{m_2}\log(\theta) - \frac{R}{m_2}P_\theta(\rho), \\ \rho_1 s_{f_1} = Lc_v\rho_1\log(\theta) - LR\rho_1 P_\theta(\rho_1) - LR\rho_1\log(\rho_1) \end{cases} \tag{1.21}$$

with

$$P_\theta(\rho) = \int_1^\rho \frac{p_\theta(z)}{z^2}\,dz.$$

The reader should contrast the form of the entropy in the present context to the form of the entropy in earlier articles (cf. [12], [13], [14], [20]) to see the effect of the presence of the individual components in the mixture.

In the case of a general nonsmooth motion now, and in the spirit of the second law of thermodynamics as given in Truesdell [32], we can only assert that

$$\begin{aligned} &\partial_t[\rho\,(s_F + s_{f_1})] + \operatorname{div}\left[\rho\,(s_F + s_{f_1})\mathbf{u}\right] + \operatorname{div}\left(-\frac{k(\theta)}{\theta}\nabla\theta - (\rho s_{f_1})Ld\nabla\log(\rho_1\theta)\right) \\ &\qquad \geq \frac{1}{\theta}\left(\mathbb{S} : \nabla u + k(\theta)\frac{|\nabla\theta|^2}{\theta} + L\theta d|\nabla\log(\rho_1\theta)|^2 - \sum_{i=1}^{2} g_i w_i\right). \end{aligned} \tag{1.22}$$

The equations which characterize the liquid-vapor *phase transition* now read

$$\partial_t\rho + \operatorname{div}(\rho\mathbf{u}) = 0, \tag{1.23}$$

$$\partial_t(\rho\mathbf{u}) + \operatorname{div}(\rho\mathbf{u}\otimes\mathbf{u}) + \nabla p = \operatorname{div}\mathbb{S}, \tag{1.24}$$

$$\partial_t(\rho s) + \operatorname{div}(\rho s\mathbf{u}) + \operatorname{div}\left[-\frac{k(\theta)}{\theta}\nabla\theta, -(\rho s_{f_1})Ld\nabla\log(\rho_1\theta)\right] = \mathbf{r}, \tag{1.25}$$

$$\partial_t(\rho_1) + \operatorname{div}(\rho_1\mathbf{u}) + \operatorname{div}\mathcal{F}_1 = w_1. \tag{1.26}$$

We assume that the mixture occupies a bounded domain $\Omega \subset \mathbb{R}^N$, $N = 3$, of class $C^{2+\nu}, \nu > 0$, and the whole physical system is both mechanically and thermally isolated. The following boundary conditions hold:

$$\mathbf{u}|_{\partial\Omega} = 0, \quad \mathbf{q}\cdot\mathbf{n}|_{\partial\Omega} = 0, \quad \mathcal{F}\cdot\mathbf{n}|_{\partial\Omega} = 0, \tag{1.27}$$

where $\mathbf{n}$ denotes the outer normal vector to $\partial\Omega$.

In accordance with the above discussion we infer that the *total energy*

$$E = \frac{1}{2}\rho\left(|\mathbf{u}|^2 + e\right) \tag{1.28}$$

is constant of motion, specifically,

$$\frac{d}{dt}\int_\Omega E(t)dx = 0. \tag{1.29}$$

We consider the following initial conditions:

$$\begin{cases} \rho(0,\cdot) = \rho_0, \\ (\rho\,\mathbf{u})(0,\cdot) = \mathbf{m}_0, \\ (\rho\, s)(0,\cdot) = \rho_0 s_0, \\ (\rho_1)(0,\cdot) = \rho_{1_0}, \end{cases} \tag{1.30}$$

together with the compatibility condition

$$\mathbf{m}_0 = 0 \quad \text{on the set} \quad \{x \in \Omega |\ \rho_0(x) = 0\}. \tag{1.31}$$

The objective of this work is to establish the global existence of weak solutions to this initial boundary value problem with large initial data. This article extends earlier work on phase transitions [12], [13] (see also [24]) since it now takes into consideration both the species concentration as well as the unique character of the individual components as given by their distinct molecular weights. The constitutive relations presented here, which differ from the ones given in [12], [13], [24], are able to accommodate both *binary* and *multicomponent* fluid *mixtures*, while at the same time yielding nonnegative *entropy production* in the spirit of the *second law of thermodynamics* as presented in Truesdell [32]. We remark that the complexity of the constitutive relations necessitates the derivation of delicate estimates on the *new* macroscopic variable $\rho_1 = \rho f$. The reader should contrast these estimates to the ones in [12], [13], [24], where the focus was on estimates of the mass fraction density. For results on the one-dimensional setting, we refer the reader to the articles [5], [6], [7]. Related results for multicomponent models are presented in [16], [23], [25], [31].

Moreover, some of the difficulties arising in the execution of our program involve the following issues:

- Obtaining boundness of the *oscillation defect measure* (essential for the *strong* convergence of the density ρ) requires the special treatment of additional terms in the pressure and the equations.
- The presence of extra terms in the entropy production can be handled only by deriving new *energy* and *entropy* estimates.
- The pressure estimates need *refinement* in dealing with *concentration* phenomena.

The methods of use are *weak convergence methods* and *compactness arguments* in the spirit of Feireisl [20] and Lions [27]. At the heart of the analysis lies the quantity

$$p - \left(\frac{4}{3}\mu + \zeta\right)\mathrm{div}\mathbf{u},$$

known as the *effective viscous pressure*. The *weak continuity* property of the *effective viscous pressure* was first shown by Lions [27] for the *barotropic* Navier–Stokes equations, where $p = p(\rho)$ and with constant viscosity coefficients. This result was extended to include the case of general temperature-dependent viscosity coefficients by Feireisl [21] with the aid of delicate commutator estimates in the spirit of Coifman and Meyer [8]. We remark that our analysis is valid for viscosity coefficients depending only on the absolute temperature and are, in the present context, independent of the viscosity. This may be viewed as a weakness of the present theory; it is, however, physically relevant in the context of sufficient diluted fluid mixtures (cf. Becker [1]).

In section 6 we establish the weak continuity of the *effective viscous pressure*, which involves showing that the following relation holds true:

$$\overline{\left(p(\rho,\theta,\rho_1) - \left(\frac{4}{3}\mu(\theta) + \zeta(\theta)\right)\mathrm{div}\mathbf{u}\right)b(\rho)}$$
$$= \left(\overline{p(\rho,\theta,\rho_1)}\,\overline{b(\rho)} - \frac{4}{3}(\overline{\mu(\theta)} + \overline{\zeta(\theta)})\mathrm{div}\mathbf{u}\right)\overline{b(\rho)}$$

for any bounded function b.

In the center of the analysis lies also the requirement that ρ is a *renormalized solution* of the continuity equation, a notion introduced by DiPerna and Lions [10], [11]. The role of this concept is twofold in the sense that it helps us deal both with the problem of *density oscillations* as well as with the problem of temperature *concentration phenomena*.

Another *key* ingredient in evaluating the propagation of density oscillations is in fact showing *boundness* of a suitable *oscillation defect measure*, which is typically expressed in terms of certain cutoff functions (cf. Feireisl [20]). We remark that the choice of the constitutive relations affects in a crucial way the form of this *measure*.

The outline of this article is as follows. In sections 2 and 3 we present the *weak formulation* of the problem and state the main result. Our analysis relies on the concept of *variational solution*, which provides us with the appropriate space setting for the *admissible* solutions of our system. The reader should contrast the notion of *variational solution* presented in this work to earlier notions for relevant phase transition models [14], [12], [13], [24].

In section 4 we introduce a series of approximating problems constituting a three-level approximating scheme. This scheme consists of a set of regularized equations (see also [12], [13], [20]). The regularization appears in terms of additional ε and δ quantities accounting for *artificial viscosity* and *artificial pressure*. In section 5 we obtain uniform bounds (*energy* and *entropy* estimates) necessary as we let the artificial viscosity ε go to zero in section 6. Here, the weak continuity of the *effective viscous pressure* needs to be established in order to obtain suitable estimates on the density component ρ.

In section 7 we recover the original system by letting δ go to zero, getting rid of the *artificial pressure* term $\{\delta\rho^\beta\}$. This process requires the introduction of a suitable family of functions

$$T_k(z) = kT\left(\frac{z}{k}\right) \quad \text{for} \quad z \in \mathbb{R}, \quad k = 1, 2, \dots,$$

where $T \in C^\infty$ is a suitable cutoff function, whose choice depends in a crucial way on the particular physical context, namely, on the choice of constitutive relations for the equation of state for the pressure and other quantities in the system as well as the assumptions on the viscosity coefficients. The main goal here is to show that the so-called *oscillation defect measure*

$$osc_{\beta+1}[\rho_\delta \to \rho]((0,T)\times\Omega) = \sup_{k\geq 1}\left(\limsup_{\delta\to 0}\int_0^T\int_\Omega |T_k(\rho_\delta) - T_k(\rho)|^{\beta+1} dxdt\right)$$

is bounded.

Related results on phase-transition models are presented in [12], [13] in the context of combustion models and in [14], [20], [21], [24] for related models in fluids and astrophysics. We refer the reader also to [23] for further discussion on *multicomponent reacting flows*.

The setting of the present article is motivated by physical considerations presented in a series of articles by Fan [18], Fan and Slemrod [19], and Slemrod [29]. For related work on van der Waals fluids and the issue of nonuniqueness of phase transition near the Maxwell line we refer the reader to Benzoni-Gavage [2] and the references therein, while for general discussion on nucleation phenomena we refer the reader to Springer [30].

2. Weak formulation. Our objective in this article is the solvability of the initial boundary value problem (1.23)–(1.26), together with (1.27) and (1.30), for *large* initial data. To this end, we need to rely on the concept of *variational solution*. Philosophically, this *weak* formulation is connected to the *balance laws* of continuum physics. *Balance laws* are typically expressed in the form of integral identities and therefore no additional requirement on the regularity of the integrand quantities is imposed.

2.1. Dissipation of energy. In the framework of weak solutions, it is common to replace the (formally derived) classical *entropy equality* by an *inequality* (cf. Truesdell [32]). This is due to the possible *loss* of some part of the kinetic energy of the system. This loss of energy appears mathematically in the form of a measure, while in the physical setting, this portion of the energy may be regarded as a new part of the spatial domain. For further remarks on the topic, we refer the reader to Feireisl [20] for relevant discussion in the context of compressible fluids, as well as to Dafermos [9] for relevant discussion in the context of hyperbolic conservation laws.

The *variational formulation* of the entropy production is given by

$$\begin{aligned}(2.1)\qquad &\int_0^T\!\!\int_\Omega \left\{(\rho s)\,\partial_t\phi + (\rho s)\mathbf{u}\cdot\nabla\phi + \left[\frac{-k(\theta)}{\theta}\nabla\theta - (\rho s_{f_1})Ld\nabla\log(\rho_1\theta)\right]\cdot\nabla\phi\right\}dxdt\\ &\leq \int_0^T\!\!\int_\Omega\left\{-\frac{\mathbb{S}:\nabla\mathbf{u}}{\theta} - \frac{k(\theta)|\nabla\theta|^2}{\theta^2} - Ld|\nabla\log(\rho_1\theta)|^2 - \sum_{i=1}^2 g_i w_i\right\}\phi\,dx\,dt\end{aligned}$$

for any nonnegative function $\phi\in\mathcal{D}((0,T)\times\mathbb{R}^N)$.

Note that the fact that θ appears in the denominator in the above relation $\left\{\frac{\mathbb{S}:\nabla\mathbf{u}}{\theta}\right\}$ indicates that the absolute temperature θ must be positive in order for (2.1) to be meaningful.

Motivated by this discussion, we introduce now the notion of a *variational solution* to the initial boundary value problem (1.23)–(1.26) together with (1.27), (1.30), and (1.31).

2.2. The class of admissible solutions. In this section we present the class of admissible solutions for our system (1.23)–(1.26) motivated by the underlying physical principles of continuum physics.

DEFINITION 2.1. *We say that* $(\rho, \mathbf{u}, \theta, \rho_1)$ *is a variational solution of the initial boundary value problem* (1.23)–(1.26) *on the interval* $(0,T)$ *if it satisfies the following properties:*

• *The continuity equation* (1.1) *is satisfied in the sense of renormalized solutions, specifically*

$$\rho \in C([0,T]; L^1(\Omega)) \cap L^\infty(0,T; L^\gamma(\Omega)),\ \rho(0,\cdot) = \rho_0$$

satisfying the continuity equation (1.1) *in the sense of* $\mathcal{D}'((0,T)\times\mathbb{R}^3)$ *provided that* $\rho, \mathbf{u}$ *were extended to be zero outside* Ω*:*

$$\mathbf{u} \in L^2(0,T; W_0^{1,2}(\Omega;\mathbb{R}^3)), \quad \rho\mathbf{u} \in L^\infty(0,T; L^1(\Omega;\mathbb{R}^3)),$$

and the integral identity

$$\int_0^T\int_\Omega (\rho B(\rho)\partial_t\phi + \rho B(\rho)\mathbf{u}\cdot\nabla_x\phi - b(\rho)\mathrm{div}_x\mathbf{u})\ dxdt = -\int_\Omega \rho_0 B(\rho_0)\phi(0,\cdot)dx$$

holds for any test function $\phi \in \mathcal{D}([0,T)\times\bar{\Omega})$ *and any*

$$b \in BC[0,\infty),\ B(\rho) = B(1) + \int_1^\rho \frac{b(z)}{z^2}\ dz. \tag{2.2}$$

• *The density of the individual component* ρ_1 *is a nonnegative measurable function belonging to the space*

$$\rho_1 \in L^\infty(0,T; L^\gamma(\Omega)), \quad \log(\rho_1) \in L^2(0,T; W^{1,2}(\Omega)),$$

while the temperature θ *is a nonnegative function such that*

$$\theta, \log(\theta) \in L^2(0,T; W^{1,2}(\Omega)),$$

and the integral identity

$$\int_0^T\int_\Omega \rho_1\partial_t\phi + \rho_1\mathbf{u}\cdot\nabla_x\phi + \mathcal{F}\cdot\nabla_x\phi\, dxdt = \int_0^T\int_\Omega w_1\phi\, dxdt - \int_\Omega \rho_{1_0}\phi(0,\cdot)dx \tag{2.3}$$

holds for any test function $\phi \in \mathcal{D}'([0,T)\times\bar{\Omega})$.

• *The momentum balance equation* (1.2) *is satisfied in the sense of distributions. Moreover, the pressure* $p \in L^1((0,T)\times\Omega)$ *is related to the state variables* ρ, ρ_1, *and* θ *through the constitutive equation* (1.8), *the viscous stress tensor* $\mathbb{S} \in L^1(0,T; L^1(\Omega;\mathbb{R}^{3\times3}))$ *is given by Newton's law of viscosity* (1.4), *while*

$$\rho\mathbf{u}\otimes\mathbf{u} \in L^1(0,T; L^1(\Omega;\mathbb{R}^{3\times3})).$$

• *The entropy* ρs *is determined by the formula* (1.21). *The density of the individual component* ρ_1 *as well as the absolute temperature are positive a.a. on* $(0,T)\times\Omega$, *and the integral inequality*

$$\begin{aligned}&\int_0^T\int_\Omega \{(\rho s)\,\partial_t\phi + (\rho s)\mathbf{u}\cdot\nabla\phi + \mathbf{q}\cdot\nabla\phi\}\,dxdt\\ &\leq \int_0^T\int_\Omega \left\{-\frac{\mathbb{S}:\nabla\mathbf{u}}{\theta} - \frac{k(\theta)|\nabla\theta|^2}{\theta^2} - Ld|\nabla\log(\rho_1\theta)|^2 - \sum_{i=1}^2 \rho g_i w_i\right\}\phi\,dx\,dt\end{aligned} \tag{2.4}$$

holds for any nonnegative function $\varphi \in \mathcal{D}((0,T)\times\mathbb{R}^N)$.

Moreover,

$$\operatorname*{ess\,lim}_{t\to 0+} \int_\Omega \rho s(t)\,\phi\,dx \geq \int_\Omega \rho_0 s_0\,\phi\,dx \ \ \textit{for any nonnegative } \phi \in \mathcal{D}(\Omega),$$

where

$$\begin{aligned}\rho_0 s_0 = {}& \frac{4a}{3}\theta_0^3 + \frac{c_v}{m_2}\rho_0\log(\theta_0) - \rho_0 P_\theta(\rho_0) + Mc_v\rho_{1_0}\log(\theta_0)\\ & - MR\rho_{1_0}P_\theta(\rho_{1_0}) - MR\rho_{1_0}\log(\rho_{1_0}).\end{aligned}$$

• *As there is no flux of energy through the kinematic boundary we require the* total energy $E(t)$ *to be constant of motion, that is,*

$$\int_0^T\int_\Omega E(t)\,\partial_t\psi\,dx\,dt = 0 \quad \textit{for any} \quad \psi\in\mathcal{D}(0,T). \tag{2.5}$$

• *The functions* $\rho, \rho\,\mathbf{u}, \rho\,\theta$, *and* ρ_1 *satisfy the initial conditions* (1.30) *in the weak sense,*

$$\begin{cases}\operatorname*{ess\,lim}_{t\to 0+}\displaystyle\int_\Omega \rho(t)\,\eta\,dx = \int_\Omega \rho_0\,\eta\,dx,\\ \operatorname*{ess\,lim}_{t\to 0+}\displaystyle\int_\Omega (\rho\,\mathbf{u})(t)\cdot\nu\,dx = \int_\Omega \mathbf{m}_0\cdot\nu\,dx,\\ \operatorname*{ess\,lim}_{t\to 0+}\displaystyle\int_\Omega (\rho\,\theta)(t)\,\eta dx = \int_\Omega \rho_0\,\theta_0\,\eta\,dx,\\ \operatorname*{ess\,lim}_{t\to 0+}\displaystyle\int_\Omega \rho_1(t)\,\eta\,dx = \int_\Omega \rho_{1_0}\,\eta\,dx\end{cases} \tag{2.6}$$

for all $\eta\in\mathcal{D}(\Omega)$.

2.3. Assumptions.

Pressure. The pressure p obeys the general pressure law (1.8), where the *elastic* pressure p_e and the *thermal* pressure p_θ are continuously differentiable functions of the density and in addition satisfy the properties (see also [20], [12], [13])

$$\begin{cases}p_e(0)=0,\\ p_e'(\rho)\geq a_1\rho^{\gamma-1}-c_1, \quad a_1>0,\\ p_e(\rho)\leq a_2\rho^\gamma + c_2,\end{cases} \qquad \begin{cases}p_\theta(0)=0,\\ p_\theta'(\rho)\geq 0,\\ p_\theta(\rho)\leq a_3\rho^G + c_3\end{cases} \tag{2.7}$$

with

$$\gamma\geq 2, \quad \gamma > \frac{4G}{3}.$$

Transport coefficients.

• The *viscosity parameters* depend on the absolute temperature in the following fashion:

$$\begin{cases}0<\underline{\mu}(1+\theta^\alpha)\leq\mu(\theta)\leq\bar{\mu}(1+\theta^\alpha),\\ 0<\underline{\zeta}\theta^\alpha\leq\zeta(\theta)\leq\bar{\zeta}(1+\theta^\alpha)\end{cases} \tag{2.8}$$

for $\alpha\geq\frac{1}{2}$.

• The *heat conductivity* obeys the rule

$$\begin{cases} k = k_C(\theta) + \sigma\theta^3, \\ 0 < \underline{k}_C \le k_C(\theta) \le \bar{k}_C(1+\theta^3), \end{cases} \tag{2.9}$$

where the term $\{\sigma\theta^3\}$ with $\sigma > 0$ accounts for the *radiative* effects.

3. Main results. We are now ready to state the existence result for the initial boundary value problem introduced in section 1.

THEOREM 3.1. *Let $\Omega \subset \mathbb{R}^3$ be a bounded domain with a boundary $\partial\Omega \in C^{2+\nu}, \nu > 0$. Suppose that the pressure p is determined by the equation of state* (1.21) *with $a > 0$ and p_e and p_θ satisfying* (1.8)*. In addition, let the viscous stress tensor $\mathbb{S}$ be given by* (1.4)*, where μ and ζ are continuous differentiable globally Lipschitz functions of θ satisfying* (2.8)*. Similarly, let the heat flux $\mathbf{q}$ be given by* (1.14) *with $k(\theta)$ satisfying* (2.9)*. Finally, assume that the initial data $\rho_0, \mathbf{m}_0, \theta_0, \rho_{1_0}$ satisfy*

$$\begin{cases} \rho_0 \ge 0,\ \rho_0 \in L^\gamma(\Omega), \\ \mathbf{m}_0 \in [L^1(\Omega)]^3,\ \frac{|\mathbf{m}_0|^2}{\rho_0} \in L^1(\Omega), \\ \theta_0 \in L^\infty(\Omega),\ 0 < \underline{\theta} \le \theta_0(x) \le \bar{\theta} \text{ for a.e. } x \in \Omega, \\ \rho_{1_0} \ge 0,\ \rho_{1_0} \in L^\gamma(\Omega). \end{cases} \tag{3.1}$$

Then, for any given $T > 0$ the initial boundary value problem (1.1)–(1.4)*, together with* (1.27)–(1.30) *and* (1.31)*, has a variational solution on $(0,T) \times \Omega$.*

In the remaining part of this paper we will carry out the strategy outlined in the introduction. The proof of this theorem will be given in detail in sections 4, 5, 6, and 7.

4. Solvability of approximating problems. In this section we construct a sequence of approximate problems by adding appropriate regularizations in (1.23)–(1.26).

Taking into account (2.7), the elastic pressure component $p_e(\rho)$ of the pressure can be decomposed in two parts p_m and p_b:

$$p_e(\rho) = p_m(\rho) + p_b(\rho),$$

where the former p_m is a nondecreasing function, the latter p_b is a bounded function on $[0,\infty)$, while both are elements in $C([0,\infty))$ (see also [14], [13]). The reason for this decomposition will be apparent in what follows, where the properties of the two components will be instrumental in obtaining useful *energy* and *entropy* estimates.

4.1. Faedo–Galerkin approximations. Let

$$X_n = \operatorname{span}\{\eta_j\}_{j=1}^n$$

be a finite-dimensional space with $\eta_j \in \mathcal{D}(\Omega;\mathbb{R}^3)$, $j = 1,2,\ldots,$ a dense subset in $C_0^1(\bar{\Omega};\mathbb{R}^3)$.

The approximate velocities $\mathbf{u}_n \in C([0,T];X_n)$ satisfy a set of integral equations of the form

$$\begin{aligned} \int_\Omega \rho\mathbf{u}_n(\tau)\cdot\eta\,dx - \int_\Omega \mathbf{m}_{0,\delta}\cdot\eta &= \int_0^\tau\int_\Omega (\rho\mathbf{u}_n\otimes\mathbf{u}_n - \mathbb{S}_n) : \nabla\eta\,dxdt \\ &+ \int_0^\tau\int_\Omega \left(p_m(\rho) + p_m(\rho_1) + \frac{a}{3}\theta^4 + \frac{R}{m_2}\theta p_\theta(\rho) + LR\theta p_\theta(\rho_1) + \delta\rho^\beta\right)\operatorname{div}\eta\,dxdt \\ &+ \int_0^\tau\int_\Omega [p_b(\rho) + p_b(\rho_1)]\operatorname{div}\eta - \varepsilon[\nabla\mathbf{u}_n\nabla\rho]\cdot\eta\,dxdt \end{aligned} \tag{4.1}$$

for any test function $\eta \in X_n = \operatorname{span}\{\eta_j\}_{j=1}^n$, $\eta_j \in \mathcal{D}(\Omega)^N$, and all $\tau \in [0,T]$.

Having replaced the momentum equation with a family of integral relations, we are ready to introduce a sequence of approximate problems obtained by adding appropriate regularization ε- and δ-terms to the set of original partial differential equations. The addition of these extra terms yields a new set of equations which are easier to solve assuming that ε and δ are fixed. The challenge in this approach arises in passing to the limit for $\varepsilon \to 0$ and $\delta \to 0$. The sequence of approximate problems is presented below.

Modified continuity equation. The density $\rho = \rho[\mathbf{u}_n]$ is determined as the (unique) solution of the initial boundary value problem

$$\begin{cases} \partial_t \rho + \operatorname{div}(\rho \mathbf{u}_n) = \varepsilon \Delta \rho, \\ \quad \nabla \rho \cdot \mathbf{n}|_{\partial\Omega} = 0, \\ \quad \rho(0,\cdot) = \rho_{0,\delta} \end{cases} \tag{4.2}$$

with the initial approximation of the density $\rho_{0,\delta} \in C^{2+\nu}(\bar{\Omega})$ also satisfying

$$\begin{cases} 0 < \delta \le \rho_{0,\delta} \le \delta^{-\frac{1}{2\beta}} \quad \text{on} \quad \Omega, \\ \rho_{0,\delta} \to \rho_0 \quad \text{in} \quad L^\gamma(\Omega), \quad |\{\rho_{0,\delta} < \rho_0\}| \to 0 \quad \text{for} \quad \delta \to 0. \end{cases}$$

With the aid of the *maximum principle* the problem (4.2) admits the a priori estimate

$$\begin{aligned} &\left(\inf_{x\in\Omega} \rho(0,x) \right) \exp\left(- \int_0^t \|\operatorname{div}\mathbf{u}_n(s)\|_{L^\infty(\Omega)}\, ds \right) \\ &\quad \le \rho(t,x) \\ &\quad \le \left(\sup_{x\in\Omega} \rho(0,x) \right) \exp\left(\int_0^t \|\operatorname{div}\mathbf{u}_n(s)\|_{L^\infty(\Omega)}\, ds \right); \end{aligned}$$

in particular the (approximate) density of the mixture has positive lower and upper bounds provided that we have control of the norm of $\{\operatorname{div}\mathbf{u}_n\}$.

For fixed ε, the solvability of the regularized problem for the continuity equation (4.2) is obtained by a simple fixed point argument (see [20]).

Modified thermal equation. Now, given $\rho, \mathbf{u}_n$, the temperature will be viewed as a solution of the regularized thermal energy equation in $(0,T) \times \Omega$ supplemented with boundary and initial conditions as follows:

(4.3)

$$\begin{cases} \partial_t \left[a\theta^4 + \frac{c_v}{m_2}\rho\theta + LR\rho_1\theta + \rho_1 P_e(\rho_1) \right] + \operatorname{div}\left[\left(a\theta^4 + \frac{c_v}{m_2}\rho\theta + L\rho_1\theta + \rho_1 P_e(\rho_1) \right) \mathbf{u}_n \right] \\ -\operatorname{div}\left[(\kappa_C(\rho,\theta) + \delta\theta^3)\nabla\theta + \frac{9\gamma-5}{4} Ld\theta \nabla \log(\rho_1\theta) \right] = \varepsilon |\nabla\rho|^2 \left(\frac{p_m'(\rho)}{\rho} + \delta\beta\rho^{\beta-2} \right) \\ + \mathbb{S}_n : \nabla\mathbf{u}_n - \left[\frac{a}{3}\theta^4 + \frac{R}{m_2}\theta p_\theta(\rho) + L\theta p_e(\rho_1) + p_e(\rho_1) \right] \operatorname{div}\mathbf{u}_n - \sum_{i=1}^2 h_i w_i, \end{cases}$$

$$\begin{aligned} &\nabla\theta \cdot \mathbf{n}|_{\partial\Omega} = 0, \\ &\theta(0,\cdot) = \theta_{0,\delta}. \end{aligned}$$

The functions $\theta_{0,\delta} \in C^{2+\nu}(\bar{\Omega})$ satisfy in addition

$$\begin{cases} \nabla\theta_{0,\delta} \cdot \mathbf{n}|_{\partial\Omega} = 0, \quad 0 < \underline{\theta} < \theta_{0,\delta} \le \bar{\theta} \quad \text{on} \quad \Omega, \\ \theta_{0,\delta} \to \theta_0 \quad \text{in} \quad L^1(\Omega) \quad \delta \to 0. \end{cases}$$

The viscous stress tensor $\mathbb{S}_n$ is given by

$$\mathbb{S}_n = \mu(\theta)\left(\nabla \mathbf{u}_n + \nabla \mathbf{u}_n{}^T - \frac{2}{3}\mathrm{div}\mathbf{u}_n\,\mathbb{I}\right) + \zeta(\theta)\,\mathrm{div}\mathbf{u}_n\,\mathbb{I}.$$

The modified thermal equation (4.3) is a nondegenerate parabolic equation with respect to $U = \theta^4$ with sublinear coefficients, and hence the existence of the solution temperature θ_n follows easily (we refer the reader also to [12], [13], [14], [26], where relevant equations were treated).

By applying the classical maximum principle argument we can conclude that the temperature $\theta = \theta(t,x)$ is strictly positive, namely,

$$\theta(t,x) \geq \bar{C} > 0 \quad \text{for } t \in [0,T], x \in \Omega.$$

We can therefore divide the modified thermal energy equation by θ in order to obtain the approximate entropy balance,

$$\begin{aligned}
&\partial_t(\rho s) + \mathrm{div}(\rho s \mathbf{u}_n) + \mathrm{div}\left(-\frac{\kappa_C(\theta_n) + \delta\theta^3}{\theta}\nabla\theta - s_{f_1} L d \nabla \log(\rho_1\theta)\right) \\
&\quad = \left(\frac{\mathbb{S}_n : \nabla \mathbf{u}_n}{\theta} + \frac{\kappa_C(\theta) + \delta\theta^3}{\theta^2}|\nabla\theta|^2 + Ld|\nabla\log(\rho_1\theta)|^2 - \sum_{k=1}^{2} g_k w_k\right) \\
&\qquad + \frac{\varepsilon\delta\Gamma}{\theta}\rho^{\Gamma-2}|\nabla\rho|^2 + \frac{c_v}{m_2}(\log(\theta_n) - 1)(\varepsilon\Delta\rho_n + w_1(\rho_1,\rho_2,\theta) - \varepsilon\Delta\rho_{1_n}).
\end{aligned} \tag{4.4}$$

Modified species conservation equation. The regularized species conservation equation is now given as

$$\begin{cases}
\partial_t(\rho_1) + \mathrm{div}(\rho_1\mathbf{u}_n) = \varepsilon\Delta\rho_1 + w_1(\rho_1,\rho_2,\theta) + \mathrm{div}(d\nabla\log(\rho_1\theta)), \\
\nabla\rho_1 \cdot \mathbf{n}|_{\partial\Omega} = 0, \\
\rho_1(0,\cdot) = \rho_{1_{0,\delta}}.
\end{cases} \tag{4.5}$$

In accordance with physical considerations, we require that the density of the individual component be less than the total density of the mixture, namely,

$$\rho_1 = \rho f_1 \leq \rho.$$

This property will be used in what follows to improve the time integrability of certain terms such as $\{\rho_1\mathbf{u}\}, \{\rho_1\log(\theta)\mathbf{u}\}$.

Finally, the initial approximations of the mass fraction of the reactant $\rho_{1_{0,\delta}} \in C^{2+\nu}(\bar{\Omega})$ satisfy

$$\begin{cases}
\nabla\rho_{1_{0,\delta}} \cdot \mathbf{n}|_{\partial\Omega} = 0, \quad 0 \leq \rho_{1_{0,\delta}} \leq \rho_0 \quad \text{on} \quad \Omega, \\
\rho_{1_{0,\delta}} \to \rho_{1_0} \quad \text{in} \quad L^1(\Omega) \quad \delta \to 0.
\end{cases}$$

The regularized species conservation equation (4.5) is a parabolic quasi-linear equation. Applying standard techniques [17], [26] one can deduce the existence of solutions (see [20], [12], [13]). Finally, the solvability of the regularized momentum equation is given by the Faedo–Galerkin method, with ρ, θ, ρ_1 obtained from (4.2), (4.3), (4.5).

We remark that an important consideration in the above construction is the necessity of keeping the total energy constant at each step of the approximation. The addition of the artificial pressure $\{\delta\rho^{\beta-2}\}$ terms, on the other hand, is essential in ensuring that the pressure estimates hold true even as the artificial viscosity vanishes and in resolving some technical issues related to temperature estimates.

This approach yields for any fixed $n = 1, 2, \ldots$ a sequence of approximate solutions $\{\rho_n, \mathbf{u}_n, \theta_n, \rho_{1_n}\}$ for the system (4.2)–(4.5) defined on the whole time interval $(0, T)$.

5. The second-level approximate solutions. Relation (4.1) can now be expressed as

$$\begin{aligned}\int_\Omega \partial_t(\rho_n\mathbf{u}_n)\cdot\eta\,dx &= \int_\Omega [\rho\mathbf{u}_n\otimes\mathbf{u}_n - \mathbb{S}_n] : \nabla\eta\,dx \\ &\quad + \int_\Omega \left[p_m(\rho) + p_m(\rho_1) + \frac{a}{3}\theta^4 + \frac{R}{m_2}\theta p_\theta(\rho) \right. \\ &\qquad \left. + L\theta p_\theta(\rho_1) + \delta\rho^\beta\right]\mathrm{div}\eta\,dx \\ &\quad + \int_\Omega \{[p_b(\rho) + p_b(\rho_1)]\mathrm{div}\eta - \varepsilon[\nabla\mathbf{u}_n\nabla\rho]\cdot\eta\}\,dx, \end{aligned} \tag{5.1}$$

supplemented with the initial conditions

$$\int_\Omega \rho_n\mathbf{u}_n\cdot\eta\,dx = \int_\Omega \mathbf{m}_{0,\delta}\cdot\eta\,dx$$

to be satisfied for any $\eta \in X_n$ with the initial momenta $\mathbf{m}_{0,\delta}$ given by

$$\mathbf{m}_{0,\delta}(x) = \begin{cases}\mathbf{m}_0 & \text{if} \quad \rho_{0,\delta}(x) \geq \rho_0(x), \\ 0 & \text{for} \quad \rho_{0,\delta}(x) < \rho_0(x).\end{cases}$$

5.1. Energy and entropy estimates. The existence of the approximate solution sequence $\{\rho_n, \mathbf{u}_n, \theta_n, \rho_{1_n}\}$ has now been established and we are ready to proceed with the program outlined in section 4.

Considering a suitable choice of test functions $\eta = \mathbf{u}_n(t)$ in (4.1), integrating in space both the (regularized) thermal energy equation (4.3) and the species conservation equation (4.5), and adding the resulting relations give rise to an energy equality of the form

$$\begin{aligned}&\frac{d}{dt}\int_\Omega \left[\frac{1}{2}\rho_n|\mathbf{u}_n|^2 + \rho_n P_m(\rho_n) + \rho_{1_n}P_e(\rho_{1_n}) + \frac{\delta}{\beta-1}\rho_n^\beta + a\theta_n^4 + \frac{c_v}{m_2}\rho_n\theta_n + Lc_v\rho_{1_n}\theta\right]dx \\ &\quad = \int_\Omega [p_b(\rho_n) + p_b(\rho_{1_n})]\mathrm{div}\mathbf{u}_n dx + \int_\Omega w_1(\rho_1,\rho_2,\theta)dx - \varepsilon\int_\Omega \Delta\rho_{1_n}dx\end{aligned} \tag{5.2}$$

with

$$P_m(\rho) = \int_1^\rho \frac{p_m(z)}{z^2}dz.$$

The approximate entropy balance (4.4) now yields the entropy inequality

$$\begin{aligned}&\partial_t(\rho_n s_n) + \mathrm{div}(\rho_n s_n\mathbf{u}_n) + \mathrm{div}\left(-\frac{\kappa_C(\theta_n) + \delta\theta_n^3}{\theta_n}\nabla\theta_n - s_{f_{1_n}}Ld\nabla\log(\rho_{1_n}\theta_n)\right) \\ &\qquad \geq \left(\frac{\mathbb{S}_n : \nabla\mathbf{u}_n}{\theta_n} + \frac{\kappa_C(\theta_n) + \delta\theta_n^3}{\theta_n^2}|\nabla\theta_n|^2 + Ld|\nabla\log(\rho_{1_n}\theta_n)|^2 - \sum_{k=1}^2 g_k w_k\right) \\ &\qquad + \frac{c_v}{m_2}(\log(\theta_n) - 1)\left(\varepsilon\Delta\rho_n + w_1(\rho_1,\rho_2,\theta) - \varepsilon\Delta\rho_{1n}\right).\end{aligned} \tag{5.3}$$

The regularized continuity equation (4.2) multiplied by ρ_n and integrated over Ω yields

$$\frac{d}{dt}\int_\Omega \frac{1}{2}\rho_n^2 dx + \varepsilon \int_\Omega |\nabla \rho_n|^2 dx = -\frac{1}{2}\int_\Omega \rho_n^2 \mathrm{div}\mathbf{u}_n dx. \tag{5.4}$$

Now (5.2), (5.3), (5.4) give rise to

$$\begin{aligned}
&\frac{d}{dt}\int_\Omega \bigg[\frac{1}{2}\rho_n|\mathbf{u}_n|^2 + \rho_n P_m(\rho_n) + \rho_{1_n} P_m(\rho_{1_n}) + \frac{\delta}{\beta-1}\rho_n^\beta \\
&\qquad + a\theta_n^4 + \frac{c_v}{m_2}\rho_n\theta_n + Lc_v\rho_{1_n}\theta_n\bigg] dx \\
&\quad + \frac{d}{dt}\int_\Omega \left(\frac{1}{2}\rho_n^2 - \rho_n(s_{F_n} + s_{f_{1_n}}) + \rho_{1n} + \varepsilon|\nabla\rho_n|^2 + \varepsilon\Delta\rho_{1n}\right) dx \\
&\quad + \int_\Omega \left(\frac{\mathbb{S}_n : \nabla\mathbf{u}_n}{\theta_n} + \frac{\kappa_C(\theta_n) + \delta\theta_n^3}{\theta_n^2}|\nabla\theta_n|^2 + Kd|\nabla\log(\rho_{1_n}\theta_n)|^2 - \sum_{k=1}^2 g_k w_k\right) dx \\
&\le \int_\Omega \left(p_b(\rho_n) + p_b(\rho_{1_n}) - \frac{1}{2}\rho_n^2\right)\mathrm{div}\mathbf{u}_n dx + \int_\Omega \frac{c_v}{m_2}(\log(\theta_n) - 1) w_1(\rho_1,\rho_2,\theta) dx \\
&\quad + \varepsilon\int_\Omega \frac{c_v}{m_2}\left(\nabla\log(\theta_n)\nabla\rho_n + (\log(\theta_n)-1)\Delta\rho_{1n}\right) dx.
\end{aligned} \tag{5.5}$$

5.2. Dissipation estimates. Starting from *Newton's law for viscosity* and using the hypotheses (2.8) on viscosity coefficients and Hölder's inequality, we deduce (see also [14], [13]) that

$$|\nabla\mathbf{u}_n + \nabla\mathbf{u}_n^t|^b \le c\left(\theta_n^{\alpha-1}|\nabla\mathbf{u}_n + \nabla\mathbf{u}_n^t|^2 + \theta_n^4\right), \quad \text{where } b = \frac{8}{5-\alpha}. \tag{5.6}$$

Moreover, the hypothesis (2.9) on the *heat conductivity* gives us the following bounds on the *absolute temperature*:

$$\int_\Omega |\nabla\log(\theta_n)|^2 + |\nabla\theta_n^{\frac{3}{2}}|^2 dx \le \int_\Omega \frac{\kappa_C(\theta_n) + \sigma\theta_n^3}{\theta_n^2}|\nabla\theta_n|^2 dx. \tag{5.7}$$

As stated in Definition 2.1 the total energy is constant of motion. Boundness of the total energy and properties (2.7) give rise to the following estimates:

$$\begin{aligned}
\sqrt{\rho_n}|\mathbf{u}_n| &\quad \text{bounded in} \quad L^\infty(0,T;L^2(\Omega)),\\
\rho_n &\quad \text{bounded in} \quad L^\infty(0,T;L^\beta(\Omega)),\\
\rho_{1n} &\quad \text{bounded in} \quad L^\infty(0,T;L^\beta(\Omega)),
\end{aligned}$$

and

$$\operatorname*{ess\,sup}_{t\in(0,T)}\int_\Omega \theta_n^4(t)\, dx \le c.$$

In addition, taking into consideration (5.5), (5.6), and (5.7) we get the following estimates:

$$\sup_{t\in[0,T]}\left\{\|\rho_n\|_{L^\beta(\Omega)}+\|\rho_n|\mathbf{u}_n|^2\|_{L^1(\Omega)}+\|\rho_n\theta_n\|_{L^1(\Omega)}+\|\rho_{1_n}\|_{L^\beta(\Omega)}\right\}\le c(\delta), \tag{5.8}$$

$$\sup_{t\in[0,T]}\left\{\|\log(\theta_n)\|_{L^1(\Omega)}+\|\rho_n\log(\theta_n)\|_{L^1(\Omega)}+\|\theta_n\|_{L^4(\Omega)}\right\}\le c(\delta), \tag{5.9}$$

$$\sup_{t\in[0,T]}\left\{\|\rho_{1_n}\log(\theta_n)\|_{L^1(\Omega)}+\|\rho_{1_n}\theta_n\|_{L^1(\Omega)}+\|\rho_{1_n}\log(\rho_{1_n})\|_{L^1(\Omega)}\right\}\le c(\delta), \tag{5.10}$$

$$\int_0^T\!\!\int_\Omega \frac{\mathbb{S}_n:\nabla\mathbf{u}_n}{\theta_n}+|\nabla\log(\theta_n)|^2+|\nabla\theta_n^{\frac{3}{2}}|^2+\varepsilon|\nabla\rho_n|^2+||\nabla\log(\rho_{1_n}\theta_n)|^2|dxdt\le c(\delta), \tag{5.11}$$

and

$$\|\mathbf{u}_n\|_{L^b(0,T;W_0^{1,b}(\Omega))}\le c(\delta)\quad \text{with } b=\frac{8}{5-\alpha}. \tag{5.12}$$

LEMMA 5.1. *For the density of the individual species ρ_{1_n}, the following hold true:*

$$\begin{cases}\log(\rho_{1_n}) & \textit{is bounded in}\quad L^2([0,T];W^{1,2}(\Omega)),\\ \varepsilon\nabla\sqrt{\rho_{1_n}} & \textit{is bounded in}\quad L^2([0,T];L^2(\Omega)).\end{cases} \tag{5.13}$$

Proof. Multiplying the regularized species conservation equations by $F'(\rho_{1_n})$ with $F(\rho_{1_n})=\rho_{1_n}\log(\rho_{1_n})$ and using the boundary conditions, we obtain

$$\begin{aligned}&\int_\Omega F(\rho_{1_n})(t)\,dx+\int_0^t\int_\Omega|\nabla\log(\rho_{1_n})|^2\,dxds+\varepsilon\int_0^t\int_\Omega\frac{|\nabla\rho_{1_n}|^2}{\rho_{1_n}}\,dxdt\\ &\qquad=\int_\Omega F(\rho_{1_n})(0)\,dx-\int_0^t\int_\Omega\rho_{1_n}\mathrm{div}\mathbf{u}_n\,dxds\\ &\qquad\quad-\int_0^t\int_\Omega\nabla\log(\rho_{1_n})\nabla\log(\theta_n)\,dx\,ds-\int_0^t\int_\Omega F'(\rho_{1_n})\omega\,dxdt.\end{aligned}$$

The last relation implies that

$$\begin{aligned}&\int_\Omega\rho_{1_n}\,\log(\rho_{1_n})(t)\,dx+c\int_0^t\int_\Omega|\nabla\log(\rho_{1_n})|^2\,dxds+\varepsilon\int_0^t\int_\Omega|\nabla\sqrt{\rho_{1_n}}|^2\,dxds\\ &\le\int_\Omega\rho_{1_n}\,\log(\rho_{1_n})(0)\,dx+\int_0^t\int_\Omega\rho_{1_n}\,\mathrm{div}\mathbf{u}_n\,dxds+\tilde{c}\int_0^t\int_\Omega|\nabla\log(\theta_n)|^2dxds\\ &\quad-\int_0^t\int_\Omega(\log(\rho_{1_n})+1)\omega\,dxds.\end{aligned}$$

The term

$$\int_0^t\int_\Omega|\nabla\log(\theta_n)|^2\,dxds$$

can be controlled using the estimates (5.11), whereas the term

$$\int_0^t\int_\Omega\rho_{1_n}\,\mathrm{div}\mathbf{u}_n\,dxds\le\|\rho_{1_n}\|_{L^2(\Omega)}\|\mathrm{div}\mathbf{u}\|_{L^2(\Omega)}$$

can be balanced using (5.8) and (5.12).

Therefore,

$$\begin{aligned} \int_\Omega \rho_{1_n} \log(\rho_{1_n})(t)\,dx + c\int_0^t \int_\Omega |\nabla \log(\rho_{1_n})|^2\,dxds + \varepsilon \int_0^t \int_\Omega |\nabla\sqrt{\rho_{1_n}}|^2\,dxdt \\ \le \int_\Omega \rho_{1_n} \log(\rho_{1_n})(0)\,dx + c_2, \end{aligned} \tag{5.14}$$

or equivalently

$$\begin{aligned} \|\rho_{1_n}\log(\rho_{1_n})(t)\|_{L^1(\Omega)} + c_1\int_0^t \|\nabla\log(\rho_{1_n})\|^2_{L^2(\Omega)}\,dt + \int_0^t \|\nabla\sqrt{\rho_{1_n}}\|^2_{L^2(\Omega)}\,dt \\ \le \|\rho_{1_n}\log(\rho_{1_n})(0)\|_{L^1(\Omega)} + c_2. \end{aligned}$$

This estimate yields the result. □

Starting from the continuity equation and taking into consideration standard compactness arguments by the aid of (5.8)–(5.12), we get

$$\rho_n \longrightarrow \rho \quad \text{in } C([0,T], L^\beta_{weak}(\Omega)),$$

and analogously

$$\rho_{1_n} \longrightarrow \rho_1 \quad \text{in } C([0,T], L^\beta_{weak}(\Omega)).$$

By using the estimates obtained in the previous steps we can assume

$$\mathbf{u}_n \longrightarrow \mathbf{u} \quad \text{weakly in } L^b(0,T;W_0^{1,b}(\Omega)),$$

$$\rho_n\mathbf{u}_n \longrightarrow \rho\mathbf{u} \quad \text{weakly-}^* \text{ in } L^\infty(0,T;L^{\frac{2\beta}{\beta+1}}(\Omega)),$$

where ρ, $\mathbf{u}$ satisfy (4.2) together in the sense of distribution. Taking now into account that the species density ρ_{1_n} is a nonnegative function satisfying the additional requirement $\rho_{1_n} \le \rho_n$ we have that

$$\int_0^t \int_\Omega \rho_{1_n}|\mathbf{u}_n|^2\,dxdt \le \int_0^t \int_\Omega \rho_n|\mathbf{u}_n|^2\,dxdt < C < \infty$$

by the boundness of the total energy. In particular,

$$\rho_{1_n}\mathbf{u}_n \longrightarrow \rho_1\mathbf{u} \quad \text{weakly-}^* \text{ in } L^\infty(0,T;L^{\frac{2\beta}{\beta+1}}(\Omega)),$$

where ρ_{1_n}, $\mathbf{u}$ satisfy the regularized species conservation equation (4.5). It is also worth noting at this point that *weak convergence preserves inequalities*, and therefore the requirement that we imposed in the construction of the approximating procedure leading to the estimate satisfied by the approximate densities, namely,

$$0 \le \rho_{1_n} \le \rho_n$$

will also be verified in what follows for the corresponding limiting quantities.

5.3. Gradient density estimate. Using standard $L^p - L^q$ estimates for the regularized continuity equation (4.2) and taking into consideration (5.8)–(5.12), we get, in the spirit of Feireisl [20],

$$\varepsilon^p \int_0^T \|\nabla \rho_n\|_{L^{\frac{2\beta}{\beta+1}}(\Omega);\mathbb{R}^3} dt \leq c(\delta, p) \quad \text{for any} \quad p \in [1, \infty).$$

Using now the Sobolev embedding theorem $W_0^{1,r}(\Omega) \subset L^{\frac{3r}{3-r}}(\Omega)$ and standard interpolation estimates, we get

$$\varepsilon \|\nabla \rho_n\|_{L^q(0,T;L^q(\Omega;\mathbb{R}^3))} \quad \text{for some} \quad q > r' \quad \text{with} \quad r = \frac{8}{5-\alpha}.$$

In fact, one can improve the estimates on ρ_n at this point using the L^p-theory of parabolic equations to obtain the following lemma.

LEMMA 5.2. *There exists $p > 1$ such that*

$$\partial_t \rho_n, \ \Delta \rho_n \quad \textit{are bounded in } L^p((0,T) \times \Omega),\ p > 1,$$

independently of n. Consequently, the limit functions ρ, $\mathbf{u}$ satisfy (4.2) a.e. on $(0,T) \times \Omega$, whereas the boundary condition and initial condition hold in the sense of traces.

Proof. The result is obtained by applying the L^p-theory of parabolic equations. This process is rather standard; we included above some of the basic steps for completeness and refer the reader to [20], [22], [14], [26]. $\square$

Similarly, the following lemma holds true.

LEMMA 5.3. *There exists $p > 1$ such that*

$$\partial_t \rho_{1_n}, \ \Delta \rho_{1_n} \quad \textit{are bounded in } L^p((0,T) \times \Omega),\ p > 1,$$

independently of n. Consequently, the limit functions ρ_1, $\mathbf{u}$ satisfy (4.5) a.e. on $(0,T) \times \Omega$, whereas the boundary condition and initial condition hold in the sense of traces.

The following lemma will be useful in order to obtain the pointwise convergence of the temperature sequence θ_n.

LEMMA 5.4. *Let $\Omega \subset \mathbb{R}^N$, $N \geq 2$, be a bounded Lipschitz domain and $\Lambda \geq 1$ a given constant. Let $\rho \geq 0$ be a measurable function satisfying*

$$0 < M \leq \int_\Omega \rho dx, \quad \int_\Omega \rho^\beta \leq K \quad \textit{for} \quad \beta > \frac{2N}{N+2}.$$

Then there exists a constant $c = c(M,K)$ such that

$$\|v\|_{L^2(\Omega)} \leq c(M,K) \left(\|\nabla v\|_{L^2(\Omega)} + \left(\int_\Omega \rho |v|^{\frac{1}{\Lambda}} \right)^\Lambda \right)$$

for any $v \in W^{1,2}(\Omega)$.

Proof. For the proof we refer the reader to Lemma 5.1 in [14]. $\square$

Using the Poincaré inequality, as presented in Lemma 5.4, as well as the estimates on $\{\nabla \log(\theta_n)\}, \{\nabla \theta_n^{\frac{3}{2}}\}$ in (5.11), we deduce uniform bounds on $\log(\theta)$ and $\theta^{\frac{3}{2}}$. Therefore, it is possible to extract a subsequence of θ_n such that

$$\begin{cases} \theta_n \longrightarrow \theta \quad \text{weakly in} \quad L^2(0,T;W^{1,2}(\Omega)), \\ \theta_n \longrightarrow \theta \quad \text{weakly-* in} \quad L^\infty(0,T;L^4(\Omega)), \\ \log(\theta_n) \longrightarrow \overline{\log(\theta)} \quad \text{weakly in} \quad L^2(0,T;W^{1,2}(\Omega)). \end{cases} \tag{5.15}$$

Using now the fact that $\{\rho_n \log(\theta_n), \rho_{1_n} \log(\theta_n)\}$ satisfy the entropy inequality (5.3), we deduce that

$$\begin{cases} \rho_n \log(\theta_n) & \text{bounded in } L^\infty(0,T;L^1(\Omega)) \cap L^2(0,T;L^{\frac{6\beta}{\beta+6}}(\Omega)), \\ \rho_n \mathbf{u}_n \log(\theta_n) & \text{bounded in } L^2(0,T;L^{\frac{6\beta}{4\beta+3}}(\Omega)) \end{cases}$$

and

$$\begin{cases} \rho_{1_n} \log(\theta_n) & \text{bounded in } L^\infty(0,T;L^1(\Omega)) \cap L^2(0,T;L^{\frac{6\beta}{\beta+6}}(\Omega)), \\ \rho_{1_n} \mathbf{u}_n \log(\theta_n) & \text{bounded in } L^2(0,T;L^{\frac{6\beta}{4\beta+3}}(\Omega)). \end{cases}$$

5.4. Pointwise convergence for the temperature. The following lemma will be very useful in what follows.

LEMMA 5.5. *Let $\Omega \subset \mathbb{R}^N$, $N \geq 2$, be a bounded Lipschitz domain. Let $\{v_n\}$ be a sequence of functions bounded in*

$$L^2(0,T;L^q(\Omega)) \cap L^\infty(0,T;L^1(\Omega)), \quad q > \frac{2N}{N+2}.$$

Furthermore, assume that

$$\partial_t v_n \geq g_n \quad in \quad \mathcal{D}'((0,T)\times\Omega),$$

where the distributions g_n are bounded in the space $L^1(0,T;W^{-m,p}(\Omega))$ for $m \geq 1$, $p > 1$. Then

$$v_n \longrightarrow v \quad in\ L^2(0,T;W^{-1,2}(\Omega))$$

passing to a subsequence as the case may be.

Proof. The proof is given in Lemma 6.3 of Chapter 6 in [20]. $\square$

By a direct application of Lemma 5.5 to the sequence

$$\left\{ \frac{4a}{3}\theta^3 + \frac{c_v}{m_2}\rho_n \log(\theta_n) + Lc_v\rho_{1_n}\log(\theta_n) \right\},$$

we get

$$\begin{cases} \frac{4a}{3}\theta^3 + \frac{c_v}{m_2}\rho_n \log(\theta_n) + Lc_v\rho_{1_n}\log(\theta_n) \\ \qquad\qquad\qquad \downarrow \\ \frac{4a}{3}\bar{\theta}3 + \frac{c_v}{m_2}\rho\,\overline{\log(\theta)} + Lc_v\rho_1\,\overline{\log(\theta)} \\ \qquad\qquad\qquad\qquad\qquad \text{weakly in} \quad L^2(0,T;W^{-1,2}(\Omega)). \end{cases}$$

Using now (5.15) we can conclude that

$$\begin{cases} \int_0^T\!\int_\Omega \left(\frac{4a}{3}\theta^3 + \frac{c_v}{m_2}\rho_n \log(\theta_n) + L\rho_{1_n}\log(\theta_n) \right)\theta_n dxdt, \\ \qquad\qquad\qquad \downarrow \\ \int_0^T\!\int_\Omega \left(\frac{4a}{3}\bar{\theta}^3 + \frac{c_v}{m_2}\rho\,\overline{\log(\theta)} + Lc_v\rho_1\,\overline{\log(\theta)} \right)\theta dxdt. \end{cases}$$

Since the function

$$y \to \left\{ \frac{4a}{3}y^3 + \frac{c_v}{m_2}\rho\log(y) + Lc_v\rho_1\log(y) \right\}$$

is nondecreasing we have

$$\theta_n \longrightarrow \theta \quad \text{strongly in} \quad L^1((0,T)\times\Omega). \tag{5.16}$$

Now by interpolation arguments we have that

$$\begin{cases} \theta_n \longrightarrow \theta & \text{strongly in} \quad L^p((0,T)\times\Omega) \quad \text{for} \quad p>4, \\ \mathbb{S}_n \longrightarrow \mathbb{S} & \text{weakly in} \quad L^q((0,T)\times\Omega) \quad \text{for} \quad q>1. \end{cases} \tag{5.17}$$

Similarly, we get

$$\rho_n \longrightarrow \rho \quad \text{in } L^p((0,T)\times\Omega) \text{ for } p>\beta. \tag{5.18}$$

By using the same argument as in [12] we have

$$\rho_n\mathbf{u}_n \longrightarrow \rho\mathbf{u} \quad \text{in } C\left([0,T]; L^{\frac{2\beta}{\beta+1}}_{weak}(\Omega)\right),$$

which allows us to pass to the limit and to get that the limit functions ρ, $\mathbf{u}$, θ satisfy the *regularized momentum equation* in $\mathcal{D}'((0,T)\times\Omega)$.

In the same way we can let $n\to\infty$ in the energy inequality (5.2) in order to get

$$\begin{aligned} &\int_0^T\!\!\int_\Omega \partial_t\psi\left(\frac{1}{2}\rho|\mathbf{u}|^2+\rho P_m(\rho)+\rho_1P_m(\rho_1)+\frac{\delta}{\beta-1}\rho^\beta+a\theta^4+\frac{c_v}{m_2}\rho\theta+Lc_v\rho_1\theta\right)dxdt \\ &\quad=\int_\Omega\left(\frac{1}{2}\frac{\mathbf{m}_{0,\delta}}{\rho_{0,\delta}}+\rho_{0,\delta}P_m(\rho_{0,\delta})+\rho_{1_0,\delta}P_m(\rho_{1_{0,\delta}})+\frac{\delta}{\beta-1}\rho_{0,\delta}^\beta+a\theta_{0,\delta}^4\right)\,dx \\ &\qquad+\int_\Omega\left(\frac{c_v}{m_2}\rho_{0,\delta}\theta_{0,\delta}+Lc_v\rho_{1_{0,\delta}}\theta_{0,\delta}\right)dx \\ &\qquad+\int_0^T\int_\Omega\psi\left([p_b(\rho)+p_b(\rho_1)]\mathrm{div}\mathbf{u}+w_1(\rho_1,\rho_2,\theta)-\varepsilon\Delta\rho_1\right)dxdt \end{aligned} \tag{5.19}$$

for any $\psi\in C^\infty[0,T]$, $\psi(0)=1$, $\psi(T)=0$, $\partial_t\psi\le 0$. The following two lemmas will be useful in what follows.

LEMMA 5.6. *Let $\Omega\subset\mathbb{R}^N$ be a bounded Lipschitz domain. Suppose that ρ is a given nonnegative function satisfying*

$$0<M\le\int_\Omega\rho\,dx,\quad \int_\Omega\rho^\beta\,dx<K,\quad \beta>\frac{2N}{N+2}.$$

(a) *Then the following two statements are equivalent:*
   (i) *The function θ is strictly positive a.e. on Ω,*

$$\rho|\log(\theta)|\in L^1(\Omega) \quad \textit{and} \quad \frac{\nabla\theta}{\theta}\in L^2(\Omega).$$

   (ii) *The function $\log(\theta)$ belongs to the Sobolev space $W^{1,2}(\Omega)$. Moreover, if this is the case, then*

$$\nabla\log(\theta)=\frac{\nabla\theta}{\theta} \quad \textit{a.e. on } \Omega.$$

(b) *The following two statements are equivalent:*

(i) *Let* ρ_1 *be a nonnegative function such that* $\rho_1 = \rho f_1$. *The quantities*

$$|\log(\rho_1\theta)| \in L^1(\Omega), \qquad \frac{\nabla(\rho_1\,\theta)}{\rho_1\theta} \in L^2(\Omega),$$

and the function $\rho_1\theta$ *is strictly positive a.e. on* Ω.

(ii) *The function* $\log(\rho_1\theta)$ *belongs to the Sobolev space* $W^{1,2}(\Omega)$. *Moreover,*

$$\nabla\log(\rho_1\theta) = \frac{\nabla\rho_1}{\rho_1} + \frac{\nabla\theta}{\theta} \quad \text{a.e. on } \Omega.$$

Proof. The proof of part (a) follows using a line of arguments similar to the ones given in [14]. We shall give only some comments on part (b). Letting $\log(\rho_1\theta) \in W^{1,2}(\Omega)$, it follows that the product $\rho_1\theta$ is positive a.e. on Ω. Then one can use the Sobolev embedding theorem to conclude that

$$|\log(\rho_1\theta)| \in L^1(\Omega) \quad \text{is integrable.}$$

The converse follows easily (see also [14]). $\square$

LEMMA 5.7. *Let* $\theta_n \to \theta$ *in* $L^2((0,T)\times\Omega$ *and let* $\log(\theta_n) \to \overline{\log(\theta)}$ *weakly in* $L^2((0,T))\times\Omega$. *Then* θ *is strictly positive a.e. on* $(0,T)\times\Omega$ *and* $\log(\theta) = \overline{\log(\theta)}$.

Proof. For the proof we refer the reader to [14]. $\square$

Using the above analysis we can now pass to the limit in the entropy inequality (5.3) to get

$$\begin{aligned}
&\int_0^T\!\!\int_\Omega \left\{\partial_t\varphi[\rho(s_F+s_{f_1})] + \nabla\varphi[\rho(s_F+s_{f_1})] - \left(\frac{\kappa_C(\theta)+\sigma\theta^3}{\theta}\nabla\theta\right)\nabla\varphi\right\}dxdt\\
&\le \int_\Omega \varepsilon\left(\frac{c_v}{m_2}\nabla[\varphi(\log\theta-1)]\nabla\rho + \Delta\rho_{1_n}\varphi\right)dx\\
&\quad - \int_0^T\!\!\int_\Omega \varphi\left(\frac{S:\nabla\mathbf{u}}{\theta} + \frac{\kappa_C(\theta)+\sigma\theta^3}{\theta^3}|\nabla\theta|^2\right)dxdt\\
&\quad - \int_0^T\!\!\int_\Omega \varphi\left(\frac{c_v}{m_2}(\log(\theta)-1)w_1(\rho_1,\rho_2,\theta)\right)dxdt\\
&\quad - \int_\Omega \varphi(0)[\rho_{0,\delta}(s_{F_{0,\delta}} + s_{f_{1_{0,\delta}}})]dx
\end{aligned} \tag{5.20}$$

for any test function φ, $\varphi \in C^\infty([0,T]\times\Omega)$, $\varphi \ge 0$, $\varphi(T) = 0$.

6. The vanishing viscosity limit. Our aim in this section is to let the artificial viscosity ε go to zero in the family of approximate solutions $\{\rho_\varepsilon, \mathbf{u}_\varepsilon, \theta_\varepsilon, \rho_{1_\varepsilon}\}$ constructed in the previous section. We remark that during this process we expect loss of regularity of ρ_ε due to the fact that the parabolic regularization $\varepsilon\Delta\rho_\varepsilon$ now vanishes. The main difficulty is to establish the strong convergence of the density ρ_ε.

6.1. Refined pressure estimates. Our goal in this section is to improve the estimates on the pressure, which so far yield only that p is bounded in the nonreflexive space $L^\infty(0,T,L^1(\Omega))$, by improving the integrability properties of the modified pressure

$$p(\rho_\varepsilon, \mathbf{u}_\varepsilon, \theta_\varepsilon, \rho_{1_\varepsilon}) + \delta\rho_\varepsilon^\beta.$$

We employ the method known as the *multipliers technique* introduced by Lions in [27] for the barotropic case and by Feireisl in [20] for handling the full system. This method involves placing a particular choice of test functions, namely,

$$\varphi(t,x) = \chi(t)\mathcal{B}[\rho_\varepsilon^\nu], \quad \chi \in \mathcal{D}(0,T),\ 0 \le \chi \le 1,$$

in the weak formulation of the momentum equation.

The quantity $\mathcal{B}[v]$ represents a suitable set of solutions to the problem (see [20])

$$\operatorname{div}(\mathcal{B}[v]) = v - \frac{1}{|\Omega|}\int_\Omega v dx \quad \text{in } \Omega, \quad \mathcal{B}[v]|_{\partial\Omega} = 0.$$

Integrating the modified pressure against suitable test functions yields the following integral relation:

$$\int_0^T \int_\Omega \left(p_e(\rho_\varepsilon) + \frac{a}{3}\theta_\varepsilon^4 + \frac{R}{m_2}\theta_\varepsilon p_\theta(\rho_\varepsilon) + L\theta_\varepsilon p_\theta(\rho_{1_\varepsilon}) + \delta\rho_\varepsilon^\beta \right) \rho_\varepsilon^\nu \, dxdt = \sum_{j=1}^8 \mathcal{L}_{j_\varepsilon}, \tag{6.1}$$

where ν is a positive constant and

$$\begin{cases} \mathcal{L}_{1_\varepsilon} = \int_0^T \chi \left(\int_\Omega p_e(\rho_\varepsilon) + \frac{a}{3}\theta_\varepsilon^4 + \frac{R}{m_2}\theta_\varepsilon p_\theta(\rho_\varepsilon) + L\theta_\varepsilon p_\theta(\rho_{1_\varepsilon}) + \delta\rho_\varepsilon^\beta dx \right) dt, \\ \mathcal{L}_{2_\varepsilon} = \int_0^T \chi \int_\Omega \mathbb{S}_\varepsilon : \nabla\mathcal{B}\left[\rho_\varepsilon - \frac{1}{|\Omega|}\right] dxdt, \\ \mathcal{L}_{3_\varepsilon} = -\int_0^T \chi \int_\Omega [\rho_\varepsilon \mathbf{u}_\varepsilon \otimes \mathbf{u}_\varepsilon] : \nabla\mathcal{B}\left[\rho_\varepsilon - \frac{1}{|\Omega|}\right] dxdt, \\ \mathcal{L}_{4_\varepsilon} = \varepsilon \int_0^T \chi \int_\Omega (\nabla\mathbf{u}_\varepsilon \nabla\rho_\varepsilon) \cdot \mathcal{B}\left[\rho_\varepsilon - \frac{1}{|\Omega|}\right] dxdt, \\ \mathcal{L}_{5_\varepsilon} = -\int_0^T \chi \int_\Omega \rho_\varepsilon \nabla\psi_\varepsilon \cdot \mathcal{B}\left[\rho_\varepsilon - \frac{1}{|\Omega|}\right] dxdt, \\ \mathcal{L}_{6_\varepsilon} = \int_0^T \partial_t\chi \int_\Omega \rho_\varepsilon \mathbf{u}_\varepsilon \cdot \mathcal{B}\left[\rho_\varepsilon - \frac{1}{|\Omega|}\right] dxdt, \\ \mathcal{L}_{7_\varepsilon} = -\varepsilon \int_0^T \chi \int_\Omega \rho_\varepsilon \mathbf{u}_\varepsilon \cdot \mathcal{B}[\Delta\rho_\varepsilon] \, dxdt, \\ \mathcal{L}_{8_\varepsilon} = \int_0^T \chi \int_\Omega \rho_\varepsilon \mathbf{u}_\varepsilon \cdot \mathcal{B}[\operatorname{div}(\rho_\varepsilon \mathbf{u}_\varepsilon)] \, dxdt. \end{cases}$$

Taking now into consideration the estimates in (5.8)–(5.12) as well as Lemma 5.1 and following the same line of arguments as in [14], [20], we show that the quantities $\mathbb{S}_\varepsilon, \rho_\varepsilon \mathbf{u} \otimes \mathbf{u}$ are bounded in $L^p((0,T)\times\Omega)$ for a certain $p > 1$. Using now the fact that the operator

$$\mathcal{B} : \left\{ f \in L^p(\Omega) | \int_\Omega f dx = 0 \right\} \to [W_0^{1,p}(\Omega)]^3$$

introduced earlier is a bounded linear operator, namely,

$$\|\mathcal{B}[f]\|_{[W_0^{1,p}(\Omega)]^3} \le c(p)\|f\|_{L^p(\Omega)} \quad \text{for any} \quad 1 < p < \infty,$$

we conclude that the integrals $\mathcal{L}_{1_\varepsilon}$–$\mathcal{L}_{3_\varepsilon}$ and $\mathcal{L}_{5_\varepsilon}$ are bounded uniformly with ε. The integral $\mathcal{L}_{4_\varepsilon}$ can be controlled using a standard density gradient estimate (see [20], [14], [12], [13]), while the other integrals can be controlled using standard embedding theorems and further properties of the operator $\mathcal{B}$. Therefore, there exist $\nu > 0$ and a positive constant $c(\delta)$ independent of ε by which

$$\int_0^T \int_\Omega \left(p_e(\rho_\varepsilon) + \frac{a}{3}\theta_\varepsilon^4 + \frac{R}{m_2}\theta_\varepsilon p_\theta(\rho_\varepsilon) + L\theta_\varepsilon p_\theta(\rho_{1_\varepsilon}) + \delta\rho_\varepsilon^\beta \right) \rho_\varepsilon^\nu \, dxdt \le c(\delta).$$

Moreover,

$$\int_0^T \int_\Omega \rho_\delta^{\gamma+\nu} + \delta \rho_\delta^{\beta+\nu}\, dxdt \le c(\delta).$$

6.2. Strong convergence of the temperature. Taking into consideration estimates (5.8)–(5.12) we may now assume that

$$\begin{cases} \theta_\varepsilon \longrightarrow \theta & \text{weakly in } L^2(0,T;W^{1,2}(\Omega)), \\ \theta_\varepsilon \longrightarrow \theta & \text{weakly-* in } L^\infty(0,T;L^4(\Omega)), \\ \log(\theta_\varepsilon) \longrightarrow \overline{\log(\theta)} & \text{weakly in } L^2(0,T;W^{1,2}(\Omega)), \end{cases} \tag{6.2}$$

$$\begin{cases} \rho_\varepsilon \longrightarrow \rho & \text{in } C([0,T], L^\beta_{weak}(\Omega)), \\ \rho_{1_\varepsilon} \longrightarrow \rho_1 & \text{in } C([0,T], L^\beta_{weak}(\Omega)), \\ \mathbf{u}_\varepsilon \longrightarrow \mathbf{u} & \text{weakly in } L^b(0,T;W_0^{1,b}(\Omega)), \\ \rho_\varepsilon \mathbf{u}_\varepsilon \longrightarrow \rho \mathbf{u} & \text{in } C([0,T], L^{\frac{2\beta}{\beta+1}}(\Omega)). \end{cases} \tag{6.3}$$

Combining (6.2) and (6.3) we obtain

$$\begin{cases} \rho_\varepsilon \log(\theta_\varepsilon)\mathbf{u}_\varepsilon \longrightarrow \rho \overline{\log\theta}\mathbf{u} & \text{weakly in } L^p((0,T)\times\Omega) \text{ for } p>1, \\ \rho_{1_\varepsilon} \log(\theta_\varepsilon)\mathbf{u}_\varepsilon \longrightarrow \rho_1 \overline{\log(\theta)}\mathbf{u} & \text{weakly in } L^p((0,T)\times\Omega) \text{ for } p>1. \end{cases} \tag{6.4}$$

Following a similar procedure to the one of the previous section we end up with

$$\theta_\varepsilon \longrightarrow \theta \quad \text{strongly in } L^2((0,T)\times\Omega). \tag{6.5}$$

6.3. Strong convergence for the density. The aim in this section is to show the strong convergence of density sequence ρ_ε. This involves understanding the time evolution of the *defect* measure

$$\mathcal{B}_{dft}[\rho_\varepsilon - \rho] = \int_\Omega [\, \overline{\rho\log(\rho)}(t) - \rho\log(\rho)(t)\,]dx. \tag{6.6}$$

Consider now the renormalized version of the regularized continuity equation

$$\begin{aligned} &\partial_t \beta(\rho_\varepsilon) + \operatorname{div}(\beta(\rho_\varepsilon)\mathbf{u}_\varepsilon) + (\beta'(\rho_\varepsilon)\rho_\varepsilon - \beta(\rho_\varepsilon))\operatorname{div}\mathbf{u}_\varepsilon \\ &\qquad = \varepsilon \operatorname{div}(\chi_\Omega \nabla \beta(\rho_\varepsilon)) - \varepsilon \chi_\Omega \beta''(\rho_\varepsilon)|\nabla\rho_\varepsilon|^2 \quad \text{in} \quad \mathcal{D}'((0,T)\times\mathbb{R}^3) \end{aligned}$$

with $\beta \in C^2[0,\infty), \beta(0)=0$, a convex function, β', β'' bounded, and χ_Ω the characteristic function on Ω. By a suitable approximation of $z \to z\log z$ by smooth convex functions we deduce for $\varepsilon \to 0$ (see also [14], [13])

$$\mathcal{B}_{dft}[\rho_\varepsilon - \rho] \le \int_0^\tau \int_\Omega [\, \rho\operatorname{div}\mathbf{u} - \overline{\rho\operatorname{div}\mathbf{u}}\,]dxdt \tag{6.7}$$

for a.e. $\tau \in [0,T]$.

In what follows we employ the *multipliers technique* as in Feireisl [20] and Lions [27], which involves placing the quantities

$$\varphi(t,x) = \psi(t)\eta(x)(\nabla\Delta^{-1})[\rho_\varepsilon], \quad \psi \in \mathcal{D}(0,T), \quad \eta \in \mathcal{D}(\Omega),$$

as test functions in the (approximate) momentum equation. Let us introduce the following quantity:

$$\mathbf{C}(\rho,\mathbf{u},\theta,\rho_1)=p_e(\rho)+p_e(\rho_1)+\frac{R}{m_2}\theta p_\theta(\rho)+L\theta p_\theta(\rho_1)+\delta\rho^\beta-\left(\left(\zeta(\theta)-\frac{2}{3}\right)+2\mu(\theta)\right)\mathrm{div}\mathbf{u}.$$

Using now the smoothing properties of the operator $\{\nabla\Delta^{-1}\}$ we get (see also [14], [13])

$$\begin{aligned}\lim_{\varepsilon\to0}\int_0^T\int_\Omega \psi\eta\left[\mathbf{C}(\rho_\varepsilon,\mathbf{u}_\varepsilon,\theta_\varepsilon,\rho_{1_\varepsilon})\right]\rho_\varepsilon\,dxdt\\ =\int_0^T\int_\Omega\psi\eta\Bigg[\overline{p_e(\rho)}+\overline{p_e(\rho_1)}+\frac{R}{m_2}\theta\overline{p_\theta(\rho)}+L\theta\overline{p_\theta(\rho_1)}+\delta\overline{\rho^\beta}\\ -\left(\left(\zeta(\theta)-\frac{2}{3}\right)+2\mu(\theta)\right)\mathrm{div}\mathbf{u}\Bigg]\rho\,dxdt\\ +\left[\left(I^1-\lim_{\epsilon\to0}I^1_\epsilon\right)+2\left(\lim_{\epsilon\to0}I^2_\epsilon-I^2\right)\right]\end{aligned}\tag{6.8}$$

with

$$\begin{aligned}I^1&=\int_0^T\int_\Omega\psi\eta\mathbf{u}\cdot(\rho\mathcal{R}[\rho\mathbf{u}]-\mathcal{R}[\rho](\rho\mathbf{u}))\,dxdt,\\ I^1_\epsilon&=\int_0^T\int_\Omega\psi\eta\mathbf{u}_\varepsilon\cdot(\rho\mathcal{R}[\rho_\varepsilon\mathbf{u}_\varepsilon]-\mathcal{R}[\rho_\varepsilon](\rho_\varepsilon\mathbf{u}_\varepsilon))\,dxdt,\\ I^2&=\int_0^T\int_\Omega\psi(\mathcal{R}[\eta\mu(\theta)\nabla\mathbf{u}]-\eta\mu(\theta)\mathcal{R}[\nabla\mathbf{u}])\rho\,dxdt,\\ I^2_\epsilon&=\int_0^T\int_\Omega\psi(\mathcal{R}[\eta\mu(\theta_\varepsilon)\nabla\mathbf{u}_\varepsilon]-\eta\mu(\theta_\varepsilon)\mathcal{R}[\nabla\mathbf{u}_\varepsilon])\rho_\varepsilon\,dxdt,\end{aligned}$$

where

$$\mathcal{R}[\mathbb{A}]=\sum_{i,j}\mathcal{R}_{i,j}[A_{i,j}],\quad \mathcal{R}=\mathcal{R}_{i,j}[v]=\mathcal{F}^{-1}_{\xi\to x}\left[\frac{\xi_i\xi_j}{|\xi|^2}\mathcal{F}_{x\to\xi}[v]\right].$$

Using now the continuity property of the bilinear form

$$[v,\mathbf{w}]\to v\mathcal{R}[\mathbf{w}]-\mathcal{R}[v]\mathbf{w},$$

one obtains as in [22], [20], [27] that

$$\lim_{\varepsilon\to0}I^1_\varepsilon=I^1.$$

The convergence

$$\lim_{\varepsilon\to0}I^2_\varepsilon=I^2$$

is obtained following the analysis presented in [20], [15] in the spirit of Coifman and Meyer [8].

Now relation (6.8), together with the strong convergence of $\{\theta_\varepsilon\}$, yields

$$\rho\mathrm{div}\mathbf{u}-\overline{\rho\mathrm{div}\mathbf{u}}\le\frac{1}{\zeta(\theta)-\frac{2}{3}+\mu(\theta)}(\mathcal{P}_1+\mathcal{P}_2+\mathcal{P}_3),$$

where

$$\mathcal{P}_1 = \overline{p_e(\rho)}\rho - \overline{p_e(\rho)\rho} + \overline{p_e(\rho_1)}\rho - \overline{p_e(\rho_1)\rho},$$

$$\mathcal{P}_2 = \frac{R\theta}{m_2}(p_\theta(\rho) - \overline{p_\theta(\rho)}) + L\theta(p_\theta(\rho_1) - \overline{p_\theta(\rho_1)}), \quad \mathcal{P}_3 = \overline{\rho^\beta}\rho - \overline{\rho^{\beta+1}}.$$

We can see immediately that $\mathcal{P}_2 \leq 0$, $\mathcal{P}_3 \leq 0$, while

$$\mathcal{P}_1 \leq \overline{p_b(\rho)}\rho - \overline{p_b(\rho)\rho} + \overline{p_b(\rho_1)}\rho - \overline{p_b(\rho_1)\rho},$$

where p_b is the bounded, nonmonotone component of the *elastic pressure*. Using (6.7) we get

$$\int_\Omega \left(\overline{\rho \log(\rho)} - \rho \log(\rho)\right)(s)dx \leq \frac{1}{\underline{\mu}} \int_0^s \int_\Omega \overline{p_b(\rho)}\rho - \overline{p_b(\rho)\rho}\, dxdt,$$

which yields

$$\mathcal{B}_{dft}[\rho_\varepsilon - \rho] \leq \frac{\Lambda}{\underline{\mu}} \int_0^\tau \mathcal{B}_{dft}[\rho_\varepsilon - \rho] dx,$$

and consequently

$$\mathcal{B}_{dft}[\rho_\varepsilon - \rho] = 0,$$

which implies that

$$\rho_\varepsilon \longrightarrow \rho \quad \text{in } L^1((0,T) \times \Omega). \tag{6.9}$$

6.4. The limit process in the field equations ($\varepsilon \to 0$). We are now ready to let $\varepsilon \to 0$ in the field equations. By the aid of the estimates obtained above, we have, passing to a subsequence if needed, that

$$\begin{cases} \rho_\varepsilon \to \rho \quad \text{in} \quad C([0,T]; L^\beta(\Omega)), \\ \mathbf{u}_\varepsilon \to \mathbf{u} \quad \text{weakly in} \quad L^2(0,T; W_0^{1,2}(\Omega; \mathbb{R}^3)), \\ \varepsilon \mathrm{div}(\chi_\Omega \nabla \rho_\varepsilon) \to 0 \quad \text{in} \quad L^2(0,T; W^{-1,2}(\mathbb{R}^N)); \end{cases} \tag{6.10}$$

therefore, the limit functions ρ, $\mathbf{u}$ satisfy the continuity equation in the sense of distributions. From the energy estimates established above, we have

$$\begin{cases} \varepsilon \nabla \mathbf{u}_\varepsilon \nabla \rho_\varepsilon \to 0 & \text{in } L^1(0,T; L^1(\Omega)), \\ \varepsilon \mathrm{div}(\chi_\Omega \nabla \rho_{1_\varepsilon}) \to 0 & \text{in } L^2(0,T; W^{-1,2}(\mathbb{R}^N)). \end{cases}$$

Keeping in mind the estimates (6.10), we get

$$\rho_\varepsilon \mathbf{u}_\varepsilon \to \rho \mathbf{u}, \quad \rho_{1_\varepsilon} \mathbf{u}_\varepsilon \to \rho_1 \mathbf{u} \quad \text{in } C\left([0,T]; L_{weak}^{\frac{2\beta}{\beta+1}}(\Omega)\right).$$

The limit functions ρ, $\mathbf{u}$, θ, and ρ_1 satisfy in the sense of distributions $\mathcal{D}'((0,T) \times \Omega)$ the momentum equation

$$\partial_t(\rho \mathbf{u}) + \mathrm{div}(\rho \mathbf{u} \otimes \mathbf{u}) + \nabla \left(\overline{p_e(\rho)} + \frac{R}{m_2}\theta \overline{p_\theta(\rho)} + \frac{a}{3}\theta^4 + LR\theta p_\theta(\overline{\rho_1}) + \delta \overline{\rho^\beta}\right) = \mathrm{div}\mathbb{S}.$$

The *species conservation equation* is also verified in $\mathcal{D}'((0,T)\times\Omega)$ by the limit function sequences ρ, $\mathbf{u}$, θ, and ρ_1 by virtue of the estimates above. Now passing to the limit in the energy equality (5.19) comes as a consequence of the estimates established earlier and so we recover the total energy balance,

(6.11)

$$-\int_0^T\int_\Omega \partial_t\psi\left(\frac{1}{2}\rho|\mathbf{u}|^2+\rho P_e(\rho)+\rho_1 P_e(\rho_1)+\frac{\delta}{\beta-1}\rho^\beta+a\theta^4+\frac{c_v}{m_2}\rho\theta+Lc_v\,\rho_1\theta\right)dxdt$$
$$=\int_\Omega\left(\frac{1}{2}\frac{\mathbf{m}_{0,\delta}}{\rho_{0,\delta}}+\rho_{0,\delta}P_e(\rho_{0,\delta})+\rho_{1_{0,\delta}}P_e(\rho_{1_{0,\delta}})+\frac{\delta}{\beta-1}\rho_{0,\delta}^\beta+a\theta_{0,\delta}^4\right.$$
$$\left.+\frac{c_v}{m_2}\rho_{0,\delta}+Lc_v\,\rho_{1_{0,\delta}}\theta_{0,\delta}\right)dx$$

for any $\psi\in C^\infty[0,T]$, $\psi(0)=1$, $\psi(T)=0$, $\partial_t\psi\le 0$. Similarly, sending $\varepsilon\to 0$ in (5.20),

$$\begin{aligned}&\int_0^T\int_\Omega \partial_t\varphi[\rho(s_F+s_{f_1})]+\nabla\varphi[\rho(s_F+s_{f_1})]-\left(\frac{\kappa_C(\theta)+\sigma\theta^3}{\theta}\nabla\theta\right)\nabla\varphi dxdt\\
&\le\int_0^T\int_\Omega\varphi\left(\frac{S:\nabla\mathbf{u}}{\theta}+\frac{\kappa_C(\theta)+\sigma\theta^3}{\theta^3}|\nabla\theta|^2\right)dxdt\\
&\quad-\int_0^T\int_\Omega\varphi\left(\frac{c_v}{m_2}(\log(\theta)-1)w_1(\rho_1,\rho_2,\theta)\right)dxdt\\
&\quad-\int_\Omega\varphi(0)[\rho_{0,\delta}(s_{F_{0,\delta}}+s_{f_{1_{0,\delta}}})]dx\end{aligned}\tag{6.12}$$

for any test function φ, $\varphi\in C^\infty([0,T]\times\Omega)$, $\varphi\ge 0$, $\varphi(T)=0$.

7. Passing to the limit in the artificial pressure term. In this last part we pass to the limit for $\delta\to 0$ in the sequence ρ_δ, $\mathbf{u}_\delta$, θ_δ, ρ_{1_δ} of the approximate solutions constructed in the previous section.

Taking into consideration the energy equality (6.11) we have

$$\begin{cases}\rho_\delta\in L^\infty(0,T;L^\gamma(\Omega)),\\ \sqrt{\rho_\delta}\mathbf{u}_\delta\in L^\infty(0,T;L^2(\Omega)),\end{cases}\qquad\begin{cases}\rho_\delta\theta_\delta\in L^\infty(0,T;L^1(\Omega)),\\ \theta_\delta\in L^\infty(0,T;L^4(\Omega)),\end{cases}\tag{7.1}$$

$$\begin{cases}\rho_{1_\delta}\in L^\infty(0,T;L^\gamma(\Omega)),\\ \sqrt{\rho_{1_\delta}}\mathbf{u}_\delta\in L^\infty(0,T;L^2(\Omega)),\\ \rho_{1,\delta}\theta\in L^\infty(0,T;L^1(\Omega)).\end{cases}\tag{7.2}$$

By applying now the same procedure as in section 5 we get the following refined estimate for ρ_δ:

$$\rho_\delta^{\gamma+\nu}+\delta\rho_\delta^{\beta+\nu}\quad\text{is bounded in } L^1((0,T)\times\Omega),\ \nu>1.\tag{7.3}$$

The relation

$$\int_0^T\int_\Omega\frac{\mathbb{S}_\delta:\nabla\mathbf{u}_\delta}{\theta_\delta}+|\nabla\log(\theta_\delta)|^2+|\nabla\theta_\delta^{\frac{3}{2}}|^2+\varepsilon|\nabla\rho_\delta|^2+||\nabla\log(\rho_{1_\delta}\theta_\delta)|^2|dxdt\le c(\delta)$$

is now valid using the energy estimates obtained in our earlier analysis. This last relation implies, in particular, that

$$\|\mathbf{u}_\delta\|_{L^b(0,T;W_0^{1,b}(\Omega))} \le c(\delta) \quad \text{with } b = \frac{8}{5-\alpha}, \tag{7.4}$$

and in combination with Lemma 5.4,

$$\|\log(\theta_\delta)\|_{L^2(0,T;W^{1,2}(\Omega))} + \|\theta_\delta^{\frac{3}{2}}\|_{L^2(0,T;W^{1,2}(\Omega))} \le \tilde{C},$$

yielding

$$\begin{cases} \theta_\delta^{3/2} \in L^2(0,T;W^{1,2}(\Omega)), \\ \log(\theta_\delta) \in L^2(0,T;W^{1,2}(\Omega)), \end{cases} \qquad \begin{cases} \mathbb{S}_\delta \in L^a(0,T;L^s(\Omega)), \\ \text{with } a = \frac{8}{5-\alpha},\ s = \frac{8}{7-\alpha}. \end{cases} \tag{7.5}$$

The bound of $\mathbb{S}_\delta \in L^a(0,T;L^s(\Omega))$ for the values of α and β specified above is a direct consequence of Hölder's inequality. Therefore,

$$\begin{cases} \rho_\delta \longrightarrow \rho & \text{in} \quad C([0,T], L^\gamma_{weak}(\Omega)), \\ \mathbf{u}_\delta \longrightarrow \mathbf{u} & \text{weakly in} \quad L^b(0,T;W_0^{1,b}(\Omega)), \end{cases} \tag{7.6}$$

where ρ, $\mathbf{u}$ satisfy (1.23) in $\mathcal{D}'((0,T)\times\mathbb{R}^3)$. We can also verify using the earlier analysis that

$$\begin{cases} \rho_\delta \mathbf{u}_\delta \longrightarrow \rho\mathbf{u} & \text{in} \quad C([0,T], L^{\frac{\gamma}{\gamma+1}}(\Omega)), \\ \log(\theta_\delta) \longrightarrow \overline{\log(\theta)} & \text{weakly in} \quad L^2(0,T;W^{1,2}(\Omega)), \\ \rho_\delta \log(\theta_\delta) \longrightarrow \rho\overline{\log(\theta)} & \text{weakly in} \quad L^2(0,T;L^{\frac{6\gamma}{6+\gamma}}(\Omega)), \\ \rho_\delta \log(\theta_\delta)\mathbf{u}_\delta \longrightarrow \rho\overline{\log(\theta)}\mathbf{u} & \text{weakly in} \quad L^2(0,T;L^{\frac{6\gamma}{3+4\gamma}}(\Omega)), \end{cases} \tag{7.7}$$

whereas estimates involving ρ_1, namely,

$$\begin{cases} \rho_{1_\delta} \mathbf{u}_\delta \longrightarrow \rho_1\mathbf{u} & \text{in} \quad C([0,T], L^{\frac{\gamma}{\gamma+1}}(\Omega)), \\ \rho_{1_\delta} \log(\theta_\delta) \longrightarrow \rho_1\overline{\log(\theta)} & \text{weakly in} \quad L^2(0,T;L^{\frac{6\gamma}{6+\gamma}}(\Omega)), \\ \rho_{1_\delta} \log(\theta_\delta)\mathbf{u}_\delta \longrightarrow \rho_1\overline{\log(\theta)}\mathbf{u} & \text{weakly in} \quad L^2(0,T;L^{\frac{6\gamma}{3+4\gamma}}(\Omega)), \end{cases} \tag{7.8}$$

are obtained using (7.2), (7.4), and (7.5).

7.1. Pointwise convergence of the temperature. Next we show the strong convergence of the temperature sequence. In analogy with the above analysis, the entropy inequality (6.12) in combination with Lemma 5.5 yields

$$\begin{cases} \frac{4a}{3}\theta_\delta^3 + \frac{c_v}{m_2}\rho_\delta \log(\theta_\delta) - \rho_\delta P_\theta(\rho_\delta) + Lc_v\rho_{1_\delta}\log(\theta_\delta) - LR\rho_{1_\delta}P_\theta(\rho_{1_\delta}) \\ \qquad\qquad\qquad\qquad\qquad\qquad \downarrow \\ \frac{4a}{3}\overline{\theta^3} + \frac{c_v}{m_2}\rho\overline{\log(\theta)} - \overline{\rho_\delta P_\theta(\rho_\delta)} + Lc_v\overline{\rho_1\log(\theta)} - LR\overline{\rho_1 P_\theta(\rho_1)} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{in} \quad L^2(0,T;W^{-1,2}(\Omega)). \end{cases}$$

We remark that in the present context

$$\begin{cases} \rho_\delta\, P_\theta(\rho_\delta) \quad \text{is bounded in} \quad L^\infty(0,T;L^{\frac{\gamma}{G}}(\Omega)) \quad \text{with} \quad \frac{\gamma}{G} > \frac{4}{3}, \\ \rho_{1_\delta}\, P_\theta(\rho_\delta) \quad \text{is bounded in} \quad L^\infty(0,T;L^{\frac{\gamma}{G}}(\Omega)) \quad \text{with} \quad \frac{\gamma}{G} > \frac{4}{3}. \end{cases} \tag{7.9}$$

In particular, we have

$$\begin{cases} \int_0^T\int_\Omega \left(\frac{4a}{3}\theta_\delta^3 + \frac{c_v}{m_2}\rho_\delta \log(\theta_\delta) - \rho_\delta P_\theta(\rho_\delta) + Lc_v\rho_{1_\delta}\log(\theta_\delta) - LR\rho_{1_\delta}P_\theta(\rho_{1_\delta})\right)\theta_\delta dxdt, \\ \downarrow \\ \int_0^T\int_\Omega \left(\frac{4a}{3}\overline{\theta^3} + \frac{c_v}{m_2}\rho\overline{\log(\theta)} - \overline{\rho P_\theta(\rho)} + Lc_v\overline{\rho_1\log(\theta)} - LR\overline{\rho_1 P_\theta(\rho_1)}\right)\theta dxdt, \end{cases}$$

where we have used (6.3) and (6.4) to obtain that as $\delta \to 0$

$$\begin{cases} \lim \int_0^T\int_\Omega \rho_\delta \log(\theta_\delta)\theta_\delta dxdt = \int_0^T\int_\Omega \rho\,\overline{\log(\theta)}\theta\, dxdt, \\ \lim \int_0^T\int_\Omega \rho_{1_\delta}\ \log(\theta_\delta)\theta_\delta\, dxdt = \int_0^T\int_\Omega \rho_1\,\overline{\log(\theta)}\theta\, dxdt. \end{cases}$$

In addition, taking into consideration that ρ is a *renormalized solution* of the continuity equation and using a standard approximation argument, we get

$$\begin{cases} \rho_\delta P_\theta(\rho_\delta) \to \overline{\rho P_\theta(\rho)} & \text{in} \quad C([0,T]; L^{\frac{\gamma}{G}}_{weak}(\Omega)), \\ \rho_{1_\delta}P_\theta(\rho_{1_\delta}) \to \overline{\rho_1 P_\theta(\rho_1)} & \text{in} \quad C([0,T]; L^{\frac{\gamma}{G}}_{weak}(\Omega)). \end{cases}$$

Therefore (6.3), (6.4) yield

$$\begin{cases} \int_0^T\int_\Omega \left(\frac{4a}{3}\theta_\delta^3 + \frac{c_v}{m_2}\rho_\delta\log(\theta_\delta) + Lc_v\rho_{1_\delta}\log(\theta_\delta)\right)\theta_\delta dxdt, \\ \downarrow \\ \int_0^T\int_\Omega \left(\frac{4a}{3}\overline{\theta^3} + \frac{c_v}{m_2}\rho\overline{\log(\theta)} + Lc_v\overline{\rho_1\log(\theta)}\right)\theta dxdt, \end{cases}$$

which in turn implies

$$\theta_\delta \longrightarrow \theta \quad \text{in } L^2((0,T)\times\Omega).$$

7.2. Pointwise convergence of the density. In order to pass to the limit we need the strong convergence of the density. The main part consists in showing that the oscillation defect measure $osc_{\beta+1}[\rho_\delta \to \rho]$ defined by

$$osc_{\beta+1}[\rho_\delta \to \rho]((0,T)\times\Omega) = \sup_{k\geq 1}\left(\limsup_{\delta\to 0}\int_0^T\int_\Omega |T_k(\rho_\delta) - T_k(\rho)|^{\beta+1}dxdt\right) \tag{7.10}$$

is bounded. Here $T_k(\rho)$ are cutoff functions:

$$T_k(y) = T\left(\frac{y}{k}\right) \qquad \text{with } T \in C^\infty(\mathbb{R}) \text{ a concave function,}$$

$$T(x) = x \quad \text{for} \quad 0 \leq x \leq 1, \quad T(x) = 2 \quad \text{if} \quad y \geq 3.$$

In order to show that $\rho, \mathbf{u}$ represent a *renormalized* solution of (1.23) we have to show that the oscillation defect measure associated with $\{\rho_\delta\}$ is bounded.

Taking into account that $\rho_1 \leq \rho$ we estimate the amplitude of oscillations as in [14] (see also [20]); namely, we write

$$\begin{cases} p_e(\rho) = p_e^{(c)}(\rho) + p_e^{(m)}(\rho) + p_e^{(b)}(\rho), \\ p_e(\rho_1) = p_e^{(c)}(\rho_1) + p_e^{(m)}(\rho_1) + p_e^{(b)}(\rho_1) \end{cases}$$

with $p_e^{(b)}$ uniformly bounded on $[0,\infty)$, $p_e^{(m)}$ nondecreasing, and $p_e^{(b)}$ a convex function satisfying

$$\begin{cases} p^{(c)}(\rho) \geq a\rho^{\gamma} \quad \text{with} \quad a > 0, \\ p^{(c)}(\rho_1) \geq a\rho_1^{\gamma} \quad \text{with} \quad a > 0. \end{cases}$$

Next, taking into account the monotonicity property of some of the quantities present in the pressure we get

$$\text{(7.11)} \quad \begin{cases} \overline{p_\theta(\rho)T_k(\rho)} \geq \overline{p_\theta(\rho)}\ \overline{T_k(\rho)}, \\ \overline{p_\theta(\rho_1)T_k(\rho)} \geq \overline{p_\theta(\rho_1)}\ \overline{T_k(\rho)}, \end{cases} \qquad \begin{cases} \overline{p_e^{(m)}(\rho)T_k(\rho)} \geq \overline{p_e^{(m)}(\rho)}\overline{T_k(\rho)}, \\ \overline{p_e^{(m)}(\rho_1)T_k(\rho)} \geq \overline{p_e^{(m)}(\rho_1)}\overline{T_k(\rho)}, \end{cases}$$

which yield first that

$$\begin{aligned} &\limsup_{\delta\to 0} \int_0^T \int_\Omega |T_k(\rho_\delta) - T_k(\rho)|^{\beta+1}\, dxdt \\ &\quad \leq \limsup_{\delta\to 0} \int_0^T \int_\Omega \Big| [p_e^{(b)}(\rho_\delta) + p_e^{(b)}(\rho_{1_\delta})] \\ &\qquad + \left(\left(\zeta(\theta) - \frac{2}{3} \right) + 2\mu(\theta) \right) \operatorname{div}\mathbf{u}_\delta \Big| |T_k(\rho_\delta) - T_k(\rho)|\, dxdt, \end{aligned}$$

and in what follows taking into consideration the properties (2.8) of the transport coefficients and the estimates derived above and following a line of argument similar to that presented in [14], [20], [13], we have that

$$osc_{\beta+1}[\rho_\delta \to \rho]((0,T)\times\Omega) < \infty.$$

Now we use the fact that $\rho, \mathbf{u}$ represent a *renormalized solution* of (1.23) on $(0,T)\times\Omega$ (cf. Proposition 6.3 in [20]).

PROPOSITION 7.1. *Let $\Omega \subset \mathbb{R}^N$ be a domain. Assume that $\rho_\delta \geq 0$, and that $\mathbf{u}_\delta$ is a sequence of renormalized solutions to* (1.23) *on $(0,T)\times\Omega$ such that*

$$\text{(7.12)} \qquad \begin{cases} \rho_\delta \to \rho \quad \textit{weakly-* in} \quad L^\infty(0,T;L^\gamma(\Omega)), \quad \gamma > 1, \\ \mathbf{u}_\delta \to \mathbf{u} \quad \textit{weakly in} \quad L^r(0,T;W^{1,r}(\Omega;\mathbb{R}^N)), \quad r > 1, \end{cases}$$

where

$$\gamma > \frac{Nr}{(N+1)r - N} \quad \textit{if} \quad r < N.$$

Furthermore, assume that

$$osc_p[\rho_\delta \to \rho]((0,T)\times\Omega) < \infty$$

for a certain p such that

$$\frac{1}{p} + \frac{1}{r} < 1.$$

Then the limit functions $\rho, \mathbf{u}$ represent a renormalized solution of (1.23) *on* $(0,T)\times\Omega$.

Note that in our case $r = \frac{8}{5-\alpha} > \frac{3}{2}$ and so all the requirements of Proposition 7.1 are valid. Therefore we get that

$$\rho_\delta \longrightarrow \rho \quad \text{strongly in } L^1((0,T)\times\Omega). \tag{7.13}$$

Property (7.10) implies now that the continuity equation (1.23) holds true in the sense of distribution. Furthermore, the bound (7.3) yields that

$$\delta\rho^\beta \longrightarrow 0 \quad \text{in } L^{\frac{\beta+\nu}{\beta}}((0,T)\times\Omega),$$

and the momentum equation (1.24) is recovered as $\delta \to 0$.

The *species conservation equation* now can be also verified using Proposition 7.1 and the estimates in (7.2). In addition, the strong convergence of the density sequence ρ_δ, together with the estimates established above, allows us to pass to the limit both in the energy equality (6.11) and in the entropy inequality (6.12). The proof of Theorem 3.1 has now been established.

Acknowledgment. The author gratefully acknowledges the hospitality and support of Mittag–Leffler Institute, where this research was performed during the semester program on "Wave Motion."

REFERENCES

[1] E. Becker, *Gasdynamik,* Teubner-Verlag, Stuttgart, Germany, 1966.
[2] S. Benzoni-Gavage, *Nonuniqueness of phase transitions near the Maxwell line*, Proc. Amer. Math. Soc., 127 (1999), pp. 1183–1190.
[3] G.-Q. Chen, *Global solutions to the compressible Navier–Stokes equations for a reacting mixture*, SIAM J. Math. Anal., 23 (1992), pp. 609–634.
[4] J. P. Cox and R. T. Giuli, *Principles of Stellar Structure* I, II, Gordon and Breach, New York, 1968.
[5] G.-Q. Chen, D. Hoff, and K. Trivisa, *Global solutions of the compressible Navier–Stokes equations with large discontinuous initial data*, Comm. Partial Differential Equations, 25 (2000), pp. 2233–2257.
[6] G.-Q. Chen, D. Hoff, and K. Trivisa, *On the Navier–Stokes equations for exothermically reacting compressible fluids*, Acta Math. Appl. Sin. (Engl. Ser.), 18 (2002), pp. 15–36.
[7] G.-Q. Chen, D. Hoff, and K. Trivisa, *Global solutions to a model for exothermically reacting, compressible flows with large discontinuous initial data*, Arch. Ration. Mech. Anal., 166 (2003), pp. 321–358.
[8] R. Coifman and Y. Meyer, *On commutators of singular integrals and bilinear singular integrals*, Trans. Amer. Math. Soc., 212 (1975), pp. 315–331.
[9] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren Math. Wiss. 325, Springer-Verlag, Berlin, 2000.
[10] R. J. DiPerna and P.-L. Lions, *On the Fokker–Planck–Bolzmann Equation*, Comm. Math. Phys., 120 (1988), pp. 1–23.
[11] R. J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory, and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.
[12] D. Donatelli and K. Trivisa, *A multidimensional model for the combustion of compressible fluids*, Arch. Ration. Mech. Anal., 185 (2007), pp. 379–408.
[13] D. Donatelli and K. Trivisa, *On the motion of a viscous, compressible, radiative-reacting gas*, Comm. Math. Phys., 265 (2006), pp. 463–491.
[14] B. Ducomet and E. Feireisl, *On the dynamics of gaseous stars*, Arch. Ration. Mech. Anal., 174 (2004), pp. 221–266.
[15] B. Ducomet and E. Feireisl, *A regularizing effect of radiation in the equations of fluid dynamics*, Math. Methods Appl. Sci., 28 (2005), pp. 661–685.
[16] A. Ern and V. Giovangigli, *Multicomponent Transport Algorithms*, Lecture Notes in Physics. New Series m: Monographs 24, Springer-Verlag, Berlin, 1994.
[17] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.
[18] H. Fan, *On a model of the dynamics of liquid/vapor phase transitions*, SIAM J. Appl. Math., 60 (2000), pp. 1270–1301.

[19] H. Fan and M. Slemrod, *Dynamic flows with liquid/vapor phase transitions*, in Handbook of Mathematical Fluid Dynamics, Vol. I, S. J. Freidlander and D. Serre, eds., North–Holland, Amsterdam 2002, pp. 373–420.

[20] E. Feireisl, *Dynamics of Viscous Compressible Fluids*, Oxford University Press, Oxford, UK, 2004.

[21] E. Feireisl, *On the motion of a viscous, compressible, and heat conducting fluid*, Indiana Univ. Math. J., 53 (2004), pp. 1705–1738.

[22] E. Feireisl, A. Novotný, and H. Petzeltová, *On the existence of globally defined weak solutions to the Navier–Stokes equations*, J. Math. Fluid Mech., 3 (2001), pp. 358–392.

[23] E. Feireisl, H. Petzeltová, and K. Trivisa, *Multicomponent Reactive Flows: Global-in-Time Existence for Large Data*, Nečas Center for Mathematical Modeling, preprint (2007).

[24] E. Feireisl and A. Novotný, *On a simple model of reacting compressible flows arising in astrophysics*, Proc. Roy. Soc. Edinburgh Sect. A, 135 (2005), pp. 1169–1194.

[25] V. Giovangigli, *Multicomponent Flow Modeling*, Model. Simul. Sci. Eng. Technol., Birkhäuser, Boston, 1999.

[26] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Uraltseva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[27] P.-L. Lions, *Mathematical Topics in Fluid Mechanics, Vol.* 2, The Clarendon Press, Oxford University Press, New York, 1998.

[28] J. Oxenius, *Kinetic Theory of Particles and Photons*, Springer Ser. Electrophys. 20, Springer-Verlag, Berlin, 1986.

[29] M. Slemrod, *Admissibility criteria for propagating phase boundaries in a van der Waals fluid*, Arch. Ration. Mech. Anal., 81 (1983), pp. 301–315.

[30] G. S. Springer, *Homogeneous nucleation,* Adv. in Heat Transfer, 14 (1978), pp. 281–345.

[31] K. Trivisa, *On Multicomponent Flow Models*, in preparation.

[32] C. Truesdell, *Rational Thermodynamics*, Springer-Verlag, New York, 1984.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 1821–1850

REACHABLE AND UNREACHABLE SETS IN THE SCATTERING PROBLEM FOR THE ACOUSTICAL EQUATION IN $\mathbb{R}^3$*

MIKHAIL I. BELISHEV† AND ALEXSEI F. VAKULENKO†

Abstract. The scattering problem is to find $u = u^f(x,t)$ satisfying $u_{tt} - \Delta u + qu = 0$, $(x,t) \in \mathbb{R}^3 \times (-\infty, \infty)$; $u\mid_{|x|<-t} = 0$, $t < 0$; $\lim_{s\to\infty} su((s+\tau)\omega, -s) = f(\tau,\omega)$, $(\tau,\omega) \in [0,\infty) \times S^2$ for a real-valued smooth compactly supported potential $q = q(x)$ and a control $f \in \mathcal{F} = L_2([0,\infty); L_2(S^2))$. The corresponding control problem is as follows: given $y \in \mathcal{H} = L_2(\mathbb{R}^3)$ find $f \in \mathcal{F}$ provided $u^f(\cdot,0) = y$; the reachable set is $\mathcal{U} = \{u^f(\cdot,0) \mid f \in \mathcal{F}\}$; the subspace of unreachable states is $\mathcal{D} = \mathcal{H} \ominus \mathcal{U}$. The main subject of the paper is the structure of $\mathcal{U}$ and $\mathcal{D}$. We present an example of the finite energy solution u^f satisfying $u^f|_{|x|<|t|} = 0$, i.e., vanishing simultaneously in the past and future cones (reversing wave) and we introduce the set of points at which such a "reverse effect" occurs. The existence of reversing waves turns out to be equivalent to the lack of controllability $\mathcal{D} \neq \{0\}$. Cauchy data of such waves belong to the classes $D_\mp$ of the incoming and outgoing data simultaneously; in other words, $D_- \cap D_+ \neq \{0\}$. Also, simple conditions on f ensuring $\|u^f(\cdot,t)\|_{\mathcal{H}} \le c\|f\|_{\mathcal{F}}$ for all $t \in (-\infty,\infty)$ are described. We plan to apply these results to the dynamical (time-domain) inverse problem, that is, determination of potential from the dynamical scattering data. The study of controllability is the first step towards solving this problem by the boundary control method.

Key words. 3D acoustical equation, time-domain scattering problem, control problem, reachable sets, reversing waves, stop points, stability conditions

AMS subject classifications. 35B, 35L, 35P25, 47A

DOI. 10.1137/060678877

Introduction. In this section, we first introduce related notions, then formulate the problem, state the main results, and finally discuss the motivation and give some other comments.

0.1. Dynamical system. We denote $B_r := \{x \in \mathbb{R}^3 \mid |x| < r\}$, $S^2 := \{\theta \in \mathbb{R}^3 \mid |\theta| = 1\}$, $\Sigma := (0,\infty) \times S^2$. Let $q = q(x)$, $x \in \mathbb{R}^3$, be a real-valued smooth (everywhere in the paper "smooth" means C^∞-smooth) compactly supported function (*potential*), $\operatorname{supp} q \subset \overline{B}_{r_*}$. The system under consideration is

$$u_{tt} - \Delta u + qu = 0, \qquad (x,t) \in \mathbb{R}^3 \times (-\infty, 0), \tag{0.1}$$

$$u\mid_{|x|<-t} = 0, \qquad t < 0, \tag{0.2}$$

$$\lim_{s\to+\infty} su((s+\tau)\theta, -s) = f(\tau,\theta), \qquad (\tau,\theta) \in \overline{\Sigma}, \tag{0.3}$$

where f is a control, $u = u^f(x,t)$ is a solution (*wave*). The value $t = 0$ is referred to as a final moment; however, in what follows we also deal with problem (0.1)–(0.3) for all $t \in (-\infty, \infty)$.

*Received by the editors December 28, 2006; accepted for publication (in revised form) August 28, 2007; published electronically February 15, 2008. This work is supported by RFBR grants 05.01.00936 and NSh–8336.2006.1, and by the Sweden Academy of Sciences.
http://www.siam.org/journals/sima/39-6/67887.html

†Department of Mathematics, Steklov Mathematical Institute, Russian Academy of Sciences, St. Petersburg 191023, Russia (belishev@pdmi.ras.ru, vak@pdmi.ras.ru).

0.2. Spaces, sets, operators. With a dynamical system (0.1)–(0.3) one associates the following:

- an *outer space* of controls $\mathcal{F} := L_2([0,\infty); L_2(S^2))$ and its subspaces
$$\mathcal{F}^\xi := \{f \in \mathcal{F} \mid \operatorname{supp} f \subset [\xi,\infty)\}, \quad \xi \geq 0,$$
consisting of the *delayed controls*;
- an *inner space* of states $\mathcal{H} := L_2(\mathbb{R}^3)$ (we consider a wave $u^f(\cdot,t)$ as a time-dependent element of $\mathcal{H}$) and its subspaces
$$\mathcal{H}^\xi := \left\{y \in \mathcal{H} \mid \operatorname{supp} y \subset \mathbb{R}^3 \setminus B_\xi\right\}, \quad \xi \geq 0;$$
- a *control operator* $W : \mathcal{F} \to \mathcal{H}$
$$Wf := u^f(\cdot,0)$$
and the subspace of *null controls*
$$\mathcal{N} := \operatorname{Ker} W = \left\{f \in \mathcal{F} \mid u^f(\cdot,0) = 0\right\};$$
- the *reachable sets*
$$\mathcal{U}^\xi := W\mathcal{F}^\xi = \left\{u^f(\cdot,0) \mid f \in \mathcal{F}^\xi\right\}, \quad \xi \geq 0$$
(we also denote $\mathcal{U}^0$ by $\mathcal{U}$); by hyperbolicity of problem (0.1)–(0.3) one has $\mathcal{U}^\xi \subset \mathcal{H}^\xi$;
- the defect subspaces (*unreachable sets*)
$$\mathcal{D}^\xi := \mathcal{H}^\xi \ominus \mathcal{U}^\xi, \quad \xi \geq 0$$
(we denote $\mathcal{D}^0 =: \mathcal{D} = \mathcal{H} \ominus \mathcal{U}$).

We introduce also the *polyharmonic subspaces* $\mathcal{A}^\xi \subset \mathcal{H}^\xi$,
$$\mathcal{A}^\xi := \operatorname{clos}\ \left\{a \in \mathcal{H}^\xi \mid \exists \,\text{integer}\, p \geq 1 :\ (-\Delta + q)^p\, a = 0 \ \text{ in } \mathbb{R}^3 \setminus \overline{B}_\xi\right\}, \quad \xi \geq 0$$
(the closure in $\mathcal{H}$; $\overline{B}_0 := \{0\}$), and denote $\mathcal{A} := \mathcal{A}^0$. It is easy to see that $\dim \mathcal{A}^\xi = \infty$ for $\xi > 0$.

The evolution of system (0.1)–(0.3) is governed by a self-adjoint operator $H : \mathcal{H} \to \mathcal{H}$, $\operatorname{Dom} H = H^2(\mathbb{R}^3)$, $Hy = (-\Delta + q)y$. This operator can have at most a finite number of negative eigenvalues of finite multiplicity, whereas its absolutely continuous spectrum fills $[0,\infty)$. The spectral point $\lambda = 0$ can also belong to the point spectrum $\sigma_p(H)$, so that $\operatorname{Ker} H$ can be a nonzero finite-dimensional subspace in $\mathcal{H}$ (see [14]).

0.3. The results. A control problem for system (0.1)–(0.3) is set up as follows: given $y \in \mathcal{H}$ find $f \in \mathcal{F}$ such that

$$u^f(\cdot,0) = y. \tag{0.4}$$

Loosely speaking, we deal with a "wave shaping": managing the control, one needs to create the incoming (from infinity) wave of prescribed shape. We consider also the same problem for delayed controls $f \in \mathcal{F}^\xi$.

To study the solvability of (0.4) is to investigate the structure of the reachable sets $\mathcal{U}, \mathcal{U}^\xi$ or, equivalently, of the defect subspaces $\mathcal{D}, \mathcal{D}^\xi$. Our main result is the following.

THEOREM 0.1. (i) *The sets* $\mathcal{U}^\xi$ *are closed. The relations*

$$\mathcal{A}^\xi \subseteq \mathcal{D}^\xi, \quad \xi \geq 0; \quad \mathrm{Ker} H \subseteq \mathcal{A} \subseteq \mathcal{D} \tag{0.5}$$

and

$$\dim \mathrm{Ker} H \leq \dim \mathcal{D} = \dim \mathcal{N} < \infty \tag{0.6}$$

hold.

(ii) *If* $||q||_{L^\infty(B_{r_*})}$ *is small enough, then* $\mathcal{U} = \mathcal{H}$ *and* $\mathcal{D} = \{0\}$.

(iii) *If the potential is radially symmetric, i.e.,* $q = q(|x|)$, *then*

$$\mathcal{D}^\xi = \mathcal{A}^\xi, \quad \xi \geq 0; \quad \mathcal{D} = \mathrm{Ker}\, H. \tag{0.7}$$

In control theory the equality $\mathcal{U} = \mathcal{H}$ is interpreted as an *exact controllability* of system (0.1)–(0.3). Thus, for small potentials the controllability occurs, whereas in the case of radial q the only obstacle for controllability is $0 \in \sigma_p(H)$. The latter is not true in the general case: we construct an example of (not radial) potential such that $0 \notin \sigma_p(H)$ (i.e., $\mathrm{Ker}\, H = \{0\}$) but $\mathcal{D} \neq \{0\}$, realizing the strict inequality in (0.6).

The presence of the defect $\mathcal{D} \neq \{0\}$ is equivalent to the existence of null controls producing the waves, which exhibit rather curious behavior. If $f \in \mathcal{N}$, then u^f is an L_2-solution of (0.1)–(0.3) satisfying $u^f(\cdot, 0) = 0$. As a consequence, the function

$$w^f(x,t) := \begin{cases} \int_{-\infty}^{t} u^f(x,s)\, ds, & t \leq 0, \\ \int_{-\infty}^{-t} u^f(x,s)\, ds, & t > 0, \end{cases}$$

turns out to be a finite (positive) energy solution of the acoustical equation (0.1) in $\mathbb{R}^3 \times (-\infty, \infty)$ satisfying

$$w^f \mid_{|x|<|t|} = 0, \quad w^f_t \mid_{t=0} = 0, \quad w^f(\cdot, -t) = w^f(\cdot, t),$$

i.e., vanishing in the past and future characteristic cones simultaneously. Thus, w^f is a wave, which comes from infinity, stops at the moment $t = 0$, and then returns back to infinity along the same trajectory. An amazing peculiarity of such a behavior is that this wave possesses a back front: it leaves $\mathrm{supp}\, q$ with no residual perturbation. A noticeable fact is that the existence of such reversing waves is equivalent to the lack of controllability of the system.

The Cauchy data of the reversing wave $\{w^f(\cdot, 0), w^f_t(\cdot, 0)\} = \{y, 0\}$ belong to the energy class (see [10], [13]). At the same time, since $w^f \mid_{|x|<|t|} = 0$, the pair $\{y, 0\}$ belongs to the classes D_- and D_+ of the incoming and outgoing data simultaneously and, thus, provides an example of $D_+ \cap D_- \neq \{0\}$. A noticeable fact is that such an intersection is possible even in the case $\mathrm{Ker}\ H = \{0\}$. Perhaps this example is new: we did not succeed in finding analogues in literature.

The origin $x = 0$ plays the role of a point which stops the reversing waves. We introduce the set $\Upsilon[q] \subset \mathbb{R}^3$ of such points and hope that it proves to be an interesting object for further investigations.

At the end of the paper we present the orthogonality conditions on a control f ensuring L_2-stability of the trajectory: $\|u^f(\cdot, t)\|_{\mathcal{H}} \leq c\|f\|_{\mathcal{F}}$ holds for all $t \in (-\infty, \infty)$ if and only if $(f, g_k)_{\mathcal{F}} = 0$, $k = 1, \ldots, p$, where g_k are determined by the negative spectrum of H in an explicit way.

0.4. Motivation and comments.

- In the future, we plan to apply the results of this paper to the dynamical (time-domain) inverse problem that is to recover the potential q from the response operator R of system (0.1)–(0.3) (see section 2.4). By hyperbolicity of the system, this operator depends *locally* on potential: its part $R\mid_{\mathcal{F}^\xi}$ is determined by the restriction $q\mid_{\mathbb{R}^3\setminus B_\xi}$. Correspondingly, the recovering procedure should be local in the following sense: given $R\mid_{\mathcal{F}^\xi}$ it must recover $q\mid_{\mathbb{R}^3\setminus B_\xi}$. In other words, knowing the response of the system on delayed controls $f\in\mathcal{F}^\xi$, one must determine the potential in the domain $\mathbb{R}^3\setminus B_\xi$ filled with delayed incoming waves $u^f(\cdot,0)$. Such a locality is a specific feature of the *boundary control (BC) method*, which is an approach to inverse problems based upon their connections with control theory [2]. It is the approach, which motivates the study of $\mathcal{U}^\xi$ and $\mathcal{D}^\xi$: by philosophy of the BC method, analysis of controllability of a system is the first step towards solving the corresponding inverse problem.
- Dealing with the scattering problem of such a kind, one cannot avoid parallels to the Lax–Phillips theory (LPT) and we do comment on such parallels in the text. However, our presentation is independent. One of the reasons is that at the moment we do not fully recognize the place of our central object $\mathcal{D}^\xi$ in the context of LPT. Also, we deal mainly with L_2-solutions u^f (not with a complete state $\{u^f,u^f_t\}$) since for the incoming waves, the component u^f_t is determined by u^f. In addition, we prefer L_2-norms avoiding the indefinite energy $\int_{\mathbb{R}^3}u_t^2+|\nabla u|^2+qu^2$.
- Our paper extends the results of [5], which deals with $q=0$, to the perturbed case.

1. Unperturbed system. Here, dealing mainly with the case $q=0$, we present some results of [5] and supplement them with geometrical optics relations used in what follows.

1.1. Expansions over spherical harmonics. Let $Y_l^m(\theta)$, $l\geq 0$, $m=-l,-l+1,\dots,0,\dots,l-1,l$, be the standard spherical functions (*harmonics*) satisfying

$$-\Delta_\omega Y_l^m=l(l+1)Y_l^m\quad\text{on}\quad S^2$$

(Δ_ω is the Beltrami–Laplace operator) and constituting an orthonormal basis in $L_2(S^2)$. The unperturbed system is

$$u_{tt}-\Delta u=0,\qquad (x,t)\in\mathbb{R}^3\times(-\infty,0),\tag{1.1}$$

$$u\mid_{|x|<-t}=0,\tag{1.2}$$

$$\lim_{s\to\infty}su((s+\tau)\theta,-s)=f(\tau,\theta),\qquad(\tau,\theta)\in[0,\infty)\times S^2;\tag{1.3}$$

its outer and inner spaces can be represented, respectively, in the form

$$\mathcal{H}=\oplus\sum_{l,m}\mathcal{H}_{lm}\ ,\quad \mathcal{H}_{lm}:=\left\{y=\frac{v(r)}{r}\,Y_l^m(\omega)\ \middle|\ \int_0^\infty v^2(r)dr<\infty\right\}\tag{1.4}$$

(we denote $\sum_{l,m}:=\sum_{l=0}^\infty\sum_{m=-l}^{m=l}$, $r:=|x|$, $\omega:=\frac{x}{|x|}$ and keep these notations in what follows) and

$$\mathcal{F}=\oplus\sum_{l,m}\mathcal{F}_{lm}\ ,\quad \mathcal{F}_{lm}=\left\{f=g(\tau)\,Y_l^m(\omega)\ \middle|\ \int_0^\infty g^2(\tau)dr<\infty\right\}.\tag{1.5}$$

1.2. Partial problems. Choose an $f \in \mathcal{F}$. Separating variables in (1.1)–(1.3) in accordance with decompositions (1.4), (1.5), we represent

$$f(\tau, \theta) = \sum_{l,m} f_{lm}(\tau)\, Y_l^m(\theta) \tag{1.6}$$

and look for the solution $u = u^f(x,t)$ in the form

$$u^f(x,t) = \sum_{l,m} \frac{u_{lm}(r,t)}{r}\, Y_l^m(\omega)\,, \tag{1.7}$$

which reduces (1.1)–(1.3) to a series of *partial problems*

$$u_{tt} - u_{rr} + \frac{l(l+1)}{r^2}\, u = 0, \qquad (r,t) \in (0,\infty) \times (-\infty, 0), \tag{1.8}$$

$$u\,|_{r<-t} = 0, \tag{1.9}$$

$$\lim_{s\to\infty} u(s+\tau, -s) = f(\tau), \qquad \tau \in [0,\infty). \tag{1.10}$$

Solving them for $f = f_{lm}(\tau)$, we get the solution $u = u_{lm}(r,t)$ and arrive at the following results (e.g., see [5]).

(i) The partial problem can be solved explicitly: for any $\phi \in C_0^\infty(0,\infty)$ the function

$$u^f(r,t) = r^l \left(\frac{\partial}{\partial r}\frac{1}{r}\right)^l \phi(r+t) \tag{1.11}$$

satisfies (1.8)–(1.10) with

$$f(\tau) = \left(\frac{d}{d\tau}\right)^l \phi(\tau). \tag{1.12}$$

For controls of the form (1.12), the inclusion $\operatorname{supp} f \subset [a,b] \subset (0,\infty)$ implies

$$\operatorname{supp} u^f(\cdot, t) \subset [a-t, b-t]$$

and, hence, the solution is compactly supported for all times t. One more important fact is that the class of controls

$$L_l := \left(\frac{d}{d\tau}\right)^l C_0^\infty(0,\infty), \quad l = 0, 1, \dots,$$

as well as each of its subclasses

$$L_l^{(k)} := \left(\frac{d}{d\tau}\right)^k L_l\ , \quad k = 0, 1, \dots,$$

are dense in $L_2(0,\infty)$. Note also the density of the classes $\left(\int_0^\tau\right) L_l^{(k)}$ for $l + k > 0$.

(ii) Representations (1.11)–(1.12) determine the map $W^l : f \mapsto u^f(\cdot, 0)$, which is an isometric operator from $L_2(0,\infty)$ to $L_2(0,\infty)$. Its extension to $L_2(0,\infty)$ turns out to be a unitary operator, which we denote also by W^l. The L_2-solutions of (1.8)–(1.10) can be introduced through this operator by

$$u^f(\cdot, t) := W^l T_t f, \quad t \le 0, \tag{1.13}$$

where T_t is a shift operator,

$$(T_t f)(\tau) := \begin{cases} 0, & 0 \le \tau \le |t|, \\ f(\tau - |t|), & \tau \ge |t|. \end{cases}$$

(iii) The reachable sets of system (1.8)–(1.10) are

$$U_l^\xi := W^l L_2(\xi, \infty), \quad \xi \ge 0,$$

where $L_2(\xi,\infty) := T_\xi L_2(0,\infty)$ is the class of delayed controls; the defect subspaces (unreachable sets) are

$$D_l^\xi := L_2(\xi,\infty) \ominus U_l^\xi, \quad \xi \ge 0.$$

The following characterization of the defect is obtained in [5].

Assign a function $a = a(r) \in L_2(0,\infty)$ to a class A_l^ξ if $\operatorname{supp} a \subset [\xi, \infty)$ and there exists an integer $p > 1$ such that

$$\left[-\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2}\right]^p a = 0, \quad r > \xi. \tag{1.14}$$

For an $a \in A_l^\xi$, define $\operatorname{rank} a$ as the minimal p providing (1.14), and put

$$\operatorname{rank} A_l^\xi := \max_{a \in A_l^\xi} \operatorname{rank} a.$$

Then, one has

$$D_0^\xi = A_0^\xi = \{0\}\ , \quad \xi \ge 0 \tag{1.15}$$

(that is commented on as "s-waves are controllable"), whereas for $l \ge 1$ the set A_l^ξ turns out to be a finite-dimensional subspace,

$$\dim A_l^\xi = \operatorname{rank} A_l^\xi = d(l) := \frac{l}{2} + \frac{1-(-1)^l}{4},$$

spanned on a basis

$$\frac{\chi(r-\xi)}{r^l},\ \frac{\chi(r-\xi)}{r^{l-2}}, \dots, \frac{\chi(r-\xi)}{r^{\frac{3+(-1)^l}{2}}}, \tag{1.16}$$

where $\chi(s) := \frac{1}{2}[1 + \operatorname{sign} s]$ is the Heaviside function. As is shown in [5], the equality

$$D_l^\xi = A_l^\xi\ , \quad l \ge 0,\ \xi > 0, \tag{1.17}$$

holds. For $\xi = 0$, the functions (1.16) do not belong to $L_2(0,\infty)$ and, hence,

$$D_l^0 = A_l^0 = \{0\}. \tag{1.18}$$

1.3. Control operator. Returning from the partial problems to system (1.1)–(1.3), we define its L_2-solutions in accordance with (1.6), (1.7), (1.13): for $f \in \mathcal{F}$ we set

$$u^f(x,t) := \sum_{l,m} \frac{1}{r} \left(W^l T_t \Pi_{lm} f\right)(r)\, Y_l^m(\omega), \tag{1.19}$$

where $\Pi_{lm} : \mathcal{F} \to L_2(0,\infty)$,

$$(\Pi_{lm} f)(\tau) := \int_{S^2} f(\tau,\theta)\, Y_l^m(\theta)\, d\theta\,, \quad \tau \geq 0.$$

By this, the control operator of the unperturbed system $W_0 : \mathcal{F} \to \mathcal{H}$, $W_0 f = u^f(\cdot, 0)$ is unitary[1] and the relation

$$u^f(\cdot,t) = W_0 T_t f, \quad t \leq 0, \tag{1.20}$$

holds (see (1.13)). It is easy to verify that an L_2-solution u^f satisfies (1.1) also in the sense of distributions and, as an $\mathcal{H}$-valued function of time (*trajectory*), belongs to the class $C((-\infty,0];\mathcal{H})$. Moreover, considering (1.1)–(1.3) for all times, one has $u^f \in C((-\infty,\infty);\mathcal{H})$. For a smooth control f vanishing near $\tau = 0$, the solution u^f is smooth and can be represented in the well-known form: assuming f extended to $\tau < 0$ by zero one has

$$u^f(x,t) = \frac{1}{2\pi} \int_{S^2} \frac{\partial f}{\partial \tau}(t + x\cdot\theta, \theta)\, d\theta \tag{1.21}$$

(e.g., see [10]). This formula yields

$$(W_0 f)(x) = \frac{1}{2\pi} \int_{S^2} \frac{\partial f}{\partial \tau}(x\cdot\theta, \theta)\, d\theta. \tag{1.22}$$

For the later use, it is convenient to introduce one more special class of controls and the corresponding solutions. Define

$$\mathcal{L}_{lm} := \{f \in \mathcal{F} \mid \ f = g(\tau)\, Y_l^m(\theta), g \in L_l\}$$

(see section 1.2(i)) and consider the class of controls

$$\mathcal{L} := \left[\oplus \sum_{l,m} \mathcal{L}_{lm}\right] \cap C_0^\infty(\Sigma),$$

which is dense in $\mathcal{F}$. Its subclasses

$$\mathcal{L}^{(k)} := \left(\frac{\partial}{\partial \tau}\right)^k \mathcal{L}\,, \quad k = 1,2,\ldots,$$

are also dense. In addition, note that $\mathcal{L}^{(k)} \cap \mathcal{F}^\xi$ is dense in $\mathcal{F}^\xi$ for every $\xi \geq 0$.

If $f \in \mathcal{L}$ and $\operatorname{supp} f \subset [a,b] \times S^2 \subset \Sigma$, then, in accordance with the results of section 1.2(i), representation (1.7) yields

$$\operatorname{supp} u^f(\cdot,t) \subset \overline{B}_{b-t} \setminus B_{a-t}, \quad t \leq 0. \tag{1.23}$$

[1]This fact also easily follows from LPT.

Thus, the controls $f \in \mathcal{L}$ produce the waves compactly supported in $\mathbb{R}^3$.

The space of controls contains the Sobolev class

$$\mathcal{F}_1 := \left\{ f \in \mathcal{F} \;\middle|\; \frac{\partial f}{\partial \tau} \in \mathcal{F} \right\}$$

equipped with the norm $\|f\|_1^2 = \|f\|_{\mathcal{F}}^2 + \|\frac{\partial f}{\partial \tau}\|_{\mathcal{F}}^2$ and its subclass

$$\mathcal{F}_{1,0} := \{ f \in \mathcal{F}_1 \;\big|\; f|_{\tau=0} = 0 \},$$

the set $\mathcal{L}$ being dense in $\mathcal{F}_{1,0}$.

Take an $f \in \mathcal{L}$. Integrating by parts with regard to the isometry of the control operator and property (1.23), we have

$$\begin{aligned}
\|f\|_1^2 &= \|f\|_{\mathcal{F}}^2 + \left(\frac{\partial f}{\partial \tau}, \frac{\partial f}{\partial \tau} \right)_{\mathcal{F}} = \|f\|_{\mathcal{F}}^2 - \left(\frac{\partial^2 f}{\partial \tau^2}, f \right)_{\mathcal{F}} \\
&= \|u^f(\cdot, 0)\|_{\mathcal{H}}^2 - \left(u^{\frac{\partial^2 f}{\partial \tau^2}}(\cdot, 0), u^f(\cdot, 0) \right)_{\mathcal{H}} = \|u^f(\cdot, 0)\|_{\mathcal{H}}^2 - \left(u^f_{tt}(\cdot, 0), u^f(\cdot, 0) \right)_{\mathcal{H}} \\
&= \|u^f(\cdot, 0)\|_{\mathcal{H}}^2 - \left(\Delta u^f(\cdot, 0), u^f(\cdot, 0) \right)_{\mathcal{H}} = \|u^f(\cdot, 0)\|_{H^1(\mathbb{R}^3)}^2 .
\end{aligned}$$

Extending to $f \in \mathcal{F}_{1,0}$, we obtain an isometric embedding

$$W_0 \mathcal{F}_{1,0} \subset H^1(\mathbb{R}^3). \tag{1.24}$$

This simple straightforward derivation demonstrates the utility of the class $\mathcal{L}$. However, (1.24) can be derived from LPT (see [10]).

1.4. Observation operator. The adjoint operator $W_0^* : \mathcal{H} \to \mathcal{F}$ plays the role of an *observation operator* of system (1.1)–(1.3). Since W_0 is unitary, one has $W_0^* = W_0^{-1}$, which relates W_0^* to the Radon transform

$$(\mathcal{R}y)(\tau, \theta) := \int_{\sigma \cdot \theta = \tau} y(\sigma)\, d\sigma, \quad (\tau, \theta) \in \Sigma,$$

as follows. For any $y \in C_0^\infty(\mathbb{R}^3)$, by the Radon inversion formula one has

$$\begin{aligned}
y(x) &= -\frac{1}{8\pi^2} \Delta_x \int_{S^2} (\mathcal{R}y)(x \cdot \theta, \theta)\, d\theta = -\frac{1}{8\pi^2} \int_{S^2} \left[\frac{\partial^2}{\partial \tau^2} \mathcal{R}y \right] (x \cdot \theta, \theta)\, d\theta \\
&= \frac{1}{2\pi} \int_{S^2} \frac{\partial}{\partial \tau} \left[-\frac{1}{4\pi} \frac{\partial}{\partial \tau} \mathcal{R}y \right] (x \cdot \theta, \theta)\, d\theta = \langle \text{see (1.22)} \rangle \\
&= \left(W_0 \left[-\frac{1}{4\pi} \frac{\partial}{\partial \tau} \mathcal{R}y \right] \right)(x);
\end{aligned}$$

hence,

$$W_0^* = W_0^{-1} = -\frac{1}{4\pi} \frac{\partial}{\partial \tau} \mathcal{R}\,, \tag{1.25}$$

whereas the isometry of W_0^* corresponds to the classical Plancherel formula for $\mathcal{R}$ (see [7], [8], [10]).

The embedding

$$W_0^* H^1(\mathbb{R}^3) \subset \mathcal{F}_1 \tag{1.26}$$

will be used below; it easily follows from the properties of the Radon transform and (1.25).

1.5. Controllability. To describe the structure of the sets $\mathcal{U}^\xi$ and $\mathcal{D}^\xi$ in the unperturbed case one needs merely to reformulate the results of section 1.2(iii).

Recall that the polyharmonic subspaces $\mathcal{A}^\xi$ are defined at the end of section 0.2. In the case $q = 0$, by this definition and in accordance with (1.4), one easily gets

$$\mathcal{A}^\xi = \oplus \sum_{l,m} \left\{ y \in \mathcal{H}_{lm} \;\middle|\; y = \frac{a(r)}{r}\, Y_l^m(\omega), a \in A_l^\xi \right\}.$$

By virtue of (1.15)–(1.18), the equalities

$$\mathcal{D}^\xi = \mathcal{A}^\xi, \quad \xi > 0; \quad \mathcal{D}^0 = \mathcal{A}^0 = \{0\} \tag{1.27}$$

hold (see [5]). Thus, for $\xi > 0$ the defect sets are infinite dimensional, whereas the reachable sets $\mathcal{U}^\xi = W_0\mathcal{F}^\xi$ are the *closed* subspaces determined by infinite number of orthogonality conditions: $y = \sum_{l,m} \frac{y_{lm}(r)}{r}\, Y_l^m(\omega)$ belongs to $\mathcal{U}^\xi$ if and only if $y_{lm} \perp A_l^\xi$.

1.6. Geometrical optics. Here we modify representations (1.21) and (1.22), extending them to the case of controls with $f\,|_{\tau=0} \neq 0$. Such controls with a jump at $\tau = 0$ produce the waves with a jump at the characteristic cone $\{(x,t) \mid |x| = |t|\}$, i.e., at the forward front of the wave. The geometrical optics formulae express the amplitude of the jump of u^f through $f\,|_{\tau=0}$.

Fix $\omega \in S^2$ and define

$$\pi_b(\omega) := \begin{cases} \left\{\theta \in S^2 \mid \omega \cdot \theta = b\right\}, & b \in [-1,1], \\ \emptyset, & |b| > 1. \end{cases}$$

The set $\pi_b(\omega)$ is a parallel on the unit sphere with the north pole ω, the length of the parallel is equal to $2\pi\sqrt{1-b^2}$; $\pi_0(\omega)$ is the equator; $\pi_{\pm 1}(\omega) = \pm\omega$. For a function g on S^2, denote by

$$[g]_b(\omega) := \begin{cases} \frac{1}{2\pi\sqrt{1-b^2}} \int_{\pi_b(\omega)} g(\theta)\, d\theta, & b \in [-1,1], \\ g(-\omega), & b = -1, \\ g(\omega), & b = 1, \\ 0, & |b| > 1, \end{cases}$$

the mean value of g on the parallel.

Lemma 1.1. *For $f \in \mathcal{F}_1$, the representation*

$$u^f(x,t) = \frac{1}{2\pi} \int_{S^2} f_\tau(t + r\omega\cdot\theta, \theta)\, d\theta + \frac{1}{r}\, [f(0,\cdot)]_{-\frac{t}{r}}(\omega) \tag{1.28}$$

holds, where $r = |x|$, $\omega = \frac{x}{|x|}$, and f_τ is understood as a classical (Sobolev) derivative of $f|_{\tau>0}$, the derivative being extended to $\tau < 0$ by zero.

Proof. Let us begin with $f \in \mathcal{F}_1 \cap C^1(\overline{\Sigma})$. Fix $\varepsilon > 0$ and denote

$$f^\varepsilon(\tau,\omega) := \begin{cases} 0, & \tau < 0, \\ \frac{f(0,\omega)}{\varepsilon}\,\tau, & 0 \le \tau < \varepsilon, \\ f(\tau - \varepsilon, \omega), & \varepsilon \le \tau < \infty. \end{cases}$$

Approximating $f^\varepsilon \in \mathcal{F}_{1,0}$ by smooth controls, we easily extend (1.21) to $f = f^\varepsilon$ and get

$$u^{f^\varepsilon}(x,t) = \frac{1}{2\pi}\int_{S^2}\frac{\partial f^\varepsilon}{\partial\tau}(t + x\cdot\theta,\theta)\,d\theta = \frac{1}{2\pi\varepsilon}\int_{0\le t+x\cdot\theta<\varepsilon} f(0,\theta)\,d\theta$$
$$+\frac{1}{2\pi}\int_{t+x\cdot\theta\ge\varepsilon} f_\tau(t + x\cdot\theta-\varepsilon,\theta)\,d\theta =: \, I_\varepsilon + II_\varepsilon. \tag{1.29}$$

To analyze I_ε, we introduce polar coordinates: choose $\omega',\omega'' \in S^2$ such that $\{\omega,\omega',\omega''\}$ form a Cartesian basis in $\mathbb{R}^3$ and define $\varphi = \varphi(\theta) \in [0, 2\pi)$, $\psi = \psi(\theta) \in [0,\pi]$ by $\cos\varphi = \omega'\cdot\theta$, $\sin\varphi = \omega''\cdot\theta$, $\cos\psi = \omega\cdot\theta$. Since the set of integration in I_ε is determined by the inequalities

$$\frac{-t}{r} \le \omega\cdot\theta \le \frac{-t}{r} + \frac{\varepsilon}{r},$$

one has

$$I_\varepsilon = \frac{1}{2\pi\varepsilon}\int_0^{2\pi}d\varphi\int_{\arccos[\frac{-t}{r}+\frac{\varepsilon}{r}]}^{\arccos\frac{-t}{r}} f\left(0,\theta(\varphi,\psi)\right)\sin\psi\,d\psi$$
$$= \frac{1}{2\pi}\int_0^{2\pi}d\varphi\,\frac{1}{\varepsilon}\int_{\frac{-t}{r}}^{\frac{-t}{r}+\frac{\varepsilon}{r}} f\left(0,\theta\left(\varphi,\arccos\eta\right)\right)\,d\eta.$$

Fix x,t such that $|x| > -t$. Sending $\varepsilon \to 0$ we get

$$I_\varepsilon \to \frac{1}{2\pi r}\int_0^{2\pi} f\left(0,\theta\left(\varphi,\arccos\frac{-t}{r}\right)\right)\,d\varphi = \frac{1}{r}\,[f(0,\cdot)]_{-\frac{t}{r}}(\omega).$$

By the continuity of f_τ, the limit passage in II_ε for fixed x,t gives

$$II_\varepsilon \to \frac{1}{2\pi}\int_{t+x\cdot\omega\ge0} f_\tau(t + x\cdot\theta,\theta)\,d\theta = \int_{S^2} f_\tau(t + r\omega\cdot\theta,\theta)\,d\theta.$$

Hence, the right-hand side (r.h.s.) of (1.29) tends pointwise in $\{(x,t) \mid 0 \le -t < |x|\}$ to the r.h.s. of (1.28). At the same time, $f_\varepsilon \to f$ in $\mathcal{F}$ that implies $u^{f^\varepsilon}(\cdot,t) \to u^f(\cdot,t)$ in $\mathcal{H}$, and we arrive at representation (1.28). For $(x,t) : |x| < -t$ the representation takes the trivial form $0 = 0$.

If $f \in \mathcal{F}_1$, then the value $f|_{\tau=0}$ is well defined as an element of $L_2(S^2)$, as well as $[f(0,\cdot)]_b(\cdot) \in L_2(S^2)$.[2] Therefore, approximating f by elements of $\mathcal{F}_1 \cap C^1(\overline{\Sigma})$ in $\mathcal{F}_1$-norm and passing to the limit, we justify (1.28). $\square$

A peculiarity of formula (1.28) is that, for a fixed t, the summands in its r.h.s. not necessarily belong to $\mathcal{H} = L_2(\mathbb{R}^3)$ individually. The formula represents the wave $u^f(\cdot,t)$ in a neighborhood of $x = 0$ and enables one to analyze its behavior near the forward front $\{x \in \mathbb{R}^3 \mid |x| = -t\}$. Namely, if x approaches the front ($|x| \to -t + 0$), then the first summand vanishes (since $\operatorname{mes}\{\theta \in S^2 \mid t + x\cdot\omega \ge 0\} \to 0$), whereas the second summand tends to the limit $r^{-1}[f(0,\cdot)]_1(\omega) = r^{-1}f(0,\omega)$. Summarizing, we get the relation

$$u^f(x,t)|_{x=(-t+0)\omega} = -\frac{f(0,\omega)}{t}\,, \tag{1.30}$$

[2] See the comments in section 1.7.

which is a typical geometrical optics formula.

Fixing $x \neq 0$ and taking $t = 0$ in (1.28), we represent the unperturbed control operator in the form

$$(W_0 f)(x) = \frac{1}{2\pi} \int_{S^2} f_\tau(r\omega \cdot \theta, \theta)\, d\theta + \frac{1}{r}\, [f(0, \cdot)]_0(\omega), \quad x \in \mathbb{R}^3. \tag{1.31}$$

The first summand belongs to $H^1_{\mathrm{loc}}(\mathbb{R}^3)$. Indeed, choose an (arbitrarily large) $r' > 0$ and $\eta = \eta(\tau) \in C^\infty[0, \infty)$ such that $\eta\mid_{[0,r']} = 1, \eta\mid_{[r'+1,\infty)} = 0$. Then, representing

$$\begin{aligned} f(\tau, \cdot) &= \eta(\tau)[f(\tau, \cdot) - f(0, \cdot)] + \{\eta(\tau) f(0, \cdot) + [1 - \eta(\tau)] f(\tau, \cdot)\} \\ &= f_1(\tau, \cdot) + f_2(\tau, \cdot) \end{aligned}$$

with $f_1 \in \mathcal{F}_{1,0}$ and $\operatorname{supp}(f_2)_\tau \subset [r', \infty)$, we have

$$\int_{S^2} f_\tau(r\omega \cdot \theta, \theta)\, d\theta = \int_{S^2} f_{1\tau}(r\omega \cdot \theta, \theta)\, d\theta + \int_{S^2} f_{2\tau}(r\omega \cdot \theta, \theta)\, d\theta = I_1(x) + I_2(x),$$

where $I_1 = W_0 f_1 \in H^1(\mathbb{R}^3)$ (see (1.24)) and $\operatorname{supp} I_2 \subset \mathbb{R}^3 \setminus B_{r'}$. The second summand in (1.31) is singular at $x = 0$ and does not belong to $H^1_{\mathrm{loc}}(\mathbb{R}^3)$.

1.7. Comments.

- In [5] we introduced one more representation of W_0^* through the so-called M transform. This transform is constructed from the jumps appearing as a result of projecting the functions on the reachable sets (see also [3]).
- The map $g \mapsto [g]_b(\cdot)$ is known as the Minkowski–Funk transform [7]; for $|b| < 1$ it is a compact operator in $L_2(S^2)$.
- Most probably, formulae (1.28), (1.30) are known but we did not succeed in finding exact references.

2. Perturbed system. We study system (0.1)–(0.3) and prepare auxiliary results required for further investigation of its controllability.

2.1. Representation of u^f. Recall that the potential is assumed compactly supported: $\operatorname{supp} q \subset \overline{B}_{r_*}$.

One possible way to establish the solvability of problem (0.1)–(0.3) for all $t \in (-\infty, \infty)$ is to present it in the form

$$(u - u_0^f)_{tt} - \Delta(u - u_0^f) = -qu, \qquad (x, t) \in \mathbb{R}^3 \times (-r_*, \infty), \tag{2.1}$$

$$(u - u_0^f)\mid_{t<-r_*} = 0, \tag{2.2}$$

where u_0^f is the solution of (1.1)–(1.3) and, applying the Poisson–Kirchhoff formula, to reduce to a Volterra-type equation

$$u(x, t) = u_0^f(x, t) - \frac{1}{4\pi} \int_{B_{t+r_*}(x)} \frac{q(\eta) u(\eta, t + r_* - |x - \eta|)}{|x - \eta|} d\eta, \tag{2.3}$$

which can be analyzed by standard methods (e.g., see [15]). If $f \in C_0^\infty(\Sigma)$, the solution $u = u^f(x, t)$ is classical (smooth).

The second representation is derived as follows. Analyzing (2.3), one establishes the well-known property of finiteness of the domain of influence, which leads to

$$\operatorname{supp}(u^f - u_0^f) \subset \{(x, t) \mid t \geq -r_*, -t \leq |x| \leq 2r_* + t\}$$

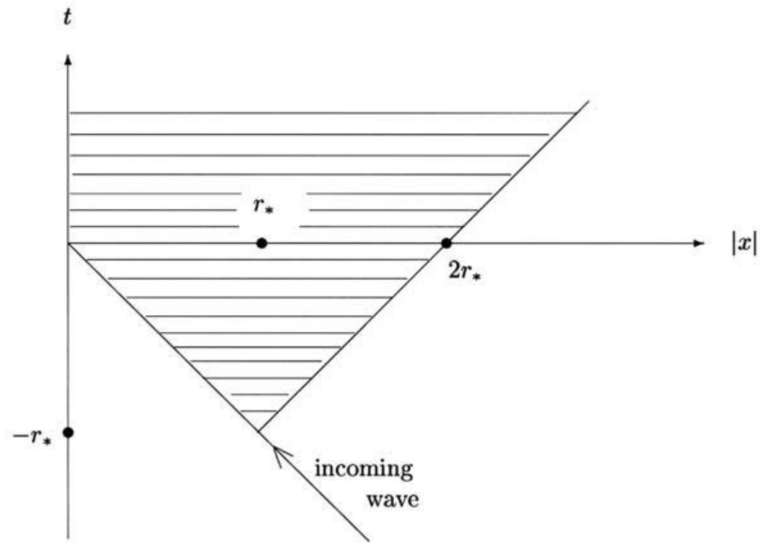


FIG. 2.1. *The set* supp $(u^f - u_0^f)$.

(this support is shadowed on Figure 2.1). Therefore, for times $t \le 0$ the problem for the difference $u^f - u_0^f$ can be reduced to the cylinder $B_{2r_*} \times [-r_*, 0]$ and written in the form

$$\begin{aligned}
&(u - u_0^f)_{tt} - \Delta(u - u_0^f) = -qu, \qquad (x,t) \in\ B_{2r_*} \times (-r_*, 0),\\
&(u - u_0^f)\mid_{t=-r_*} = (u - u_0^f)_t \mid_{t=-r_*} = 0, \qquad x \in \overline{B}_{2r_*},\\
&(u - u_0^f)\mid_{\Sigma^{r_*}} = 0,
\end{aligned}$$

where $\Sigma^{r_*} := \partial B_{2r_*} \times [-r_*, 0]$. Then, by an operator version of the Duhamel formula, one has

$$\left[u(\cdot,t) - u_0^f(\cdot,t)\right]\Big|_{B_{2r_*}\times[-r_*,0]} = -L^{-\frac{1}{2}} \int_{-r_*}^{t} \sin\left[(t-s)L^{\frac{1}{2}}\right] Y\left[q(\cdot)u(\cdot,s)\right] ds,$$

where $L : L_2(B_{2r_*}) \to L_2(B_{2r_*})$, Dom $L = H^2(B_{2r_*}) \cap H_0^1(B_{2r_*})$, $Ly = -\Delta y$; $Y : L^2(\mathbb{R}^3) \to L^2(B_{2r_*})$ is a reduction, $Yy := y\mid_{B_{2r_*}}$. Returning to $\mathbb{R}^3$, we get

$$u(\cdot,t) = u_0^f(\cdot,t) - Y^* L^{-\frac{1}{2}} \int_{-r_*}^{t} \sin\left[(t-s)L^{\frac{1}{2}}\right] Y\left[q(\cdot)u(\cdot,s)\right] ds. \tag{2.4}$$

The third representation uses a spectral decomposition of the operator $H = -\Delta + q$ (see the end of section 0.2). Let $\{E_\lambda\}_{\lambda\in\mathbb{R}} : E_\lambda = E_{\lambda-0}$ be its spectral measure, $\mathcal{H}_- := E_0\mathcal{H}$ be the subspace of negative spectrum, $\mathcal{H}_0 := (E_{+0} - E_0)\mathcal{H} = \operatorname{Ker} H$, and $\mathcal{H}_+ := (I - E_{+0})\mathcal{H}$, so that $\mathcal{H} = \mathcal{H}_- \oplus \mathcal{H}_0 \oplus \mathcal{H}_+$. Let $\varphi_1, \varphi_2, \dots, \varphi_p$ and $\varphi_1^0, \varphi_2^0, \dots, \varphi_n^0$ be the orthonormal bases in $\mathcal{H}_-$ and $\mathcal{H}_0$, respectively, $H\varphi_k = -\varkappa_k^2\varphi_k$, $\varkappa_k > 0$, $H\varphi_k^0 = 0$. Solving the problem

$$\begin{aligned}
&u_{tt} + Hu = 0, \qquad t \in (-\infty, \infty),\\
&u\mid_{t=0} = u^f(\cdot,0), \quad u_t\mid_{t=0} = u_t^f(\cdot,0)
\end{aligned}$$

by the Fourier method, we get

$$
\begin{aligned}
u^f(\cdot,t) = \sum_{k=1}^{p}&\left\{\cosh \varkappa_k t\, \left(u^f(\cdot,0),\varphi_k\right)_{\mathcal{H}} + \frac{\sinh \varkappa_k t}{\varkappa_k}\left(u^f_t(\cdot,0),\varphi_k\right)_{\mathcal{H}}\right\}\varphi_k \\
&+ \sum_{k=1}^{n}\left\{\mathbf{1}(t)\, \left(u^f(\cdot,0),\varphi^0_k\right)_{\mathcal{H}} + t(u^f_t(\cdot,0),\varphi^0_k)_{\mathcal{H}}\right\}\varphi^0_k \\
&+ \int_{+0}^{\infty}\left\{\cos\sqrt{\lambda}\,t\; dE_\lambda\, u^f(\cdot,0) + \frac{\sin\sqrt{\lambda}\,t}{\sqrt{\lambda}}\, dE_\lambda\, u^f_t(\cdot,0)\right\}. \qquad (2.5)
\end{aligned}
$$

Each of the representations mentioned above can be used for introducing a generalized solution of (0.1)–(0.3) and proving its existence and uniqueness. For instance, considering (2.4) as a second-order Volterra equation in the space $C([-r_*,0];\mathcal{H})$, it is easy to establish its solvability and the estimates

$$
\begin{aligned}
\|u^f(\cdot,t)\|_{\mathcal{H}} \;&\le\; \|u^f\|_{C([-r_*,0];\mathcal{H})} \;\le\; c_q\|u^f_0\|_{C([-r_*,0];\mathcal{H})} \\
&= \langle \text{see } (1.20)\rangle \;=\; c_q \sup_{t\in[-r_*,0]}\|W_0T_tf\|_{\mathcal{H}} = c_q\|f\|_{\mathcal{F}} \qquad (2.6)
\end{aligned}
$$

(with regard to isometry of W_0 and T_t), where a constant c_q is determined by r_* and $\|q\|_{L_\infty(B_{r_*})}$. Then the solution of (2.4) is claimed to be an L_2-solution of (0.1)–(0.3) for $f\in\mathcal{F}$.

Note in addition that for $f\in\mathcal{L}$,

$$
\operatorname{diam}\operatorname{supp} u^f(\cdot,t) < \infty, \quad t\in(-\infty,\infty), \qquad (2.7)
$$

holds. Indeed, for $t<-r_*$ one has $u^f(\cdot,t) = u^f_0(\cdot,t)$ and, hence, $\operatorname{supp} u^f(\cdot,t)$ is a compact set (see (1.23)). For $t\in[-r_*,\infty)$ this property is preserved by finiteness of domain of influence. Surely, the diameter tends to infinity as $t\to\infty$.

2.2. Control operator. Along with L_2-solution u^f of (0.1)–(0.3), the operator $W:\mathcal{F}\to\mathcal{H}$, $Wf = u^f(\cdot,0)$ is well defined and bounded: $\|W\|\le c_q$ (see (2.6)). The solution can be represented through the control operator in the form

$$
u^f(\cdot,t) = WT_tf, \qquad t\le 0. \qquad (2.8)
$$

Putting $t=0$ in (2.4) and changing $s\to -s$ in the integral, we obtain

$$
u^f(\cdot,t) = u^f_0(\cdot,t) - Y^*L^{-\frac12}\int_0^{r_*}\sin\left[sL^{\frac12}\right]Y\left[q(\cdot)u(\cdot,-s)\right]\,ds.
$$

By (2.8), this is equivalent to an operator relation

$$
W = W_0 - K, \qquad (2.9)
$$

where

$$
K = Y^*L^{-\frac12}\int_0^{r_*}\sin\left[sL^{\frac12}\right]Y\hat{q}\,WT_s\,ds
$$

and $\hat{q}$ multiplies functions by $q(\cdot)$. Since $L^{-\frac12}$ maps $L_2(B_{2r_*})$ onto $H^1_0(B_{2r_*})$ continuously, the operator $K:\mathcal{F}\to\mathcal{H}$ is compact, $\operatorname{Ran}K\subset Y^*H^1_0(B_{2r_*})\subset H^1(\mathbb{R}^3)$, and the evident estimate

$$
\|K\| \le \lambda_*^{-\frac12} r_* c_q \|q\|_{L_\infty(B_{r_*})} \qquad (2.10)
$$

holds, where $\lambda_* > 0$ is the lowest eigenvalue of L.

Recall the definitions $\mathcal{N} := \operatorname{Ker} W$, $\mathcal{U} := W\mathcal{F} = \operatorname{Ran} W$, and $\mathcal{D} := \mathcal{H} \ominus \mathcal{U}$ (see section 0.2). Since W_0 is unitary and K is compact, (2.9) and Fredholm's theorems lead to $\dim \mathcal{N} = \dim \mathcal{D} < \infty$, proving part of (0.6). Recall also that $\mathcal{F}^\xi = T_\xi \mathcal{F}$ is the subspace of delayed controls. Later we will use the following property of the control operator.

LEMMA 2.1. *The relation $\mathcal{F}^\xi \cap \mathcal{N} = \{0\}$, $\xi > 0$, holds.*

Proof. Fix $\xi > 0$ and choose $f \in \mathcal{F}^\xi \cap \mathcal{N}$. We shall show that $f = 0$.

Since $f \in \mathcal{N}$, one has $u^f(\cdot, 0) = 0$ and, hence, the odd extension

$$u(\cdot, t) = \begin{cases} u^f(\cdot, t), & t < 0, \\ -u^f(\cdot, -t), & t \geq 0, \end{cases}$$

turns out to be a solution of the class $C((-\infty, \infty); \mathcal{H})$ of the equation

$$u_{tt} - \Delta u + qu = 0, \qquad (x, t) \in \mathbb{R}^3 \times (-\infty, \infty). \tag{2.11}$$

Since $f \in \mathcal{F}^\xi$ implies $\operatorname{supp} u^f \subset \{(x, t) \mid |x| \geq -t + \xi\}$, such a solution satisfies $\operatorname{supp} u \subset \{(x, t) \mid |x| \geq |t| + \xi\}$ and, hence,

$$u = 0, \qquad (x, t) \in B_\xi \times (-\infty, \infty). \tag{2.12}$$

Applying the Fourier transform in (2.11), (2.12), we see that the function

$$\tilde{u}(x, k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ikt} u(x, t)\, dt$$

satisfies

$$-k^2 \tilde{u} - \Delta \tilde{u} + q\tilde{u} = 0, \qquad x \in \mathbb{R}^3,$$

and

$$\tilde{u}(x, k) = 0, \quad x \in B_\xi, \tag{2.13}$$

for all $k \in (-\infty, \infty)$. By the well-known uniqueness theorem for homogeneous second-order elliptic equations (see [9]), (2.13) implies $\tilde{u} = 0$ everywhere in $\mathbb{R}^3$. Hence, $u = 0$, $u^f = 0$, and finally $f = 0$. ∎

2.3. Observation operator. A dynamical system

$$v_{tt} - \Delta v + qv = 0, \qquad (x, t) \in \mathbb{R}^3 \times (-\infty, 0), \tag{2.14}$$

$$v\,|_{t=0} = 0, \quad v_t\,|_{t=0} = y, \qquad x \in \mathbb{R}^3, \tag{2.15}$$

is called *dual* to system (0.1)–(0.3). For $y \in \mathcal{H}$, its L^2-solution $v = v^y(x, t)$ can be introduced through an equivalent integral equation

$$v(x, t) = -\frac{1}{4\pi|t|} \int_{\partial B_{|t|}(x)} y(\sigma)\, d\sigma - \frac{1}{4\pi} \int_{B_{|t|}(x)} \frac{q(\eta) v(\eta, |t| - |x - \eta|)}{|x - \eta|}\, d\eta \tag{2.16}$$

($B_r(x)$ is a ball centered at x) obtained by applying the Poisson–Kirchhoff formula to (2.14), (2.15). This Volterra-type equation has a unique solution $v^y \in C_{\mathrm{loc}}((-\infty, 0]; \mathcal{H})$.

With the dual system one associates the observation operator $O : \mathcal{H} \to \mathcal{F}$, $\mathrm{Dom}\, O = \mathcal{S}$ (the Schwartz class of the smooth rapidly decaying functions):

$$(Oy)(\tau,\omega) := \lim_{s\to\infty} s \left[\frac{\partial v^y}{\partial t} + \frac{\partial v^y}{\partial r}\right]((s+\tau)\omega, -s),$$

where $\frac{\partial}{\partial r} := \frac{\partial}{\partial |x|} = \frac{x}{|x|}\cdot\nabla$. The well-posedness of this definition (i.e., the existence of the pointwise limit belonging to $\mathcal{F}$) can be seen from (2.16): for smooth y, the solution v^y is classical (smooth) and straightforward calculations lead to

$$\begin{aligned}
&\lim_{s\to\infty} s \left[\frac{\partial v^y}{\partial t} + \frac{\partial v^y}{\partial r}\right]((s+\tau)\omega, -s)\\
&\quad = -\frac{1}{4\pi}\frac{\partial}{\partial\tau}\int_{\omega\cdot\sigma=\tau} y(\sigma)\, d\sigma + \frac{1}{4\pi}\int_{\omega\cdot\eta\geq\tau} q(\eta)\, v_t^y(\eta, \tau-\omega\cdot\eta)\, d\eta,
\end{aligned}\tag{2.17}$$

where the first summand in the r.h.s. is $W_0^* y$ (see (1.25)), and the second one is smooth and compactly supported on $\overline{\Sigma}$. Since $|y|, |\nabla y|$ rapidly decrease at infinity, the limit turns out to be uniform with respect to $(\tau,\omega)\in\overline{\Sigma}$.

The following result clarifies a duality of systems (0.1)–(0.3) and (2.14), (2.15).

Lemma 2.2. *The relation $O^* = W$ is valid.*

Proof. For $y \in \mathrm{Dom}\, O$ and $f \in \mathcal{L}$, the solutions u^f and v^f are smooth and (2.7) holds. Integrating by parts, one has

$$\begin{aligned}
0 &= \int_{B_s\times[-s,0]} \left[u_{tt}^f - \Delta u^f + q u^f\right] v^y\, dxdt\\
&= \int_{B_s} dx \left[u_t^f v^y - u^f v_t^f\right]\Big|_{t=-s}^{t=0} - \int_{\partial B_s\times[-s,0]} \left[\frac{\partial u^f}{\partial r} v^y - u^f \frac{\partial v^y}{\partial r}\right] d\sigma dt\\
&\quad + \int_{B_s\times[-s,0]} u^f \left[v_{tt}^y - \Delta v^y + q v^y\right] dxdt\\
&= \langle \text{see } (2.14), (2.15)\rangle\\
&= -\left\{\int_{B_s} u^f(\cdot,0) y\, dx + \int_{B_s} \left[u_t^f v^y - u^f v_t^y\right]\Big|_{t=-s} dx\right\}\\
&\quad - \left\{\int_{-s}^0 dt \int_{\partial B_s} \left[\frac{\partial u^f}{\partial r} v^y - u^f \frac{\partial v^y}{\partial r}\right] d\sigma\right\}\\
&=: -I(s) - II(s),
\end{aligned}\tag{2.18}$$

where $s > 0$ is a large parameter. Since $u^f|_{|x|<-t} = 0$, one has $u^f(\cdot,-s)|_{B_s} = u_t^f(\cdot,-s)|_{B_s} = 0$ and, hence,

$$I(s) \to \int_{\mathbb{R}^3} u^f(\cdot,0) y\, dx = (Wf, y)_{\mathcal{H}}\tag{2.19}$$

as $s\to\infty$. Changing variables $t \to \tau - s$, $\sigma \to s^2\omega$, one gets

$$II(s) = \int_{[0,s]\times S^2} s^2 \left[\frac{\partial u^f}{\partial r} v^y - u^f \frac{\partial v^f}{\partial r}\right](s\omega, \tau-s)\, d\tau d\omega.\tag{2.20}$$

For $y \in \mathcal{L}$, the set $\mathrm{supp}\, f \subset \overline{\Sigma}$ is compact and one can show that the limit in (0.3) is uniform with respect to $(\tau,\omega)\in\overline{\Sigma}$, whereas the asymptotic

$$s u^f(s\omega, \tau - s) = f(\tau,\omega) + o(1)\tag{2.21}$$

is differentiable:

$$\frac{d}{ds}[su^f(s\omega,\tau-s)]=o(1).$$

This implies

$$u^f(s\omega,\tau-s)+s\left[\frac{\partial u^f}{\partial r}-\frac{\partial u^f}{\partial t}\right](s\omega,\tau-s)=o(1)$$

and, combining these asymptotics, we get

$$\begin{aligned} s\frac{\partial u^f}{\partial r}(s\omega,\tau-s)&=s\frac{\partial u^f}{\partial t}(s\omega,\tau-s)+o(1)\\ &=\frac{\partial}{\partial\tau}u^f(s\omega,\tau-s)+o(1)\end{aligned}$$

uniformly with respect to $(\tau,\omega)\in\overline{\Sigma}$. Substituting the asymptotics in (2.20), we have

$$\begin{aligned} II(s)=\int_{[0,s]\times S^2}\Bigg\{&\left[s\frac{\partial}{\partial\tau}u^f(s\omega,\tau-s)\right][sv^y(s\omega,\tau-s)]\\ &-[su^f(s\omega,\tau-s)]\left[s\frac{\partial v^y}{\partial r}(s\omega,\tau-s)\right]\Bigg\}\,d\tau d\omega+o(1).\end{aligned}$$

Integrating by parts in the first summand in $\{\cdots\}$ and taking into account the relations

$$u^f(s\omega,0)\big|_{s\gg1}=0,\quad u^f(s\omega,-s)=0$$

provided by $f\in\mathcal{L}$, one obtains

$$II(s)=-\int_{[0,s]\times S^2}\left[su^f(s\omega,\tau-s)\right]s\left[\frac{\partial v^y}{\partial t}+\frac{\partial v^y}{\partial r}\right](s\omega,\tau-s)\,d\tau d\omega+o(1).$$

Passing to the limit with regard to (2.21), we get

$$II(s)\to-\int_{[0,s]\times S^2}f(\tau,\omega)\left\{\lim_{s\to\infty}s\left[\frac{\partial v^y}{\partial t}+\frac{\partial v^y}{\partial r}\right](s\omega,\tau-s)\right\}d\tau d\omega=-(f,Oy)_{\mathcal{F}}. \tag{2.22}$$

Finally, sending $s\to\infty$ in (2.18) and taking into account (2.19), (2.22), we arrive at

$$0=-(Wf,y)_{\mathcal{H}}+(f,Oy)_{\mathcal{F}},$$

which leads to the assertion of the lemma in view of density of the used f's and y's in $\mathcal{F}$ and $\mathcal{H}$, respectively. ∎

Let us discuss some consequences of this result. The eigenfunctions φ_k of the negative spectrum of H possess the well-known asymptotic as $|x|\to\infty$:

$$\varphi_k(x)=\alpha_k(\omega)\,\frac{e^{-\varkappa_k r}}{r}\,\left[1+o(1)\right]. \tag{2.23}$$

Put $y=\varphi_k$ in (2.14), (2.15); the solution of the dual problem is

$$v^{\varphi_k}(x,t)=\frac{\sinh\varkappa_k t}{\varkappa_k}\,\varphi_k(x). \tag{2.24}$$

Using (2.23), (2.24), one can find the limit, which defines the observation operator, by straightforward calculation and then, applying Lemma 2.2, obtain

$$(W^*\varphi_k)(\tau,\omega) = \alpha_k(\omega)\, e^{-\varkappa_k \tau}, \quad (\tau,\omega) \in \Sigma. \tag{2.25}$$

In the case of $0 \in \sigma_p(H)$ the eigenfunctions φ_k^0 have the asymptotics

$$\varphi_k^0(x) = \frac{\alpha_k^0(\omega)}{r^2}\,[1 + o(1)]\,. \tag{2.26}$$

Considering the dual problem with $y = \varphi_k^0$, we have

$$v^{\varphi_k^0}(x,t) = t\,\varphi_k^0(x). \tag{2.27}$$

Using (2.26), (2.27), one can obtain the equality

$$(W^*\varphi_k^0)(\tau,\omega) = 0, \quad (\tau,\omega) \in \Sigma\,,$$

which yields

$$\operatorname{Ker} H \subset \operatorname{Ker} W^*. \tag{2.28}$$

One more general fact is the embedding $W^*H^1(\mathbb{R}^3) \subset \mathcal{F}_1$, which can be derived from (1.26), (2.17), and Lemma 2.2.

2.4. Response operator. Dealing with system (0.1)–(0.3) for all times $t \in (-\infty,\infty)$, one associates with it a *response operator* $R : \mathcal{F} \to \mathcal{F}$, $\operatorname{Dom} R = \mathcal{L}$,

$$(Rf)(\tau,\omega) := \lim_{s\to+\infty} s u^f((s+\tau)\omega, s), \quad (\tau,\omega) \in \Sigma.$$

The well-posedness of this definition can be verified as follows. If $f \in \mathcal{L}$ and $\operatorname{supp} f \subset [a,b]\times S^2$ $(a > 0)$, then property (1.23) and the finiteness of domain of influence imply

$$u_0^f\big|_{0\le t\le |x|+a} = 0, \quad u^f\big|_{0\le t\le |x|-2r_*} = 0$$

for a large enough $|x|$ (see Figure 2.2: $\operatorname{supp} u_0^f$ and $\operatorname{supp} u^f$ are shadowed; $l_\tau = \{(x,t) \mid x = (s+\tau)\omega,\ t = s;\ s \ge 0\}$ is a space-time ray; the dotted line is the cone $|x| = |t|$).

As a result, $s u_0^f((s+\tau)\omega, s)\,|_{\tau\ge 0}= 0$ for large enough s and, taking into account the asymptotic

$$s - |(s+\tau)\omega - \eta| \to \omega\cdot\eta - \tau + o(1),$$

the passage to the limit in (2.3) easily leads to

$$\begin{aligned}(Rf)(\tau,\omega) &= \lim_{s\to\infty} s u^f((s+\tau)\omega, s)\\ &= -\frac{1}{4\pi}\int_{\omega\cdot\eta\ge\tau-r_*} q(\eta)\, u^f(\eta, \omega\cdot\eta - \tau + r_*)\, d\eta.\end{aligned} \tag{2.29}$$

This justifies the definition and provides also a representation of the response operator. Note a simple consequence of (2.29): since for $\tau > 2r_*$ the domain of integration does not intersect $\operatorname{supp} q$, one has $Rf\,|_{\tau>2r_*}= 0$, i.e., $\operatorname{supp} Rf \subset [0,2r_*]$. Using the same representation, one can show that R acts continuously from $\mathcal{F}$ to $C([0,2r_*]\times S^2)$ and

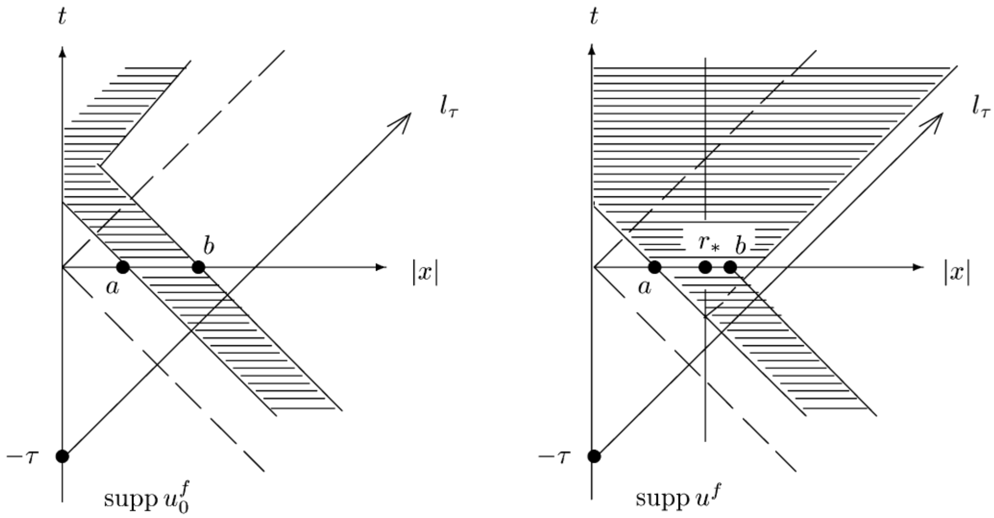


Fig. 2.2. *The supports.*

is a compact operator in $\mathcal{F}$. Hence, R can be extended to $\mathcal{F}$, representation (2.29) remaining in force for $f \in \mathcal{F}$.

An important property of R is the local character of its dependence on the potential: the response of the system on delayed controls $R\mathcal{F}^\xi$ is determined by $q\,|_{\mathbb{R}^3 \setminus B^\xi}$, $\xi \geq 0$. One more general property, following from the time independence of q is the relation $RT_\xi = T_\xi^* R$, $\xi \geq 0$ and its consequence

$$R\,\frac{\partial}{\partial\tau} f = -\,\frac{\partial}{\partial\tau}\,Rf, \quad f \in \mathcal{F}_{1,0}. \tag{2.30}$$

Note in addition that the operator R can be identified with some block of the scattering operator of LPT in the translation representation.

2.5. Connecting operator. A map $C : \mathcal{F} \to \mathcal{F}$,

$$C := W^* W,$$

is called a *connecting operator* of system (0.1)–(0.3). By this definition, for $f, g \in \mathcal{F}$, one has

$$(Cf, g)_{\mathcal{F}} = (Wf, Wg)_{\mathcal{H}} = \left(u^f(\cdot, 0), u^g(\cdot, 0)\right)_{\mathcal{H}}, \tag{2.31}$$

i.e., C connects the Hilbert metrics of the outer and inner spaces.

Lemma 2.3. *The relation*

$$C = I + R \tag{2.32}$$

holds.

Proof. Take $f, g \in \mathcal{L}$ and let u^f, u^g be the corresponding solutions of problem (0.1)–(0.3) considered for all times $t \in (-\infty, \infty)$. Blagovestchenskii's function

$$b(s,t) := \left(u^f(\cdot, s),\, u^g(\cdot, t)\right)_{\mathcal{H}}, \quad -\infty < s, t < \infty,$$

is well defined by (2.7) and possesses the following properties.

(i) Function b satisfies the wave equation

$$b_{tt} - b_{ss} = 0. \tag{2.33}$$

Indeed

$$\begin{aligned}\left[\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial s^2}\right] b(s,t) &= \int_{\mathbb{R}^3} \left[u^f(x,s)u^g_{tt}(x,t) - u^f_{ss}(x,s)u^g(x,t)\right]\, dx \\ &= \int_{\mathbb{R}^3} \left\{u^f(x,s)\left[\Delta u^g(x,t) - q(x)u^g(x,t)\right]\right. \\ &\qquad \left. - \left[\Delta u^f(x,s) - q(x)u^f(x,s)\right] u^g(x,t)\right\}\, dx \\ &= \int_{\mathbb{R}^3} \left[u^f(x,s)\Delta u^g(x,t) - \Delta u^f(x,s)u^g(x,t)\right]\, dx = 0\end{aligned}$$

by compactness of $\operatorname{supp} u^f(\cdot,t)$ and $\operatorname{supp} u^g(\cdot,t)$.

(ii) The relation

$$\lim_{s\to\infty} b(-s,-s) = (f,g)_{\mathcal{F}} \tag{2.34}$$

holds. Indeed, let $\operatorname{supp} f, \operatorname{supp} g \subset (0,\beta] \times S^2$ $(\beta < \infty)$. Since u^f and u^g vanish in the past cone one has

$$\begin{aligned} b(-s,-s) &= \int_{\mathbb{R}^3} u^f(x,-s)\, u^g(x,-s)\, dx = \int_{|x|>-s} \\ &= \langle \text{see } (1.23)\rangle = \int_s^{s+\beta} dr\ r^2 \int_{S^2} u^f(r\omega,-s)\, u^g(r\omega,-s)\, d\omega \\ &= \int_0^\beta d\tau (s+\tau)^2 \int_{S^2} u^f((s+\tau)\omega,-s)\, u^g((s+\tau)\omega,-s)\, d\omega \\ &= \int_0^\beta d\tau (1+\frac{\tau}{s})^2 \int_{S^2} \left[su^f((s+\tau)\omega,-s)\right]\left[su^g((s+\tau)\omega,-s)\right]\, d\omega \\ &= \langle \text{see } (0.3)\rangle \to (f,g)_{\mathcal{F}} \quad \text{as} \quad s\to\infty. \end{aligned}$$

(iii) The relation

$$\lim_{s\to\infty} b(s,-s) = (Rf,g)_{\mathcal{F}} \tag{2.35}$$

holds. Indeed,

$$\begin{aligned} b(s,-s) &= \int_{\mathbb{R}^3} u^f(x,s)\, u^g(x,-s)\, dx = \langle \text{see } (1.23)\rangle \\ &= \int_s^{s+\beta} dr\ r^2 \int_{S^2} u^f(r\omega,s)\, u^g(r\omega,-s)\, d\omega \\ &= \int_0^\beta d\tau \left[1+\frac{\tau}{s}\right]^2 \int_{S^2} \left[su^f((s+\tau)\omega,s)\right]\left[su^g((s+\tau)\omega,-s)\right] d\omega \\ &\to (Rf,g)_{\mathcal{F}} \quad \text{as } s\to\infty. \end{aligned}$$

(iv) The relation

$$\lim_{s\to\infty} b(0,-s) = 0 \tag{2.36}$$

is valid. Indeed,

$$b(0,-s) = \int_{\mathbb{R}^3} u^f(x,0)\, u^g(x,-s)\, dx\Big|_{s\gg 1} = 0$$

since $\operatorname{supp} u^f(\cdot,0) \subset B_\beta$ and $\operatorname{supp} u^f(\cdot,-s) \subset \mathbb{R}^3 \setminus B_s$.

Since b satisfies (2.33), it is a sum of two D'Alembert solutions:

$$b(s,t) = \Phi(s+t) + \Psi(s-t).$$

At the same time, relations (2.34)–(2.36) imply

$$(f,g)_{\mathcal{F}} = \Phi(-\infty) + \Psi(0), \quad (Rf,g)_{\mathcal{F}} = \Phi(0) + \Psi(\infty), \quad 0 = \Phi(-\infty) + \Psi(\infty).$$

Combining these equalities, one gets

$$\begin{aligned}(f,g)_{\mathcal{F}} + (Rf,g)_{\mathcal{F}} &= \Phi(0) + \Psi(0) = b(0,0)\\ &= \big(u^f(\cdot,0), u^g(\cdot,0)\big)_{\mathcal{H}} = \ \langle \text{see (2.31)} \rangle \ = (Cf,g)_{\mathcal{F}}\end{aligned}$$

and, hence, $Cf = f + Rf$ for $f \in \mathcal{L}$. By density of $\mathcal{L}$, we arrive at (2.32). $\square$

The result of the lemma connects null controls $f \in \mathcal{N}$ with the response operator: since $\operatorname{Ker} C = \operatorname{Ker} W =: \mathcal{N}$, representation (2.32) implies

$$\mathcal{N} = \operatorname{Ker}\,(I + R)\,. \tag{2.37}$$

Therefore $f = -Rf$ and, by virtue of the inclusion $\operatorname{supp}\, Rf \subset [0, 2r_*]$, equality (2.37) implies $\operatorname{supp}\, f \subset [0, 2r_*]$.

One more consequence is the representation

$$R = -W_0^* K - K^* W_0 + K^* K$$

following from (2.32) and (2.9).

3. Controllability. Here we prove Theorem 0.1 and study the structure of the defect subspaces in more detail.

3.1. Sets $\mathcal{U}^\xi$. Recall that the objects in use are defined in section 0.2.

Fix $\xi \geq 0$. Since $W = W_0 - K$, where W_0 is unitary and K is compact, it follows from the result of Lemma 2.1 and the Fredholm theory that the operator W maps $\mathcal{F}^\xi$ onto $W\mathcal{F}^\xi = \mathcal{U}^\xi$ isomorphically. Hence, the reachable sets $\mathcal{U}^\xi$ are closed subspaces and $\mathcal{U}^\xi \subset \mathcal{H}^\xi$ holds.

Let us prove the first of inclusions (0.5). Let $a \in \mathcal{A}^\xi$ be such that $(-\Delta + q)^p a = 0$ in $\mathbb{R}^3 \setminus \overline{B}_\xi$; take $g \in \mathcal{F}^\xi \cap \mathcal{L}$ and put $f = (\frac{\partial}{\partial \tau})^{2p} g$. Since the potential q does not depend on time, the solution u^f of (0.1)–(0.3) satisfies

$$u^f(\cdot,0) = u^{(\frac{\partial}{\partial\tau})^{2p} g}(\cdot,0) = \left(\frac{\partial^2}{\partial t^2}\right)^p u^g(\cdot,0) = (\Delta - q)^p\, u^g(\cdot,0).$$

Therefore

$$\begin{aligned}\big(a, u^f(\cdot,0)\big)_{\mathcal{H}} &= \int_{\mathbb{R}^3} a(x)\,(\Delta - q)^p\, u^g(x,0)\, dx\\ &= \langle \text{see (2.7)} \rangle = \int_{\mathbb{R}^3} [(\Delta - q)^p\, a(x)]\, u^g(x,0)\, dx = 0.\end{aligned} \tag{3.1}$$

Since the set of f's used in this calculation is dense in $\mathcal{F}^\xi$ (see section 1.3), the set of the corresponding waves $u^f(\cdot,0)$ is dense in $\mathcal{U}^\xi = W\mathcal{F}^\xi$. Thus, $a \perp \mathcal{U}^\xi$ and, hence, $\mathcal{A}^\xi \subset \mathcal{H}^\xi \ominus \mathcal{U}^\xi = \mathcal{D}^\xi$. In addition, if $\xi > 0$, then $\dim \mathcal{A}^\xi = \infty$ and, hence, $\dim \mathcal{D}^\xi = \infty$.

For proving the second relation of (0.5), it is enough to take $p = 1$ in (3.1). Another argument is relation (2.28) implying $\operatorname{Ker} H \subset \operatorname{Ker} W^* = \mathcal{H} \ominus \operatorname{Ran} W = \mathcal{H} \ominus \mathcal{U} = \mathcal{D}$. Thus, part (i) of the theorem is proved.

By (2.10), for small enough $\|g\|_{L_\infty(B_{r^*})}$ one has $\|K\| < 1$ and $W = W_0 - K$ turns out to be an isomorphism. This implies $\mathcal{U} = \mathcal{H}$, $\mathcal{D} = \{0\}$, and proves part (ii).

Postponing the further proof, let us comment on the obtained results. Thus, at the moment $t = 0$ the delayed incoming waves fill the subdomain $\mathbb{R}^3 \setminus B_\xi$, but the set of such waves $\mathcal{U}^\xi$ is not complete in $\mathcal{H}^\xi = L_2(\mathbb{R}^3 \setminus B_\xi)$ as $\xi > 0$ and, moreover, $\dim \mathcal{H}^\xi \ominus \mathcal{U}^\xi = \infty$. In the meantime, such a completeness occurs in the case of a bounded domain. Namely, consider the problem

$$u_{tt} - \Delta u + q = 0, \quad (x,t) \in B_T \times (-T,0), \tag{3.2}$$

$$u\,|_{t=-T} = u_t\,|_{t=-T} = 0 \quad \text{in } B_T, \tag{3.3}$$

$$u\,|_{\Sigma^T} = f, \tag{3.4}$$

where $\Sigma^T = \partial B_T \times [-T,0]$, $u = u^f(x,t)$ is a solution. Denote

$$\mathcal{F}^T := L_2(\Sigma^T),\ \ \mathcal{F}^{T,\xi} := \left\{ f \in \mathcal{F}^T \;\middle|\; \operatorname{supp} f \subset \partial B_T \times [-T+\xi, 0] \right\},$$

$$\mathcal{H} := L_2(B_T),\ \ \mathcal{H}^\xi := \{ y \in \mathcal{H} \mid \operatorname{supp} y \subset \overline{B}_T \setminus B_\xi \};$$

introduce the reachable sets

$$\mathcal{U}^\xi := \{ u^f(\cdot,0) \mid f \in \mathcal{F}^{T,\xi} \}.$$

By finiteness of the domains of influence, the inclusion $\mathcal{U}^\xi \subset \mathcal{H}^\xi$ holds. The well-known fact derived from the fundamental Holmgren–John–Tataru theorem (e.g., see [2]) is the density of this inclusion:

$$\operatorname{clos} \mathcal{U}^\xi = \mathcal{H}^\xi, \quad 0 \le \xi \le T, \tag{3.5}$$

which is interpreted as an *approximate controllability* of system (3.2)–(3.5) in subdomains filled with waves. Problem (0.1)–(0.3) may be considered as a limit case of (3.2)–(3.4) as $T \to \infty$: loosely speaking, system (0.1)–(0.3) is governed by controls located at the sphere of infinite radius. However, the properties of the reachable sets turn out to be different: comparing (3.5) and (0.5), we state a lack of controllability in the limit case.[3] One more noteworthy difference is that in a finite ball equality (3.5) fails without the closure (see [1]), whereas in the limit case $\mathcal{U}^\xi$ is closed.

The following result shows that, studying the general properties of the reachable sets, it is hardly possible to find a simple characterization of their elements.

LEMMA 3.1. *If $\xi > 0$ and u is a compactly supported element of $\mathcal{U}^\xi$, then the set $\mathbb{R}^3 \setminus \operatorname{supp} u$ is disconnected.*

Sketch of the proof. Assume the opposite: there exists a nonzero compactly supported $u \in \mathcal{U}^\xi$ such that $\mathbb{R}^3 \setminus \operatorname{supp} u$ is an (open) connected set.

[3]This lack of controllability is in a sense partially compensated by the following property. It can be shown that for any ξ, ξ' such that $0 \le \xi < \xi' < \infty$, one has $\operatorname{clos}\{u^f(\cdot,0)|_{B_{\xi'} \setminus B_\xi} \mid f \in \mathcal{F}^\xi\} = L_2(B_{\xi'} \setminus B_\xi)$, i.e., the forward parts of delayed incoming waves possess a local completeness.

Since $\mathcal{U}^\xi$ is closed, one has $u = u^f(\cdot,0)$ for some $0 \neq f \in \mathcal{F}^\xi$. The function

$$w(x,t) = \begin{cases} u^f(x,t), & t \le 0, \\ -u^f(x,-t), & t > 0, \end{cases}$$

is a solution of

$$w_{tt} - \Delta w + qw = -2u^f(\cdot,0)\,\delta'$$

($\delta = \delta(t)$ is the Dirac function) satisfying

$$w(\cdot,t)\big|_{B_\xi} = 0, \quad t \in (-\infty,\infty).$$

Its Fourier transform $\tilde{w}^f(\cdot,k)$ satisfies

$$\Delta\tilde{w} + (k^2 - q)\tilde{w} = 2ik\,u^f(\cdot,0)$$

and

$$\tilde{w}\big|_{B_\xi} = 0, \quad k \in (-\infty,\infty).$$

Hence, out of $\operatorname{supp} u^f(\cdot,0)$ the function $\tilde{w}$ is a solution of a homogeneous elliptic equation, which vanishes identically on B_ξ. By [9], $\tilde{w} = 0$ everywhere in $\mathbb{R}^3 \setminus \operatorname{supp} u^f(\cdot,0)$. In particular,

$$\tilde{w}(\cdot,k)\big|_{\mathbb{R}^3\setminus B_{\xi'}} = 0, \quad k \in (-\infty,\infty),$$

for large enough ξ' provided $\operatorname{supp}\, u^f(\cdot,0) \subset \overline{B}^{\xi'} \setminus B^\xi$. This leads to

$$w(\cdot,t)\big|_{\mathbb{R}^3\setminus B_{\xi'}} = 0, \quad t \in (-\infty,\infty),$$

yields $u^f = 0$ in $[\mathbb{R}^3 \setminus B_{\xi'}] \times (-\infty, 0]$ and implies $f = 0$, which contradicts the assumption $f \neq 0$. $\square$

This rather subtle result is of general character for systems governed by hyperbolic equations (see [1], [3], [4]).

3.2. Radial potential. Here we prove relations (0.7). Thus, in this section we deal with the case $q = q(r)$.

In this case, the sets $\mathcal{A}^\xi$, $\mathcal{D}^\xi$, and $\mathcal{U}^\xi$ are invariant with respect to rotations and, hence, are reduced by spherical harmonics. Namely,

$$\mathcal{A}^\xi = \oplus\sum_{l,m} \mathcal{A}^\xi{}_{lm}\,, \quad \mathcal{A}^\xi{}_{lm} := \left\{ y \in \mathcal{H}^\xi \;\middle|\; y(x) = \frac{a(r)}{r}\, Y_l^m(\omega), a \in A_l^\xi \right\},$$

where A_l^ξ are defined in the same way as in unperturbed case, replacing (1.14) by

$$\left(-\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2} + q(r)\right)^p a \;=\; 0, \quad r > \xi.$$

As well as in the unperturbed case, the ODE theory implies $\operatorname{rank} A_l^\xi = \dim A_l^\xi = d(l)$ as $\xi > 0$. Quite analogously,

$$\mathcal{U}^\xi = \oplus\sum_{l,m} \mathcal{U}^\xi{}_{lm}, \quad \mathcal{U}^\xi{}_{lm} = W\mathcal{F}^\xi{}_{lm},$$

$$\mathcal{D}^\xi = \oplus\sum_{l,m} \mathcal{D}^\xi{}_{lm}, \quad \mathcal{D}^\xi{}_{lm} = \mathcal{H}^\xi{}_{lm} \ominus \mathcal{U}^\xi{}_{lm} \supset \mathcal{A}^\xi{}_{lm},$$

where

$$\mathcal{F}^{\xi}{}_{lm} := \mathcal{F}_{lm} \cap \mathcal{F}^{\xi}, \quad \mathcal{H}^{\xi}{}_{lm} := \mathcal{H}_{lm} \cap \mathcal{H}^{\xi}$$

(see (1.4), (1.5)). Marking the objects corresponding to $q = 0$ by the index $\{\,\mathring{}\,\}$, one has evidently

$$\mathcal{A}^{\xi}{}_{lm}\;\big|_{\;\xi > r_*} = \mathring{\mathcal{A}}^{\xi}_{lm}, \quad \mathcal{U}^{\xi}{}_{lm}\;\big|_{\;\xi > r_*} = \mathring{\mathcal{U}}^{\xi}_{lm}, \quad \mathcal{D}^{\xi}{}_{lm}\;\big|_{\;\xi > r_*} = \mathring{\mathcal{D}}^{\xi}_{lm}.$$

We say that a family of subspaces $\{\mathcal{B}^{\xi}\}$ is continuous with respect to ξ if the orthogonal projection on $\mathcal{B}^{\xi}$ is continuous in the strong operator topology. By Lemma 2.1, as $\xi > 0$ the control operator W maps $\mathcal{F}^{\xi}{}_{lm}$ onto $\mathcal{U}^{\xi}{}_{lm}$ isomorphically. Therefore, since $\{\mathcal{F}^{\xi}{}_{lm}\}_{\xi>0}$ is continuous, the family $\{\mathcal{U}^{\xi}{}_{lm}\}_{\xi>0}$ is also continuous. Accordingly, $\{\mathcal{D}^{\xi}{}_{lm}\}_{\xi>0}$ is continuous, which implies $\dim \mathcal{D}^{\xi}{}_{lm} = \text{const} \le \infty$. In the meantime,

$$\begin{aligned}\dim \mathcal{D}^{\xi}{}_{lm}\;\big|_{\;\xi > r_*} &= \dim \mathring{\mathcal{D}}^{\xi}_{lm} = \langle \text{see (1.27)} \rangle = \dim \mathring{\mathcal{A}}^{\xi}{}_{lm} \\ &= (2l+1) \dim A^{\xi}_{l} = \langle \text{see section 1.2(iii)} \rangle = (2l+1)\, d(l) < \infty;\end{aligned}$$

hence, $\dim \mathcal{D}^{\xi}{}_{lm} = (2l+1)d(l)$ for all $\xi > 0$. On the other hand, by ODE theory one has $\dim \mathcal{A}^{\xi}{}_{lm} = \dim \mathring{\mathcal{A}}^{\xi}_{lm}$ for all $\xi > 0$. Therefore, the evident inclusion $\mathcal{D}^{\xi}{}_{lm} \supset \mathcal{A}^{\xi}{}_{lm}$ leads to $\mathcal{D}^{\xi}{}_{lm} = \mathcal{A}^{\xi}{}_{lm}$ and, finally, to $\mathcal{D}^{\xi} = \mathcal{A}^{\xi}$, $\xi > 0$.

Let us prove that $\mathcal{D} = \mathcal{A} = \operatorname{Ker} H$. If $h \in \mathcal{D}$, then $h\big|_{\mathbb{R}^3 \setminus B_\xi} \in \mathcal{D}^{\xi}$ and, hence, for $\xi > 0$ we have $h\big|_{\mathbb{R}^3 \setminus B_\xi} \in \mathcal{A}^{\xi}$. Therefore, representing $h(x) = \sum_{l,m} \frac{h_{lm}(r)}{r}\, Y^m_l(\omega)$ for $g := h_{lm}$ we have

$$\left(-\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2} + q(r)\right)^{p} g = 0, \quad r > 0, \tag{3.6}$$

with an integer $p > 0$. If $\xi > r_*$, then $g|_{(\xi,\infty)} \in \mathring{A}^{\xi}{}_{l}$ and, hence, $g|_{(\xi,\infty)}$ satisfies

$$\left(-\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2}\right)^{d(l)} g = 0, \quad r > \xi \tag{3.7}$$

(see (1.14)). Comparing (3.6) with (3.7), we conclude that (3.6) holds for $p = d(l)$.

Since g is an L_2-solution of (3.6) with $p = d(l)$, by the ODE theory it has the differentiable asymptotic $g|_{r\to 0} = O(r^{l+1})$, whereas $g|_{r>r_*}$ is a linear combination of the solutions (1.16). From this it easily follows that $\frac{g(r)}{r}\, Y^m_l(\omega) \in \operatorname{Dom} H^j$ as $j = 1, 2, \dots, d(l)$ and, hence, $\frac{g(r)}{r}\, Y^m_l(\omega) \in \operatorname{Ker} H^{d(l)}$. Since H is a self-adjoint operator, the last inclusion implies $\frac{g(r)}{r}\, Y^m_l(\omega) \in \operatorname{Ker} H$ and, returning to the element h, we get $h \in \operatorname{Ker} H$. Thus, $\mathcal{D} \subset \operatorname{Ker} H$, whereas the converse inclusion has been proved in section 3.1. Thus, we arrive at $\mathcal{D} = \operatorname{Ker} H$.

The proof of Theorem 0.1 is completed.

3.3. Example of $\mathcal{D} \neq \operatorname{Ker} H = \{0\}$. The result of this section shows that, in the general (not radial) case, the defect subspace is not exhausted by Ker H. Here, dealing with the operator H for different q's, we denote it by H_q.

Take a smooth compactly supported q' such that $\operatorname{Ker} H_{q'} \neq \{0\}$ (it does exist; see [14]) and choose $\varphi' \in \operatorname{Ker} H_{q'}$, so that

$$(-\Delta + q')\, \varphi' = 0 \quad \text{in } \mathbb{R}^3.$$

By ellipticity, one has $\varphi' \in C^\infty(\mathbb{R}^3)$. Let x_0 and a (small) $\varepsilon > 0$ be such that $\varphi'(x) \geq \text{const} > 0$, $x \in B_\varepsilon(x_0)$. Construct a function

$$\varphi''(x) = \begin{cases} \varphi'(x), & x \in \mathbb{R}^3 \setminus B_\varepsilon(x_0), \\ \geq \frac{\text{const}}{2}, & x \in B_\varepsilon(x_0) \setminus B_{\frac{\varepsilon}{2}}(x_0), \\ \frac{1}{|x-x_0|}, & x \in B_{\frac{\varepsilon}{2}}(x_0) \setminus \{x_0\}, \end{cases}$$

such that $\varphi'' \in C^\infty(\mathbb{R}^3 \setminus \{x_0\})$ and define

$$q''(x) = \begin{cases} q'(x), & x \in \mathbb{R}^3 \setminus B_\varepsilon(x_0), \\ \frac{\Delta\varphi''(x)}{\varphi''(x)}, & x \in B_\varepsilon(x_0) \setminus \{x_0\}, \\ 0, & x = x_0. \end{cases}$$

It is easy to see that $q'' \in C_0^\infty(\mathbb{R}^3)$ and φ'' satisfies

$$(-\Delta + q'')\,\varphi'' = 0 \quad \text{in } \mathbb{R}^3 \setminus \{x_0\}.$$

Note also that $q'' = 0$ in $B_{\frac{\varepsilon}{2}}(x_0)$.

Defining $\varphi(x) := \varphi''(x - x_0)$, $q(x) := q''(x - x_0)$, we get

$$\varphi \in \mathcal{H}, \quad (-\Delta + q)\,\varphi = 0 \quad \text{in } \mathbb{R}^3 \setminus \{0\},$$

which obviously implies $\varphi \in \mathcal{A} \subset \mathcal{D}$. In the meantime, $\varphi \notin \operatorname{Ker} H_q$ because φ is singular at $x = 0$, whereas $\operatorname{Ker} H_q \subset C^\infty(\mathbb{R}^3)$.

If $\operatorname{Ker} H_q = \{0\}$, then the required example is provided. If, accidentally, $\operatorname{Ker} H_q \neq \{0\}$, then one can deform q, preserving nontrivial defect but gaining $\operatorname{Ker} H_q = \{0\}$. This can be done in the following way. For simplicity we assume that $\lambda = 0$ is an ordinary eigenvalue: $\dim \operatorname{Ker} H_q = 1$.

Take $0 \neq \psi \in \operatorname{Ker} H_q$. Such an element satisfies $(-\Delta + q)\,\psi = 0$ and, hence,

$$\psi(x) = -\frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{q(y)\psi(y)}{|x-y|}\,dy, \qquad x \in \mathbb{R}^3. \tag{3.8}$$

If $\operatorname{supp} q \subset \overline{B}_{r_*}$, then $\Delta\psi = 0$ in $\mathbb{R}^3 \setminus \overline{B}_{r_*}$; therefore, $\psi \in L_2(\mathbb{R}^3)$ implies $\psi = O(r^{-2})$ as $r \to \infty$.

Take a ball $B \subset B_{2r_*} \setminus \overline{B}_{r_*}$ (ensuring $q|_B = 0$) such that $\varphi|_B \neq 0$, where φ has been constructed above. Choose $\eta \in C_0^\infty(\mathbb{R}^3)$, $\operatorname{supp} \eta \subset B$; define a function $\varphi_\varepsilon := \varphi + \varepsilon\eta$ and a deformation of q by

$$q_\varepsilon := \frac{\Delta\varphi_\varepsilon}{\varphi_\varepsilon} = \frac{\Delta\varphi + \varepsilon\Delta\eta}{\varphi + \varepsilon\eta} = \frac{\Delta\varphi}{\varphi} + \varepsilon\frac{\Delta\eta}{\varphi} + o(\varepsilon) = q + \varepsilon\frac{\Delta\eta}{\varphi} + o(\varepsilon).$$

Note that q_ε is analytic with respect to ε and $q_\varepsilon = q$ out of B. We shall show that the assumption $\operatorname{Ker} H_{q_\varepsilon} \neq \{0\}$, $\varepsilon \in (0, \varepsilon_0)$, leads to a contradiction.

Reducing (3.8) to B_{2r_*}, we get an integral equation

$$\psi(x) = (G_q\psi)(x) := -\frac{1}{4\pi} \int_{B_{2r_*}} \frac{q(y)\psi(y)}{|x-y|}\,dy, \qquad x \in B_{2r_*}.$$

By the same arguments, for $\psi_\varepsilon \in \operatorname{Ker} H_{q_\varepsilon}$, we have

$$\psi_\varepsilon(x) = -\frac{1}{4\pi} \int_{B_{2r_*}} \frac{q_\varepsilon(y)\psi_\varepsilon(y)}{|x-y|}\,dy, \qquad x \in B_{2r_*}.$$

The analyticity implies $\psi_\varepsilon = \psi + \varepsilon h + o(\varepsilon)$ as $\varepsilon \to 0$. Substituting this asymptotic and the asymptotic of q_ε in the last equation, we easily obtain

$$h(x) + \frac{1}{4\pi}\int_{B_{2r_*}} \frac{q(y)h(y)}{|x-y|}\,dy = -\frac{1}{4\pi}\int_{B_{2r_*}} \frac{\frac{\Delta\eta}{\varphi}(y)\psi(y)}{|x-y|}\,dy =: g(x), \quad x \in B_{2r_*},$$

so that h solves this integral equation. Note also the asymptotic $g|_{r\to\infty} = r^{-1}$.

The solvability of the last equation is equivalent to the orthogonality condition $g \perp \mathrm{Ker}\,(I - G_q{}^*)$ in $L_2(B_{2r_*})$. In the meantime, writing $(-\Delta + q)\psi = 0$ in the form

$$\Delta\psi(x) = -\frac{q(x)}{4\pi}\int_{\mathbb{R}^3} \frac{\Delta\psi(y)}{|x-y|}\,dy \ = -\frac{q(x)}{4\pi}\int_{B_{2r_*}} \frac{\Delta\psi(y)}{|x-y|}\,dy, \qquad x \in \mathbb{R}^3,$$

we conclude that $\Delta\psi|_{B_{2r_*}} \in \mathrm{Ker}\,(I - G_q{}^*)$. Therefore, by the orthogonality one has

$$0 = (\Delta\psi, g)_{L_2(B_{2r_*})} = \int_{B_{2r_*}} \Delta\psi(x)g(x)\,dx = \int_{\mathbb{R}^3} \Delta\psi(x)g(x)\,dx.$$

Integrating by parts with regard to the asymptotics of ψ and g (as $r \to \infty$), we have

$$\begin{aligned} 0 &= \int_{\mathbb{R}^3} \psi(x)\Delta g(x)\,dx = \int_{\mathbb{R}^3} dx\ \psi(x)\left[\frac{\Delta\eta}{\varphi}\psi\right](x)\ dx \\ &= \int_B \eta(x)\left[\Delta\frac{\psi^2}{\varphi}\right](x)\,dx = 2\int_B \eta(x)\frac{|\psi(x)\nabla\varphi(x) - \varphi(x)\nabla\psi(x)|^2}{\varphi^3(x)}\,dx. \end{aligned}$$

In the last equality we have used the fact that φ and ψ are harmonic in B in view of $q|_B = 0$. By arbitrariness of η, we have $|\psi\nabla\varphi - \varphi\nabla\psi|\,|_B = 0$. Hence, φ and ψ are proportional in B and, by ellipticity, everywhere. The latter is impossible since ψ is smooth, whereas φ has a singularity at $x = 0$.

This contradiction shows that one can choose $\varepsilon > 0$ provided $\mathrm{Ker}\,H_{q_\varepsilon} = \{0\}$, and we assume such an ε chosen. Then, the corresponding function φ_ε belongs to $\mathcal{H}$ and satisfies $(-\Delta + q_\varepsilon)\,\varphi_\varepsilon = 0$ in $\mathbb{R}^3 \setminus \{0\}$, and $\varphi_\varepsilon = \frac{1}{r}$ near $x = 0$. Therefore, $\varphi_\varepsilon \in \mathcal{A}_{q_\varepsilon} \subset \mathcal{D}_{q_\varepsilon}$ and we have $\mathcal{D}_{q_\varepsilon} \neq \{0\}$ although $\mathrm{Ker}\ H_{q_\varepsilon} = \{0\}$. Thus, the condition $0 \in \sigma_p(H)$ is sufficient but not necessary for the lack of controllability.

3.4. General properties of $h \in \mathcal{D}$. Here we detail the behavior of unreachable elements near $x = 0$.

Lemma 3.2. *For $h \in \mathcal{D}$, the representation*

$$h(x) = \frac{\alpha(\omega)}{r} + \tilde{h}(x) \tag{3.9}$$

holds with $\alpha \in L_2(S^2)$ and $\tilde{h} \in H^1_{\mathrm{loc}}(\mathbb{R}^3)$.

Proof. By virtue of (2.9), an element $h \in \mathcal{D} = \mathrm{Ker}\,W^*$ satisfies $(W_0 - K)^*\,h = 0$; this implies

$$h = W_0 K^* h. \tag{3.10}$$

The control $f = K^*h$ can be written in the form

$$f = \int_0^{r^*} T_s^* W^*\ \hat{q}\ Y^* \sin\left[sL^{\frac{1}{2}}\right] L^{-\frac{1}{2}} Y h\,ds$$

(see (2.9)). Analyzing step by step this representation, we have $L^{-\frac{1}{2}}Yh \in H^1(B_{2r^*})$; a function $\hat{q}\,Y^* \sin[sL^{\frac{1}{2}}]L^{-\frac{1}{2}}Yh$ belongs to $H^1(\mathbb{R}^3)$ and is supported in B_{2r^*} for any s; W^* transfers this function to a control of the class $\mathcal{F}_1$ supported in $[0, 2r^*]$; this control is transferred by the left (reduced) shift T_s^* to a control of the same class supported on $[0, 2r^* - s]$. At last, the integration gives $f \in \mathcal{F}_1$, $\operatorname{supp} f \subset [0, 2r^*]$.

Representing by (3.10) $h = W_0 f$, in accordance with (1.31) one obtains

$$h = \frac{\alpha(\omega)}{r} + \tilde{h},$$

where $\alpha = [f(0, \cdot)]_0$ and $\tilde{h} = \frac{1}{2\pi} \int_{S^2} f_\tau(x \cdot \theta, \theta)\, d\theta \in H^1_{\mathrm{loc}}(\mathbb{R}^3)$. $\square$

If $h \in \operatorname{Ker} H \subset \mathcal{D}$, then h is smooth and α in (3.9) equals zero. The singular element $\varphi \in \mathcal{D}$ constructed in section 3.3 is of the form $\varphi = \frac{1}{r} + \tilde{h}$, so that its α equals 1. In this connection, the following question remains open: Do there exist the defect elements with variable α? Furthermore, we know no simple conditions (besides a smallness of q) ensuring $\mathcal{D} = \{0\}$. For instance, does $q \geq 0$ imply $\mathcal{D} = \{0\}$? Is the presence of the negative spectrum of H necessary for $\mathcal{D} \neq \{0\}$? These natural questions are so far open.

3.5. Reversing waves and s-points. Let the potential q be such that $\mathcal{D} \neq \{0\}$. In this case, $\mathcal{N} \neq \{0\}$ (see (0.6)) and there exists a null control $f \neq 0$ such that $u^f(\cdot, 0) = Wf = 0$. It is easily proved that, owing to the last equality, the function

$$u^f_-(\cdot, t) := \begin{cases} u^f(\cdot, t), & t \leq 0, \\ -u^f(\cdot, -t), & t > 0, \end{cases}$$

turns out to be an L_2-solution of problem (0.1)–(0.3) in $\mathbb{R}^3 \times (-\infty, \infty)$. This solution is odd with respect to time and vanishes in $\{(x, t) \mid |x| < |t|\}$, i.e., in the past and future light cones simultaneously.

As a consequence, the function

$$w^f(x, t) = \int_{-\infty}^{t} u^f(x, s)\, ds, \quad (x, t) \in \mathbb{R}^3 \times (-\infty, \infty),$$

is an even L_2-solution of the acoustical equation in $\mathbb{R}^3 \times (-\infty, \infty)$ also vanishing in both of the cones. As one can show, w^f is a finite energy solution:

$$\int_{\mathbb{R}^3} \left\{ \left[w^f_t(x, t) \right]^2 + \left| \nabla w^f(x, t) \right|^2 + q(x) \left[w^f(x, t) \right]^2 \right\} dx = \mathrm{const} > 0$$

and, hence, its Cauchy data $\{w^f(\cdot, 0), w^f_t(\cdot, 0)\} = \{y, 0\}$ belong to the energy class (see [10], [13]). At the same time, since $w^f \mid_{|x|<|t|} = 0$, the pair $\{y, 0\}$ belongs to the classes D_- and D_+ of the incoming and outgoing data simultaneously. Therefore, we get an example of $\{y, 0\} \in D_+ \cap D_- \neq \{0\}$. Note that such an intersection is possible even in the case Ker $H = \{0\}$.

From the physical point of view, the behavior of the waves u^f_- and w^f is very curious. They come from infinity, penetrate into supp q, and interact with the potential. Then the waves return back to infinity, leaving supp q with no trace remaining. The absence of the residual perturbation in supp q is amazing and, probably, may be interpreted as a subtle interference effect. The existence of such reversing waves is equivalent to the lack of controllability of system (0.1)–(0.3).

In the effect described above, the origin $x = 0$ plays the role of a point, which in a sense stops the reversing waves. This observation motivates the following definition. For $s \in \mathbb{R}^3$, we denote the shift of q by $q_s(x) = q(x - s)$ (recall that $\sigma(H_{q_s}) = \sigma(H_q)$). An $s \in \mathbb{R}^3$ is said to be a *stop point* (s-*point*) of the potential q if $\operatorname{Ker} H_q = \{0\}$ and $\mathcal{D}_{q_s} \neq \{0\}$. The set of s-points is denoted by $\Upsilon[q]$.

Loosely speaking, the potential q makes the space inhomogeneous and the set $\Upsilon[q]$ consists of points that can stop incoming waves. We hope this set is worthy of further investigation. At the moment our knowledge about $\Upsilon[q]$ is exhausted by the following observations.

- *Connection with the operator extensions*: Assuming again $\operatorname{Ker} H = \{0\}$, fix $s \in \mathbb{R}^3$, and consider the operators $L_p : \mathcal{H} \to \mathcal{H}$, $\operatorname{Dom} L_p = \{y \in C_0^\infty(\mathbb{R}^3) \mid y = 0 \text{ near } s\}$, $L_p y = (-\Delta + q)^p y$; $p = 1, 2, \ldots$. Every L_p is a densely defined symmetric semibounded operator and hence has the self-adjoint semibounded extensions. If at least one L_p possesses an extension $\tilde{L_p}$ with $\operatorname{Ker} \tilde{L_p} \neq \{0\}$, then $s \in \Upsilon[q]$. Perhaps this gives a characterization of s-points.
- *Connection with the factorization problem*: If $x = 0 \in \Upsilon[q]$, then the s-matrix associated with potential q does not admit the factorization in the sense of Newton [11].
- *Model example*: Consider a singular potential $q = \delta(x)$. Understanding $-\Delta - \delta$ in a proper sense (see [6], [12]), one can show that this operator has only one eigenvalue $\lambda = -1$ and $\Upsilon[\delta] = S^2$. Other examples of this sort motivate the following conjecture: any potential with negative spectrum has a nonempty set of s-points.

3.6. Stability. Here we present necessary and sufficient conditions providing a global L_2-boundedness of a trajectory of system (0.1)–(0.3). We use the same notations as in (2.5), (2.23) and denote $g_k := W^* \varphi_k \in \mathcal{F}$ (see (2.25)).

LEMMA 3.3. *The estimate $\|u^f(\cdot, t)\|^2_{\mathcal{H}} \le c\|f\|_{\mathcal{F}}$ holds for all $t \in (-\infty, \infty)$ if and only if $(f, g_k)_{\mathcal{F}} = 0$, $k = 1, 2, \ldots, p$.*

Proof. (i) *Orthogonality $\Rightarrow$ stability*: For $t \le 0$ one has

$$\|u^f(\cdot, t)\|_{\mathcal{H}} = \langle \text{see } (2.8) \rangle = \|W T_t f\|_{\mathcal{H}} \le \|W\| \|f\|_{\mathcal{F}},$$

so that $u^f \mid_{t<0}$ is bounded. Suppose $t > 0$. Take $f \in C_0^\infty(\Sigma)$ and find the Fourier coefficients in (2.5):

$$\begin{aligned}
&\cosh \varkappa_k t \left(u^f(\cdot, 0), \varphi_k\right)_{\mathcal{H}} + \frac{\sinh \varkappa_k t}{\varkappa_k} \left(u^f_t(\cdot, 0), \varphi_k\right)_{\mathcal{H}} \\
&= \frac{e^{\varkappa_k t}}{2} \left(u^f(\cdot, 0) + \frac{1}{\varkappa_k}\, u^f_t(\cdot, 0), \varphi_k\right)_{\mathcal{H}} + \frac{e^{-\varkappa_k t}}{2} \left(u^f(\cdot, 0) - \frac{1}{\varkappa_k}\, u^f_t(\cdot, 0), \varphi_k\right)_{\mathcal{H}} \\
&= \frac{e^{\varkappa_k t}}{2} \left(W \left[f + \frac{1}{\varkappa_k}\, f_\tau\right], \varphi_k\right)_{\mathcal{H}} + \frac{e^{-\varkappa_k t}}{2} \left(W \left[f - \frac{1}{\varkappa_k}\, f_\tau\right], \varphi_k\right)_{\mathcal{H}} \\
&= \langle \text{see } (2.25) \rangle = \frac{e^{\varkappa_k t}}{2} \left(f + \frac{1}{\varkappa_k}\, f_\tau, g_k\right)_{\mathcal{F}} + \frac{e^{-\varkappa_k t}}{2} \left(f - \frac{1}{\varkappa_k}\, f_\tau, g_k\right)_{\mathcal{F}} = e^{\varkappa_k t} \left(f, g_k\right)_{\mathcal{F}}
\end{aligned}$$

(we have used $(\frac{1}{\varkappa_k} f_\tau, g_k)_{\mathcal{F}} = (f, g_k)_{\mathcal{F}}$) and

$$\begin{aligned}
&1(t) \left(u^f(\cdot, 0), \varphi^0_k\right)_{\mathcal{H}} + t \left(u^f_t(\cdot, 0), \varphi^0_k\right)_{\mathcal{H}} \\
&\quad = 1(t) \left(W f, \varphi^0_k\right)_{\mathcal{H}} + t \left(W f_\tau, \varphi^0_k\right)_{\mathcal{H}} = \langle \text{see } (2.28) \rangle = 0.
\end{aligned}$$

Now (2.5) can be written in the form

$$u^f(\cdot,t) = \sum_{k=1}^{p} e^{\varkappa_k t}(f, g_k)_{\mathcal F}\ \varphi_k$$
$$+\int_{+0}^{\infty}\left\{\cos\sqrt{\lambda}t\, dE_\lambda\, u^f(\cdot,0) + \frac{\sin\sqrt{\lambda}t}{\sqrt{\lambda}}\, dE_\lambda\, u^f_t(\cdot,0)\right\}, \tag{3.11}$$

whereas the orthogonality condition leads to

$$u^f(\cdot,t) = \int_{+0}^{\infty}\left\{\cos\sqrt{\lambda}t\, dE_\lambda\, u^f(\cdot,0) + \frac{\sin\sqrt{\lambda}t}{\sqrt{\lambda}}\, dE_\lambda u^f_t(\cdot,0)\right\}, \tag{3.12}$$

so that $u^f(\cdot,t)$ belongs to the subspace $\mathcal{H}_+ = (I - E_{-0})\,\mathcal{H}$ of absolutely continuous spectrum of H. Since $f \perp g_k$ implies $f_\tau \perp g_k$, the same is valid for $u^{f_\tau(\cdot,t)} = u^f_t(\cdot,t) \in \mathcal{H}_+$.

To estimate the norm of the wave u^f represented by (3.12) let us assume in addition that $(Jf)(\tau,\omega) = \int_0^\tau f(s,\omega)\,ds \in C_0^\infty(\Sigma)$. The set of such f's is dense in $\mathcal{F}$ and $u^{Jf}(\cdot,0) \in \operatorname{Dom} H$, $u^{Jf}(\cdot,0) \perp \operatorname{Ker} H$. Now, we have the inequality

$$\|u^f(\cdot,t)\|^2_{\mathcal H} \le 2\left[\left\|\int_{+0}^{\infty}\cos\sqrt{\lambda}t\, dE_\lambda\, u^f(\cdot,0)\right\|^2_{\mathcal H}\right.$$
$$\left. + \left\|\int_{+0}^{\infty}\frac{\sin\sqrt{\lambda}t}{\sqrt{\lambda}}\, dE_\lambda\, u^f_t(\cdot,0)\right\|^2_{\mathcal H}\right] =: 2\,[I(t) + II(t)] \tag{3.13}$$

and the evident estimate

$$I(t) \le \|u^f(\cdot,0)\|^2_{\mathcal H} \le \|W\|^2 \|f\|^2_{\mathcal F}. \tag{3.14}$$

The part of the operator H in $\mathcal{H} \ominus \operatorname{Ker} H$ is injective; its part $H_+ := H\,(I - E_{-0})$ acting in $\mathcal{H}_+$ is a positive injective operator. This enables one to represent

$$u^{Jf}(\cdot,0) = H^{-1}Hu^{Jf}(\cdot,0) = H^{-1}\left(-u^{Jf}_{tt}(\cdot,0)\right)$$
$$= -H^{-1}u^{(Jf)_{tt}}(\cdot,0) = -H^{-1}u^{f_\tau}(\cdot,0) = -H^{-1}u^f_t(\cdot,0)$$

and to get

$$\left(u^{Jf}(\cdot,0), u^f_t(\cdot,0)\right)_{\mathcal H} = -\left(H^{-1}u^f_t(\cdot,0), u^f_t(\cdot,0)\right)_{\mathcal H}$$
$$= -\left\|H_+^{-\frac12}u^f_t(\cdot,0)\right\|^2_{\mathcal H}.$$

On the other hand,

$$\left(u^{Jf}(\cdot,0), u^f_t(\cdot,0)\right)_{\mathcal H} = \left(u^{Jf}(\cdot,0), u^{f_\tau}(\cdot,0)\right)_{\mathcal H}$$
$$= \langle\text{see } (2.31), (2.32)\rangle = \left([I+R]\,Jf, f_\tau\right)_{\mathcal F}$$
$$= -\left(\frac{\partial}{\partial\tau}\,[I+R]\,Jf, f\right)_{\mathcal F} = \langle\text{see } (2.30)\rangle$$
$$= -\left([I-R]\,f, f\right)_{\mathcal F} = \langle\text{see } (2.32)\rangle$$
$$= -\left([2I - C]\,f, f\right)_{\mathcal F},$$

and we obtain

$$\begin{aligned}\left\|H_+^{-\frac{1}{2}} u_t^f(\cdot,0)\right\|_{\mathcal{H}}^2 &= ([2I-C]\,f,f)_{\mathcal{F}} \\ &\leq (2+\|C\|)\,\|f\|_{\mathcal{F}}^2 = \left(2+\|W\|^2\right)\|f\|_{\mathcal{F}}^2.\end{aligned}$$

Thereafter, we can estimate the second norm in (3.13):

$$\begin{aligned}II(t) &= \left\|\int_{+0}^{\infty} \frac{\sin\sqrt{\lambda}t}{\sqrt{\lambda}}\,dE_\lambda\,u_t^f(\cdot,0)\right\|_{\mathcal{H}}^2 = \left\|\int_{+0}^{\infty} \sin\sqrt{\lambda}t\,dE_\lambda H_+^{-\frac{1}{2}}\,u_t^f(\cdot,0)\right\|_{\mathcal{H}}^2 \\ &\leq \left\|H_+^{-\frac{1}{2}} u_t^f(\cdot,0)\right\|_{\mathcal{H}}^2 \leq \left(2+\|W\|^2\right)\|f\|_{\mathcal{F}}^2.\end{aligned}$$

Substituting this and (3.14) in (3.13), we get the estimate

$$\left\|u^f(\cdot,t)\right\|_{\mathcal{H}}^2 \leq \left(3+\|W\|^2\right)\|f\|_{\mathcal{F}}^2 \tag{3.15}$$

on a dense set of controls f. Extending (3.15) to $\mathcal{F}$, we conclude that $u^f(\cdot,t)$ is bounded in $\mathcal{H}$ uniformly with respect to $t\in(-\infty,\infty)$.

(ii) *Stability* $\Rightarrow$ *orthogonality*: If the orthogonality does not hold, by (3.11) we have

$$\left\|u^f(\cdot,t)\right\|_{\mathcal{H}}^2 \geq \sum_{k=1}^{p} e^{2\varkappa_k t}(f,g_k)_{\mathcal{F}}^2 \to \infty \quad \text{as } t\to\infty,$$

so that $u^f\,|_{t>0}$ is unbounded, i.e., the trajectory is not stable. $\square$

3.7. Comments.

- Probably, the equality $\mathcal{A}^\xi = \mathcal{D}^\xi$, $\xi>0$ (see (0.5)), is valid and not only for radial potentials.
- Let us repeat once again an open problem of the principal nature: Is $\sigma_p(H)\neq\{\emptyset\}$ equivalent to $\mathcal{D}\neq\{0\}$?
- As we saw in section 3.5, if $0\in\Upsilon[q]$ and $f\in\mathcal{N}$, then the solution u_-^f is a reversing wave satisfying $u_-^f\,|_{|x|<|t|}=0$. Its Fourier transform $\tilde{u}_-^f(x,k) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ikt}u_-^f(x,t)\,dt = \frac{1}{2\pi}\int_{-|x|}^{|x|} e^{ikt}u_-^f(x,t)\,dt$ satisfies

  $$\left[-\Delta+(q-k^2)\right]\tilde{u}(\cdot,k)=0 \quad \text{in } \mathbb{R}^3$$

  and, hence, is an eigenfunction of continuous spectrum of H, which is an *entire function* of k. Such eigenfunctions are associated with each point of $\Upsilon[q]$. It would be interesting to compare them with the entire (with respect to k) solutions introduced in [11] and investigate their properties: completeness, orthogonality, asymptotic as $|x|,|k|\to\infty$, etc.

Acknowledgments. We thank W. Lionheart for kind consultations and are grateful to the referees for many suggested corrections and helpful remarks.

REFERENCES

[1] S. A. Avdonin, M. I. Belishev, and S. I. Ivanov, *Controllability in the filled domain for the wave equation with a singular boundary control*, Zap. Nauch. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI), 210 (23) (1994), pp. 3–14 (in Russian); J. Sov. Math., 83 (1997), pp. 165–174 (in English).

[2] M. I. Belishev, *Boundary control in reconstruction of manifolds and metrics (the BC-method)*, Inverse Problems, 13 (1997), pp. R1–R45.

[3] M. I. Belishev and A. K. Glasman, *Dynamical inverse problem for the Maxwell system: Recovering the velocity in a regular zone (the BC–method)*, St. Petersburg Math. J., 12 (2001), pp. 279–316.

[4] M. I. Belishev and I. Lasiecka, *The dynamical Lame system: Regularity of solutions, boundary controllability and boundary data continuation*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 143–167.

[5] M. I. Belishev and A. F. Vakulenko, *On a control problem for the wave equation in $\mathbb{R}^3$*, Zap. Nauch. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI), 332 (2006), pp. 19–37 (in Russian); J. Math. Sci. (N.Y.), to appear.

[6] F. A. Berezin and L. D. Faddeev, *Remark on the Schrödinger equation with singular potential*, Dokl. Akad. Nauk SSSR, 137 (1961), pp. 1011–1014 (in Russian).

[7] I. M. Gelfand, S. G. Gindikin, and M. I. Graev, *Selected Topics in Integral Geometry*, Transl. Math. Monogr. 220, AMS, Providence RI, 2003.

[8] S. Helgason, *The Radon Transform*, Birkhäuser, Boston, Basel, Stuttgart, 1980.

[9] E. M. Landis, *On some properties of solutions of the elliptic equations*, Dokl. Akad. Nauk SSSR, 107 (1961), pp. 640–643 (in Russian).

[10] P. Lax and R. Phillips, *Scattering Theory*, Academic Press, New York, London, 1967.

[11] R. Newton, *Inverse scattering.* II. *Three dimensions*, J. Math. Phys., 21 (1980), pp. 1698–1715.

[12] B. S. Pavlov, *The theory of extensions and explicitly solvable models*, Uspekhi Mat. Nauk, 42 (1987), pp. 99–131 (in Russian).

[13] R. S. Phillips, *Scattering theory for the wave equation with a short range potential*, Indiana Univ. Math. J., 31 (1982), pp. 609–635.

[14] M. Reed and B. Simon, *Methods of Modern Mathematical Physics.* IV: *Analysis of Operators*, Academic Press, New York, San Francisco, London, 1978.

[15] V. S. Vladimirov, *Equations of Mathematical Physics*, Pure Appl. Math. 3, Marcel Dekker, New York, 1971.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 1851–1862

DECAY ESTIMATES OF A TANGENTIAL DERIVATIVE TO THE LIGHT CONE FOR THE WAVE EQUATION AND THEIR APPLICATION*

SOICHIRO KATAYAMA† AND HIDEO KUBO‡

Abstract. We consider wave equations in three space dimensions and obtain new weighted L^∞-L^∞ estimates for a tangential derivative to the light cone. As an application, we give a new proof of the global existence theorem, which was originally proved by Klainerman and Christodoulou, for systems of nonlinear wave equations under the null condition. Our new proof has the advantage of using neither the scaling nor the Lorentz boost operators.

Key words. nonlinear wave equation, null condition, global existence

AMS subject classification. 35L70

DOI. 10.1137/070694417

1. Introduction.

Solutions to the Cauchy problem for nonlinear wave equations with quadratic nonlinearity in three space dimensions may blow up in finite time no matter how small initial data are, and we have to impose some special condition on the nonlinearity to get global solutions. The null condition is one of such conditions and is associated with the null forms Q_0 and Q_{ab}, which are given by

$$Q_0(v,w;c) = (\partial_t v)(\partial_t w) - c^2(\nabla_x v)\cdot(\nabla_x w), \tag{1.1}$$

$$Q_{ab}(v,w) = (\partial_a v)(\partial_b w) - (\partial_b v)(\partial_a w) \quad (0 \le a < b \le 3) \tag{1.2}$$

for $v = v(t,x)$ and $w = w(t,x)$, where c is a positive constant corresponding to the propagation speed, $\partial_0 = \partial_t = \partial/\partial t$, and $\partial_j = \partial/\partial x_j$ $(j = 1,2,3)$. More precisely, let $c > 0$ and consider the Cauchy problem for

$$\square_c u_i = F_i(u, \partial u, \nabla_x \partial u) \quad \text{in } (0,\infty)\times\mathbb{R}^3 \quad (1 \le i \le m) \tag{1.3}$$

with initial data

$$u = \varepsilon f \text{ and } \partial_t u = \varepsilon g \quad \text{at } t = 0, \tag{1.4}$$

where $\square_c = \partial_t^2 - c^2\Delta_x$, $u = (u_j)$, $\partial u = (\partial_a u_j)$, and $\nabla_x \partial u = (\partial_k \partial_a u_j)$ with $1 \le j \le m$, $1 \le k \le 3$, and $0 \le a \le 3$, while ε is a positive parameter. Let $F = (F_i)_{1\le i\le m}$ be quadratic around the origin in its arguments and the system be quasi-linear. In other words, we assume that each F_i has the form

$$F_i(u, \partial u, \nabla_x \partial u) = \sum_{\substack{1\le j\le m\\ 1\le k\le 3,\ 0\le a\le 3}} c_{ka}^{ij}(u,\partial u)\partial_k\partial_a u_j + d_i(u,\partial u), \tag{1.5}$$

*Received by the editors June 13, 2007; accepted for publication (in revised form) October 30, 2007; published electronically February 22, 2008.
http://www.siam.org/journals/sima/39-6/69441.html

†Department of Mathematics, Wakayama University, 930 Sakaedani, Wakayama 640-8510, Japan (katayama@center.wakayama-u.ac.jp). This author's research was partially supported by Grant-in-Aid for Young Scientists (B) (16740094), MEXT.

‡Department of Mathematics, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan (kubo@math.sci.osaka-u.ac.jp). This author's research was partially supported by Grant-in-Aid for Science Research (17540157), JSPS.

where $c^{ij}_{ka}(u, \partial u) = O(|u| + |\partial u|)$ and $d_i(u, \partial u) = O(|u|^2 + |\partial u|^2)$ around $(u, \partial u) = (0, 0)$. Without loss of generality, we may assume $c^{ij}_{k\ell} = c^{ij}_{\ell k}$ for $1 \le i, j \le m$ and $1 \le k, \ell \le 3$. In addition, we always assume the symmetry condition

$$c^{ij}_{ka} = c^{ji}_{ka} \quad \text{for } 1 \le i, j \le m, \quad 1 \le k \le 3, \text{ and } 0 \le a \le 3.$$

Then it is well known that the null condition (for the above system (1.3)) is satisfied if and only if the quadratic terms of F_i $(1 \le i \le m)$ can be written as linear combinations of the null forms $Q_0(u_j, \partial^\alpha u_k; c)$ and $Q_{ab}(u_j, \partial^\alpha u_k)$ with $1 \le j, k \le m$, $0 \le a < b \le 3$, and $|\alpha| \le 1$, where $\partial^\alpha = \partial_0^{\alpha_0} \partial_1^{\alpha_1} \partial_2^{\alpha_2} \partial_3^{\alpha_3}$ for a multi-index $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ (refer to [3] and [14] for the precise description of the null condition). Klainerman [14] and Christodoulou [3] proved the following global existence theorem independently by different methods.

THEOREM 1.1 (Klainerman [14], Christodoulou [3]). *Suppose that the null condition is satisfied. Then, for any f, $g \in C_0^\infty(\mathbb{R}^3; \mathbb{R}^m)$, there exists a positive constant ε_0 such that the Cauchy problem* (1.3)–(1.4) *admits a unique global solution $u \in C^\infty([0, \infty) \times \mathbb{R}^3; \mathbb{R}^m)$ for any $\varepsilon \in (0, \varepsilon_0]$.*

Christodoulou used the so-called conformal method which is based on Penrose's conformal compactification of Minkowski space. On the other hand, Klainerman used the vector field method and showed the above theorem by deriving some decay estimates in the original coordinates. In Klainerman's proof, he introduced vector fields

$$L_{c,j} = \frac{x_j}{c}\partial_t + ct\partial_j \quad (1 \le j \le 3), \quad \Omega_{ij} = x_i \partial_j - x_j \partial_i \quad (1 \le i < j \le 3),$$

which are the generators of the Lorentz group, and the scaling operator

$$S = t\partial_t + x \cdot \nabla_x.$$

These vector fields play an important role in getting Klainerman's weighted L^1-L^∞ estimates for wave equations (see also Hörmander [5]). In addition, using them, we can see that an extra decay factor is expected from the null forms. For example, we have

$$(1.6) \qquad Q_0(v, w; c) = \frac{1}{t + r}\Big\{(\partial_t v)\big(Sw + cL_{c,r}w\big) - c\sum_{j=1}^{3}(L_{c,j}v)(\partial_j w) - c^2(Sv)(\partial_r w) + c^2 \sum_{j \ne k} \omega_k (\Omega_{jk} v)(\partial_j w)\Big\},$$

where $r = |x|$, $\omega = (\omega_1, \omega_2, \omega_3) = x/r$, $\partial_r = \sum_{j=1}^3 \omega_j \partial_j$, $L_{c,r} = \sum_{j=1}^3 \omega_j L_{c,j}$, and $\Omega_{ij} = -\Omega_{ji}$ for $1 \le j < i \le 3$.

Among the above vector fields, the Lorentz boost fields $L_{c,j}$ depend on the propagation speed c, and they are unfavorable when we consider the multiple speed case. Thus, the vector field method without the Lorentz boost fields was developed by many authors (see Kovalyov [17, 18], Klainerman and Sideris [16], Yokoyama [25], Kubota and Yokoyama [19], Sideris and Tu [23], Sogge [24], Hidano [4], Katayama [9, 11], and Katayama and Yokoyama [13], for example). In place of (1.6), the following identity was used in the above works relating to the null condition for the multiple speed case:

$$
\begin{aligned}
(1.7)\qquad Q_0(v,w;c) = {} & \frac{1}{t^2}(Sv+(ct-r)\partial_r v)(Sw-(ct+r)\partial_r w) \\
& + \frac{c}{t}\left\{(Sv)(\partial_r w)-(\partial_r v)(Sw)\right\} + \frac{c^2}{r}\sum_{j\neq k}\omega_k(\partial_j v)(\Omega_{jk}w),
\end{aligned}
$$

whose variant was introduced by Hoshiga and Kubo [6]. Equation (1.7) leads to a good estimate in the region $r > \delta t$ with some small $\delta > 0$, because r is equivalent to $t+r$ in this region. Note that the operator S is still used in (1.7), and this is the only reason why S was adopted in [9, 19, 25], because these works are based on variants of L^∞-L^∞ estimates due to John [7] and Kovalyov [17], where only ∂_a and Ω_{ij} are used (see Lemma 3.2 below).

Our aim here is to get rid of not only $L_{c,j}$, but also S from the estimate of the null forms, and prove Theorem 1.1 using only ∂_a and Ω_{jk}. Though the usage of the scaling operator S has not caused any serious difficulty in the study of the Cauchy problem for nonlinear wave equations so far, we believe that it is worthwhile developing a simple approach with a smaller set of vector fields. For this purpose, we make use of the identity

$$
\begin{aligned}
(1.8)\qquad Q_0(v,w;c) = {} & \frac{1}{2}\big\{(D_{+,c}v)(D_{-,c}w)+(D_{-,c}v)(D_{+,c}w)\big\} \\
& + \frac{c^2}{r}\sum_{j\neq k}\omega_k(\partial_j v)(\Omega_{jk}w),
\end{aligned}
$$

where $D_{\pm,c} = \partial_t \pm c\partial_r$. Note that this identity was already used implicitly to obtain identities like (1.7) (see [23], for example). In view of (1.8), what we need to treat the null forms is an enhanced decay estimate for the tangential derivative $D_{+,c}$ to the light cone. We can say that, in the previous works, this enhanced decay has been observed through

$$
D_{+,c} = \frac{1}{t}\big(S+(ct-r)\partial_r\big) \text{ or } D_{+,c} = \frac{1}{ct+r}\big(cS+cL_{c,r}\big)
$$

with the help of S or also $L_{c,r} = \sum_{j=1}^3 \omega_j L_{c,j}$.

In this paper, we take a different approach. We will establish the enhanced decay of $D_{+,c}u$ for the solution u to the wave equation directly. We formulate it as a weighted L^∞-L^∞ estimate in Theorem 2.1 below, which is our main ingredient in this paper. The point is that such an estimate can be derived by using only ∂_a and Ω_{ij}. This type of approach to $D_{+,c}$ goes back to the work of John [8].

2. The main result. Before stating our result precisely, we introduce several notations. We put $Z = \{Z_a\}_{1\le a\le 7} = \{(\partial_a)_{0\le a\le 3}, (\Omega_{jk})_{1\le j<k\le 3}\}$. For a multi-index $\alpha = (\alpha_1,\dots,\alpha_7)$, we define $Z^\alpha = Z_1^{\alpha_1}Z_2^{\alpha_2}\cdots Z_7^{\alpha_7}$. For a function $v = v(t,x)$ and a nonnegative integer s, we define

$$
(2.1)\qquad |v(t,x)|_s = \sum_{|\alpha|\le s}|Z^\alpha v(t,x)| \text{ and } \|v(t,\cdot)\|_s = \big\||v(t,\cdot)|_s\big\|_{L^2(\mathbb{R}^3)}.
$$

We put $\langle a\rangle = \sqrt{1+a^2}$ for $a \in \mathbb{R}$. Let c be a positive constant, and we fix arbitrary positive constants c_j $(1 \le j \le N)$ (our theorem is true for any choice of these constants c_j, but when we apply our estimate to nonlinear problems, we usually choose c_j as the propagation speeds and N as the number of different propagation speeds in the system; c is also chosen from these propagation speeds). We define

$$w(t,r) = w(t,r;c_1,\dots,c_N) = \min_{0\le j\le N} \langle c_j t - r\rangle \tag{2.2}$$

with $c_0 = 0$, and we define

$$A_{\rho,\mu,s}[G;c](t,x) = \sup_{(\tau,y)\in\Lambda_c(t,x)} |y|\,\langle \tau+|y|\rangle^{\rho}\, w(\tau,|y|)^{1+\mu}|G(\tau,y)|_s \tag{2.3}$$

for $\rho,\mu \ge 0$, a nonnegative integer s, and a smooth function $G = G(t,x)$, where $\Lambda_c(t,x) = \{(\tau,y)\in[0,t]\times\mathbb{R}^3\,;\, |y-x|\le c(t-\tau)\}$. We also define

$$B_{\rho,s}[\phi,\psi;c](t,x) = \sup_{y\in\Lambda'_c(t,x)} \langle |y|\rangle^{\rho}\left(|\phi(y)|_{s+1} + |\psi(y)|_s\right) \tag{2.4}$$

for $\rho \ge 0$, a nonnegative integer s, and smooth functions ϕ and ψ on $\mathbb{R}^3$, where $\Lambda'_c(t,x) = \{y\in\mathbb{R}^3\,;\,|y-x|\le ct\}$.

The following theorem is our main result.

Theorem 2.1. *Assume $1\le\kappa\le 2$ and $\mu>0$.*

(i) *Let u be the solution to*

$$\Box_c u = G \quad \text{in } (0,\infty)\times\mathbb{R}^3$$

with initial data $u = \partial_t u = 0$ at $t=0$. Then there exists a positive constant C, depending on κ and μ, such that

$$\langle |x|\rangle\,\langle t+|x|\rangle\,\langle ct-|x|\rangle^{\kappa-1}\{\log(2+t+|x|)\}^{-1}\,|D_{+,c}u(t,x)| \le CA_{\kappa,\mu,2}[G;c](t,x) \tag{2.5}$$

for $(t,x)\in(0,\infty)\times\mathbb{R}^3$ with $x\ne 0$, where $A_{\kappa,\mu,2}$ is given by (2.3).

Moreover, if $1<\kappa<2$, then for any $\delta>0$, there exists a constant C, depending on κ, μ, and δ, such that

$$\langle t+|x|\rangle^2\,\langle ct-|x|\rangle^{\kappa-1}\,|D_{+,c}u(t,x)| \le CA_{\kappa,\mu,2}[G;c](t,x) \tag{2.6}$$

for $(t,x)\in(0,\infty)\times\mathbb{R}^3$ satisfying $|x|>\delta t$.

(ii) *Let u^* be the solution to*

$$\Box_c u^* = 0 \text{ in } (0,\infty)\times\mathbb{R}^3$$

with initial data $u^* = \phi$ and $\partial_t u^* = \psi$ at $t=0$. Then we have

$$\langle |x|\rangle\,\langle t+|x|\rangle\,\langle ct-|x|\rangle^{\kappa-1}\,|D_{+,c}u^*(t,x)| \le CB_{\kappa+\mu+1,2}[\phi,\psi;c](t,x) \tag{2.7}$$

for $(t,x)\in(0,\infty)\times\mathbb{R}^3$ with $x\ne 0$, where $B_{\kappa+\mu+1,2}$ is given by (2.4).

Remark.

1. Similar estimates for radially symmetric solutions are obtained by Katayama [11].

2. Suppose that $A_{\kappa,\mu,2}[G;c](t,x)$ is bounded on $[0,\infty)\times\mathbb{R}^3$ for some $\kappa\in[1,2)$ and $\mu>0$ and that u solves $\Box_c u=G$ with zero initial data. Then, from Lemma 3.2 below, we see that u and ∂u decay like $\langle t\rangle^{-1}\,\Psi_{\kappa-1}(t)$ along the light cone $ct=|x|$, where $\Psi_\rho(t)=\log(2+t)$ if $\rho=0$, and $\Psi_\rho(t)=1$ if $\rho>0$. Compared with this decay rate, we find from (2.5) and (2.6) that $D_{+,c}u$ gains extra decay of $\langle t\rangle^{-1}$ and behaves like $\langle t\rangle^{-2}\,\Psi_{\kappa-1}(t)$ along the light cone.

3. For tangential derivatives $T_{c,j}=(x_j/|x|)\partial_t+c\partial_j$ $(1\le j\le 3)$, Alinhac showed that

$$\left(\int_0^t\int_{\mathbb{R}^3}\big(1+\big|c\tau-|x|\big|\big)^{-\rho}|T_{c,j}u(\tau,x)|^2dxd\tau\right)^{1/2}$$

with $\rho>1$ is bounded by $\|\partial u(0,\cdot)\|_{L^2(\mathbb{R}^3)}+\int_0^t\|\Box_c u(\tau,\cdot)\|_{L^2(\mathbb{R}^3)}d\tau$ (see [1], for example). Observe that $T_{c,j}$ is closely connected to $D_{+,c}$. In fact, we have $D_{+,c}=\sum_{j=1}^3(x_j/|x|)T_{c,j}$. Though Alinhac's estimate does not need S and means enhanced decay of tangential derivatives implicitly, it seems difficult to recover a pointwise decay estimate from his weighted space-time estimate. On the other hand, Sideris and Thomases [22] obtained the estimate for $\big\|\big(1+\big|ct+|\cdot|\big|\big)T_{c,j}u(t,\cdot)\big\|_{L^2(\mathbb{R}^3)}$; however, S is used in their estimate.

4. The exterior problem for systems of nonlinear wave equations with the single or multiple speed(s) is also widely studied (see Metcalfe, Nakamura, and Sogge [20] and Metcalfe and Sogge [21] and the references cited therein). In the exterior domains, because of their unbounded coefficients on the boundary, the Lorentz boosts are unlikely to be applicable even for the single speed case. This is another reason why the vector field method without the Lorentz boosts is widely studied. In addition, S also causes a technical difficulty in the exterior problems. We will discuss the exterior problem in a subsequent paper, and we will not go into further details here.

We will prove Theorem 2.1 in the next section, after stating some known weighted L^∞-L^∞ estimates for wave equations. Though we can apply our theorem to exclude S from the proof of the multiple speed version of Theorem 1.1 in [9, 19, 25], we concentrate on the single speed case for simplicity, and we will give a new proof, without using S and $L_{c,j}$, of Theorem 1.1 in section 4 as an application of our main theorem.

Throughout this paper, various positive constants, which may change line by line, are denoted just by the same letter C.

3. Proof of Theorem 2.1. For $c>0$, $\phi=\phi(x)$, and $\psi=\psi(x)$, we write $U_c^*[\phi,\psi]$ for the solution u to the homogeneous wave equation $\Box_c u=0$ in $(0,\infty)\times\mathbb{R}^3$ with initial data $u=\phi$ and $\partial_t u=\psi$ at $t=0$. Similarly, for $c>0$ and $G=G(t,x)$, we write $U_c[G]$ for the solution u to the inhomogeneous wave equation $\Box_c u=G$ in $(0,\infty)\times\mathbb{R}^3$ with initial data $u=\partial_t u=0$ at $t=0$.

For $U_c^*[\phi,\psi]$ we have the following.

Lemma 3.1. *Let $c>0$. Then, for $\kappa>1$, we have*

$$\begin{aligned}(3.1)\qquad &\langle t+|x|\rangle\,\langle ct-|x|\rangle^{\kappa-1}\,|U_c^*[\phi,\psi](t,x)|\\ &\qquad\le C\sup_{y\in\Lambda_c'(t,x)}\langle|y|\rangle^{\kappa}\,(\langle|y|\rangle\,|\phi(y)|_1+|y|\,|\psi(y)|)\end{aligned}$$

for $(t,x)\in[0,\infty)\times\mathbb{R}^3$.

For the proof, see Katayama and Yokoyama [13, Lemma 3.1] (see also Asakura [2] and Kubota and Yokoyama [19]).

After the pioneering work of John [7], a wide variety of weighted L^∞-L^∞ estimates for $U_c[G]$ and $\partial U_c[G]$ have been obtained (see [2, 9, 10, 12, 13, 17, 18, 19, 25]). Here we restrict our attention to what will be used directly in our proofs of Theorems 1.1 and 2.1.

LEMMA 3.2. *Let* $c>0$. *Define*

$$\Phi_\rho(t,r)=\begin{cases}\log\big(2+\langle t+r\rangle\,\langle t-r\rangle^{-1}\big) & \text{if } \rho=0,\\ \langle t-r\rangle^{-\rho} & \text{if } \rho>0,\end{cases} \tag{3.2}$$

$$\Psi_\rho(t)=\begin{cases}\log(2+t) & \text{if } \rho=0,\\ 1 & \text{if } \rho>0.\end{cases} \tag{3.3}$$

Assume $\kappa\ge 1$ *and* $\mu>0$. *Then we have*

$$\langle t+|x|\rangle\,\Phi_{\kappa-1}(ct,|x|)^{-1}|U_c[G](t,x)|\le CA_{\kappa,\mu,0}[G;c](t,x), \tag{3.4}$$

$$\langle |x|\rangle\,\langle ct-|x|\rangle^{\kappa}\,\Psi_{\kappa-1}(t)^{-1}|\partial U_c[G](t,x)|\le CA_{\kappa,\mu,1}[G;c](t,x) \tag{3.5}$$

for $(t,x)\in[0,\infty)\times\mathbb{R}^3$, *where* $A_{\kappa,\mu,s}[G;c]$ *is given by* (2.3).

Proof. For the proof of (3.4), see Katayama and Yokoyama [13, equation (3.6) in Lemma 3.2, and section 8] for $\kappa>1$ and Katayama [11] for $\kappa=1$.

Next we consider (3.5) with $\kappa>1$. From Lemma 8.2 in [13], we find that (3.5) with $\partial U_c[G]$ replaced by $U_c[\partial G]$ is true. Now (3.5) follows immediately from Lemma 3.1, because we have $\partial_a U_c[G]=U_c[\partial_a G]+\delta_{a0}U_c^*[0,G(0,\cdot)]$ for $0\le a\le 3$ with the Kronecker delta δ_{ab}, and $\langle |y|\rangle^{\kappa+1}\,|y|\,|G(0,y)|\le CA_{\kappa,\mu,1}[G;c](t)$ (note that we have $w(0,r)=\langle r\rangle$). Equation (3.5) for the case $\kappa=1$ can be treated similarly (see [19] and [9]). $\square$

Note that we will use (3.5) in the proof of Theorem 1.1 but not in that of Theorem 2.1.

Now we are in a position to prove Theorem 2.1. Suppose that all the assumptions in Theorem 2.1 are fulfilled. Without loss of generality, we may assume $c=1$.

For simplicity of exposition, we write $D_\pm$ for $D_{\pm,1}=\partial_t\pm\partial_r$. Similarly, $U^*[\phi,\psi]$, $U[G]$, $A_{\rho,\mu,s}(t,x)$, and $B_{\rho,s}(t,x)$ denote $U_1^*[\phi,\psi]$, $U_1[G]$, $A_{\rho,\mu,s}[G;1](t,x)$, and $B_{\rho,s}[\phi,\psi;1](t,x)$, respectively.

First we prove (2.5). Assume $0<r=|x|\le 1$. We have

$$|D_+u|\le|\partial_t u|+|\nabla_x u|\le\sum_{0\le a\le 3}|U[\partial_a G]|+|U^*[0,G(0,\cdot)]|\,.$$

From (3.4) in Lemma 3.2, we get

$$\langle t+r\rangle\,\Phi_{\kappa-1}(t,r)^{-1}\,|U[\partial_a G](t,x)|\le CA_{\kappa,\mu,1}(t,x), \tag{3.6}$$

while Lemma 3.1 leads to

$$\begin{aligned}\langle t+r\rangle\,\langle t-r\rangle^{\kappa}\,|U^*[0,G(0,\cdot)](t,x)|&\le C\sup_{y\in\Lambda_1'(t,x)}|y|\,\langle |y|\rangle^{\kappa+1}\,|G(0,y)|\\&\le CA_{\kappa,\mu,0}(t,x).\end{aligned}$$

Thus we obtain (2.5) for $0<|x|\le 1$.

We set $v(t,r,\omega) = ru(t,r\omega)$ for $r > 0$ and $\omega \in S^2$. Then we have

$$D_-D_+v(t,r,\omega) = rG(t,r\omega) + \frac{1}{r}\sum_{1\le j<k\le 3}\Omega_{jk}^2 u(t,r\omega). \tag{3.7}$$

Let $r = |x| \ge 1$ and $1 \le \kappa \le 2$. From (3.4), we get

$$\begin{aligned}\frac{1}{r}\sum_{1\le j<k\le 3}|\Omega_{jk}^2 u(t,r\omega)| &\le C\langle r\rangle^{-1}\langle t+r\rangle^{-1}\Phi_{\kappa-1}(t,r)A_{\kappa,\mu,2}(t,r\omega)\\ &\le C\langle t+r\rangle^{-\kappa}(\langle r\rangle^{-1}+\langle t-r\rangle^{-1})A_{\kappa,\mu,2}(t,r\omega),\end{aligned} \tag{3.8}$$

where $\Phi_{\kappa-1}$ is from (3.2). It is easy to see that

$$|rG(t,r\omega)| \le \langle t+r\rangle^{-\kappa} w(t,r)^{-1-\mu}A_{\kappa,\mu,0}(t,r\omega). \tag{3.9}$$

Note that we have

$$A_{\kappa,\mu,s}(\tau,(t+r-\tau)\omega) \le A_{\kappa,\mu,s}(t,r\omega) \quad \text{for } 0 \le \tau \le t.$$

Therefore, by (3.7), (3.8), and (3.9), we get

$$\begin{aligned}|D_+v(t,r,\omega)| &= \left|\int_0^t \frac{d}{d\tau}(D_+v)(\tau,t+r-\tau,\omega)d\tau\right|\\ &= \left|\int_0^t (D_-D_+v)(\tau,t+r-\tau,\omega)d\tau\right|\\ &\le C\langle t+r\rangle^{-\kappa}A_{\kappa,\mu,2}(t,r\omega)\int_0^t\langle t+r-\tau\rangle^{-1}d\tau\\ &\quad + C\langle t+r\rangle^{-\kappa}A_{\kappa,\mu,2}(t,r\omega)\int_0^t\langle t+r-2\tau\rangle^{-1}d\tau\\ &\quad + C\langle t+r\rangle^{-\kappa}A_{\kappa,\mu,0}(t,r\omega)\int_0^t w(\tau,t+r-\tau)^{-1-\mu}d\tau\\ &\le C\langle t+r\rangle^{-\kappa}A_{\kappa,\mu,2}(t,r\omega)\log(2+t+r).\end{aligned} \tag{3.10}$$

Since we have

$$rD_+u(t,r\omega) = D_+v(t,r,\omega) - u(t,r\omega),$$

from (3.10) and (3.4), we obtain

$$\langle r\rangle\langle t+r\rangle\langle t-r\rangle^{\kappa-1}|D_+u(t,x)| \le C\log(2+t+|x|)A_{\kappa,\mu,2}(t,x)$$

for $r = |x| \ge 1$. This completes the proof of (2.5).

To prove (2.6), we first note that $\langle t+r\rangle \le C\langle r\rangle$ for $r > \delta t$. Let $1 < \kappa < 2$. By the first line of (3.8), we have

$$\frac{1}{r}\sum_{1\le j<k\le 3}|\Omega_{jk}^2 u(t,r\omega)| \le C\langle t+r\rangle^{-2}\langle t-r\rangle^{-\kappa+1}A_{\kappa,\mu,2}(t,r\omega) \tag{3.11}$$

for $r > \max\{\delta t, 1\}$. Obviously $r > \max\{\delta t, 1\}$ yields $t + r - \tau > \max\{\delta\tau, 1\}$ for $0 \le \tau \le t$. Hence following similar lines to (3.10), we obtain

$$|D_+ v(t, r, \omega)| \le C \langle t + r\rangle^{-\kappa} A_{\kappa,\mu,2}(t, r\omega) \quad \text{ for } r \ge \max\{\delta t, 1\}.$$

This immediately implies (2.6), because we already know that $|D_+u|$ (resp., $|D_+u - r^{-1}D_+v|$) has the desired bound for $(\delta t <)r \le 1$ (resp., $r \ge \max\{\delta t, 1\}$).

Now we are going to prove (2.7). Lemma 3.1 immediately implies

$$\langle t + |x|\rangle \langle t - |x|\rangle^{\kappa+\mu-1} |D_+u^*(t, x)| \le CB_{\kappa+\mu+1,1}(t, x),$$

which is better than (2.7) for $0 < |x| \le 1$. Lemma 3.1 also implies

$$\begin{aligned} \frac{1}{r} \sum_{1\le j<k\le 3} &|\Omega^2_{jk} u^*(t, x)| \\ &\le C \langle r\rangle^{-1} \langle t + r\rangle^{-1} \langle t - r\rangle^{1-\kappa-\mu} B_{\kappa+\mu+1,2}(t, x) \\ &\le C \langle t + r\rangle^{-\kappa} (\langle r\rangle^{-1-\mu} + \langle t - r\rangle^{-1-\mu})B_{\kappa+\mu+1,2}(t, x) \end{aligned} \tag{3.12}$$

for $r = |x| \ge 1$. Set $v^*(t, r, \omega) = ru^*(t, r\omega)$ for $r \ge 0$ and $\omega \in S^2$. For $r \ge 1$, similarly to (3.10), we get

$$\begin{aligned} |D_+v^*(t, r, \omega)| &= \left|(D_+v^*)(0, t + r, \omega) + \int_0^t (D_-D_+v^*)(\tau, t + r - \tau, \omega)d\tau\right| \\ &\le C \langle t + r\rangle^{-\kappa} B_{\kappa+1,0}(t, r\omega) \\ &\quad + C \langle t + r\rangle^{-\kappa} B_{\kappa+\mu+1,2}(t, r\omega) \int_0^t \langle t + r - \tau\rangle^{-1-\mu} d\tau \\ &\quad + C \langle t + r\rangle^{-\kappa} B_{\kappa+\mu+1,2}(t, r\omega) \int_0^t \langle t + r - 2\tau\rangle^{-1-\mu} d\tau \\ &\le C \langle t + r\rangle^{-\kappa} B_{\kappa+\mu+1,2}(t, r\omega), \end{aligned}$$

which ends up with

$$\langle r\rangle \langle t + r\rangle \langle t - r\rangle^{\kappa-1} |D_+u^*(t, x)| \le CB_{\kappa+\mu+1,2}(t, x)$$

for $r = |x| \ge 1$. This completes the proof of (2.7). ∎

4. Proof of Theorem 1.1. As an application of Theorem 2.1, we give a new proof of Theorem 1.1. First we derive estimates for the null forms.

LEMMA 4.1. *Let c be a positive constant, and $v = (v_1, \dots, v_M)$. Suppose that Q is one of the null forms. Then, for a nonnegative integer s, there exists a positive constant C_s, depending only on c and s, such that*

$$\begin{aligned} |Q(v_j, v_k)|_s \le C_s\Bigg\{&|\partial v|_{[s/2]} \sum_{|\alpha|\le s} |D_{+,c}Z^\alpha v| + |\partial v|_s \sum_{|\alpha|\le[s/2]} |D_{+,c}Z^\alpha v| \\ &+ \frac{1}{r}\big(|\partial v|_{[s/2]}|v|_{s+1} + |v|_{[s/2]+1}|\partial v|_s\big)\Bigg\}. \end{aligned}$$

Proof. The case $Q = Q_0$ and $s = 0$ follows immediately from (1.8). We can obtain similar identities for other null forms by using

$$(\partial_t, \nabla_x) = \left(\frac{1}{2}, -\frac{x}{2cr}\right) D_{-,c} + \left(\frac{1}{2}, \frac{x}{2cr}\right) D_{+,c} - \left(0, \frac{x}{r^2} \wedge \Omega\right)$$

with $\Omega = (\Omega_{23}, -\Omega_{13}, \Omega_{12})$ (see (5.2) in Sideris and Tu [23, Lemma 5.1]), and we can show the desired estimate for $s = 0$. Since $Z^\alpha Q(v_j, v_k)$ can be written in terms of $Q_0(Z^\beta v_j, Z^\gamma v_k; c)$ and $Q_{ab}(Z^\beta v_j, Z^\gamma v_k)$ $(0 \le a < b \le 3)$ with $|\beta| + |\gamma| \le |\alpha|$, the desired estimate for general s follows immediately. $\square$

Now we are going to prove Theorem 1.1. Without loss of generality, we may assume $c = 1$. Assume that the assumptions in Theorem 1.1 are fulfilled. Let u be the solution to (1.3)–(1.4) on $[0, T) \times \mathbb{R}^3$, and we set

$$\begin{aligned} e_{\rho,k}(t,x) = &\langle t + |x|\rangle \langle t - |x|\rangle^\rho |u(t,x)|_{k+2} + \langle |x|\rangle \langle t - |x|\rangle^{\rho+1} |\partial u(t,x)|_{k+1} \\ &+ \chi(t,x) \langle t + |x|\rangle^2 \langle t - |x|\rangle^\rho \sum_{|\alpha| \le k} |D_{+,1} Z^\alpha u(t,x)| \end{aligned}$$

for $\rho > 0$ and a positive integer k, where $\chi(t,x) = 1$ if $|x| > (1+t)/2$, while $\chi(t,x) = 0$ if $|x| \le (1+t)/2$. We fix $\rho \in (1/2, 1)$ and $s \ge 8$, and assume that

$$\sup_{0 \le t < T} \|e_{\rho,s}(t,\cdot)\|_{L^\infty(\mathbb{R}^3)} \le M\varepsilon \tag{4.1}$$

holds for some large $M (> 0)$ and small $\varepsilon (> 0)$, satisfying $M\varepsilon \le 1$. Our goal here is to get (4.1) with M replaced by $M/2$. Once such an estimate is established, it is well known that we can obtain Theorem 1.1 by the so-called bootstrap (or continuity) argument.

In the following we always assume M is large enough, and ε is sufficiently small. For simplicity of exposition, we will not write dependence of nonlinearities on the unknowns explicitly. Namely we abbreviate $F(u, \partial u, \nabla_x \partial u)(t,x)$ as $F(t,x)$, and so on.

First we evaluate the energy. For any nonnegative integer $k \le 2s$, (4.1) implies

$$|F^{(2)}(t,x)|_k \le CM\varepsilon \langle |x|\rangle^{-1} \langle t - |x|\rangle^{-1-\rho} |\partial u(t,x)|_{k+1}, \tag{4.2}$$

where $F^{(2)}$ denotes the quadratic terms of F. Put $H = F - F^{(2)}$, and $Zu = (Z_1 u, \dots, Z_7 u)$. Since we have

$$\langle r\rangle^{-1} \langle t - r\rangle^{-1} \le C \langle t + r\rangle^{-1} \quad \text{for any } (t,r) \in [0,\infty) \times [0,\infty), \tag{4.3}$$

and since $\langle |x|\rangle^{-1} |Zu| \le C|\partial u|$, from (4.1) we obtain

$$\begin{aligned} |H(t,x)|_k &\le C\big(|u|^3 + |(u, \partial u)|^2_{[k/2]+1} (|Zu|_{k-1} + |\partial u|_{k+1})\big) \\ &\le CM^3\varepsilon^3 \langle t + |x|\rangle^{-3} \langle t - |x|\rangle^{-3\rho} \\ &\quad + CM^2\varepsilon^2 \langle t + |x|\rangle^{-1} \langle t - |x|\rangle^{-2\rho} |\partial u(t,x)|_{k+1} \end{aligned} \tag{4.4}$$

for any nonnegative integer $k \le 2s$. Similarly to (4.2) and (4.4), using (4.3), we obtain

$$|F_{i,\alpha}(t,x)| \le CM\varepsilon(1+t)^{-1} |\partial u(t,x)|_{2s} + CM^3\varepsilon^3 \langle t + |x|\rangle^{-3} \langle t - |x|\rangle^{-3\rho} \tag{4.5}$$

for $|\alpha| \le 2s$, where

$$F_{i,\alpha} = Z^\alpha F_i - \sum_{j,k,a} c^{ij}_{ka} \partial_k \partial_a (Z^\alpha u_j)$$

with c^{ij}_{ka} coming from (1.5). It is easy to see that

$$\| \langle t + |\cdot| \rangle^{-3} \langle t - |\cdot| \rangle^{-3\rho} \|_{L^2(\mathbb{R}^3)} \le C(1+t)^{-2} \tag{4.6}$$

for $\rho > 1/2$. Therefore, from (4.5), we obtain

$$\|F_{i,\alpha}(t,\cdot)\|_{L^2} \le CM\varepsilon(1+t)^{-1}\|\partial u(t,\cdot)\|_{2s} + CM^3\varepsilon^3(1+t)^{-2}$$

for $|\alpha| \le 2s$. We also have

$$\sum_{j,k,a} |c^{ij}_{ka}(t,x)|_1 \le CM\varepsilon(1+t)^{-1}.$$

Now, applying the energy inequality for the systems of perturbed wave equations $\Box_1(Z^\alpha u_i) - \sum_{j,k,a} c^{ij}_{ka}\partial_k\partial_a(Z^\alpha u_j) = F_{i,\alpha}$, we find

$$\frac{d}{dt}\|\partial u(t,\cdot)\|_{2s} \le CM\varepsilon(1+t)^{-1}\|\partial u(t,\cdot)\|_{2s} + CM^3\varepsilon^3(1+t)^{-2},$$

and the Gronwall lemma leads to

$$\|\partial u(t,\cdot)\|_{2s} \le C(\varepsilon + M^3\varepsilon^3)(1+t)^{C_0 M\varepsilon} \le CM\varepsilon(1+t)^{C_0 M\varepsilon} \tag{4.7}$$

with an appropriate positive constant C_0 which is independent of M (note that the energy inequality for the systems of perturbed wave equations is available because of the symmetry condition).

In the following, we repeatedly use Theorem 2.1 and Lemmas 3.1 and 3.2 with the choice of $N = 1$ and $c_1 = 1 (= c)$. In other words, from now on we put $w(t,r) = \min\big\{\langle r \rangle, \langle t - r \rangle\big\}$. Note that we have

$$\langle r \rangle^{-1} \langle t - r \rangle^{-1} \le C \langle t + r \rangle^{-1} w(t,r)^{-1}, \tag{4.8}$$

which is more precise than (4.3).

By (4.7) and the Sobolev-type inequality

$$\langle |x| \rangle \, |v(t,x)| \le C\|v(t,\cdot)\|_2,$$

whose proof can be found in Klainerman [15], we see that

$$\langle |x| \rangle \, |\partial u(t,x)|_{2s-2} \le CM\varepsilon(1+t)^{C_0 M\varepsilon}. \tag{4.9}$$

Using (4.8) and (4.9), from (4.2) and (4.4) with $k = 2s-3$, we obtain

$$|F(t,x)|_{2s-3} \le CM^2\varepsilon^2 \langle r \rangle^{-1} \langle t + |x| \rangle^{-1} w(t,|x|)^{-2\rho}(1+t)^{C_0 M\varepsilon},$$

which implies

$$A_{1+\nu,2\rho-1,2s-3}[F;1](t,x) \le CM^2\varepsilon^2 \langle t + |x| \rangle^{C_0 M\varepsilon + \nu}, \tag{4.10}$$

where ν is a positive constant to be fixed later (note that we have $\langle \tau+|y|\rangle \le \langle t+|x|\rangle$ for $(\tau,y)\in\Lambda_1(t,x)$). Since $2\rho>1$ and $1+\nu>1$, by Lemmas 3.1 and 3.2 with Theorem 2.1, we obtain

$$\begin{aligned} (4.11)\qquad e_{0,2s-5}(t,x) &\le e_{\nu,2s-5}(t,x) \le C\varepsilon + CM^2\varepsilon^2 \langle t+|x|\rangle^{C_0M\varepsilon+\nu} \\ &\le CM\varepsilon \langle t+|x|\rangle^{C_0M\varepsilon+\nu}. \end{aligned}$$

Finally, we are going to estimate $e_{\rho,s}(t,x)$. By (4.11) and (4.2) with $k=2s-6$, we have

$$|F^{(2)}(t,x)|_{2s-6} \le CM^2\varepsilon^2 \langle t+|x|\rangle^{-2-\rho+C_0M\varepsilon+\nu} \langle |x|\rangle^{-2}$$

for (t,x) satisfying $|x|\le (t+1)/2$. On the other hand, (4.1), (4.11), and Lemma 4.1 imply

$$|F^{(2)}(t,x)|_{2s-6} \le CM^2\varepsilon^2 \langle t+|x|\rangle^{-3+C_0M\varepsilon+\nu} \langle t-|x|\rangle^{-1-\rho}$$

for (t,x) satisfying $|x|\ge (t+1)/2$. Summing up, we obtain

$$(4.12)\qquad |F^{(2)}(t,x)|_{2s-6} \le CM^2\varepsilon^2 \langle |x|\rangle^{-1} \langle t+|x|\rangle^{-2+C_0M\varepsilon+\nu} w(t,|x|)^{-1-\rho}.$$

By the first line of (4.4) with $k=2s-6$, using (4.1) and (4.11), we get

$$(4.13)\qquad |H(t,x)|_{2s-6} \le CM^3\varepsilon^3 \langle |x|\rangle^{-1} \langle t+|x|\rangle^{-2+C_0M\varepsilon+\nu} w(t,|x|)^{-2\rho}.$$

Equations (4.12) and (4.13) yield

$$(4.14)\qquad |F(t,x)|_{2s-6} \le CM^2\varepsilon^2 \langle |x|\rangle^{-1} \langle t+|x|\rangle^{-2+C_0M\varepsilon+\nu} w(t,|x|)^{-2\rho}.$$

Now we fix some ν satisfying $0<\nu<1-\rho$, and assume that ε is sufficiently small to satisfy $-2+C_0M\varepsilon+\nu\le -1-\rho$. Then from (4.14) we find that

$$(4.15)\qquad A_{1+\rho,2\rho-1,2s-6}[F;1](t,x) \le CM^2\varepsilon^2.$$

Since we have $s+2\le 2s-6$, $1+\rho>1$, and $2\rho>1$, from Theorem 2.1, Lemmas 3.1 and 3.2, we obtain

$$(4.16)\qquad e_{\rho,s}(t,x) \le C_1\big(\varepsilon + M^2\varepsilon^2\big)$$

for $(t,x)\in[0,T)\times\mathbb{R}^3$, with an appropriate positive constant C_1 which is independent of M. Finally, if M is large enough to satisfy $4C_1\le M$, and ε is small enough to satisfy $C_1M\varepsilon\le 1/4$, by (4.16) we obtain

$$(4.17)\qquad \sup_{0\le t<T} \|e_{\rho,s}(t,\cdot)\|_{L^\infty(\mathbb{R}^3)} \le \frac{M}{2}\varepsilon,$$

which is the desired result. This completes the proof. $\square$

REFERENCES

[1] S. Alinhac, *Remarks on energy inequalities for wave and Maxwell equations on a curved back ground*, Math. Ann., 329 (2004), pp. 707–722.

[2] F. Asakura, *Existence of a global solution to a semi-linear wave equation with slowly decreasing initial data in three space dimensions*, Comm. Partial Differential Equations, 11 (1986), pp. 1459–1487.
[3] D. Christodoulou, *Global solutions of nonlinear hyperbolic equations for small initial data*, Comm. Pure Appl. Math., 39 (1986), pp. 267–282.
[4] K. Hidano, *The global existence theorem for quasi-linear wave equations with multiple speeds*, Hokkaido Math. J., 33 (2004), pp. 607–636.
[5] L. Hörmander, *L^1, L^∞ estimates for the wave operator*, in Analyse Mathématique et Applications, Contributions en l'Honneur de J. L. Lions, Gauthier–Villars, Paris, 1988, pp. 211–234.
[6] A. Hoshiga and H. Kubo, *Global small amplitude solutions of nonlinear hyperbolic systems with a critical exponent under the null condition*, SIAM J. Math. Anal., 31 (2000), pp. 486–513.
[7] F. John, *Blow-up of solutions of nonlinear wave equations in three space dimensions*, Manuscripta Math., 28 (1979), pp. 235–268.
[8] F. John, *Lower bounds for the life span of solutions of nonlinear wave equations in three space dimensions*, Comm. Pure Appl. Math., 36 (1983), pp. 1–35.
[9] S. Katayama, *Global and almost-global existence for systems of nonlinear wave equations with different propagation speeds*, Differential Integral Equations, 17 (2004), pp. 1043–1078.
[10] S. Katayama, *Global existence for systems of wave equations with nonresonant nonlinearities and null forms*, J. Differential Equations, 209 (2005), pp. 140–171.
[11] S. Katayama, *Lifespan for radially symmetric solutions to systems of semilinear wave equations with multiple speeds*, Osaka J. Math, to appear.
[12] S. Katayama and A. Matsumura, *Sharp lower bound for the lifespan of the systems of semilinear wave equations with multiple speeds*, J. Math. Kyoto Univ., 45 (2005), pp. 391–403.
[13] S. Katayama and K. Yokoyama, *Global small amplitude solutions to systems of nonlinear wave equations with multiple speeds*, Osaka J. Math., 43 (2006), pp. 283–326.
[14] S. Klainerman, *The null condition and global existence to nonlinear wave equations*, in Nonlinear Systems of Partial Differential Equations in Applied Mathematics, Part 1, Lectures in Appl. Math. 23, AMS, Providence, RI, 1986, pp. 293–326.
[15] S. Klainerman, *Remarks on the global Sobolev inequalities in the Minkowski space* $\mathbf{R}^{n+1}$, Comm. Pure Appl. Math., 40 (1987), pp. 111–117.
[16] S. Klainerman and T. C. Sideris, *On almost global existence for nonrelativistic wave equations in 3D*, Comm. Pure Appl. Math., 49 (1996), pp. 307–321.
[17] M. Kovalyov, *Long-time behavior of solutions of a system of nonlinear wave equations*, Comm. Partial Differential Equations, 12 (1987), pp. 471–501.
[18] M. Kovalyov, *Resonance-type behaviour in a system of nonlinear wave equations*, J. Differential Equations, 77 (1989), pp. 73–83.
[19] K. Kubota and K. Yokoyama, *Global existence of classical solutions to systems of nonlinear wave equations with different speeds of propagation*, Japan. J. Math. (N.S.), 27 (2001), pp. 113–202.
[20] J. Metcalfe, M. Nakamura, and C. D. Sogge, *Global existence of quasilinear, nonrelativistic wave equations satisfying the null condition*, Japan. J. Math. (N.S.), 31 (2005), pp. 391–472.
[21] J. Metcalfe and C. D. Sogge, *Global existence of null-form wave equations in exterior domains*, Math. Z., 256 (2007), pp. 521–549.
[22] T. C. Sideris and B. Thomases, *Local energy decay for solutions of multidimensional isotropic symmetric hyperbolic systems*, J. Hyperbolic Differ. Equ., 3 (2006), pp. 673–690.
[23] T. C. Sideris and S.-Y. Tu, *Global existence for systems of nonlinear wave equations in 3D with multiple speeds*, SIAM J. Math. Anal., 33 (2001), pp. 477–488.
[24] C. D. Sogge, *Global existence for nonlinear wave equations with multiple speeds*, in Harmonic Analysis at Mount Holyoke, W. Beckner, A. Nagel, A. Seeger, and H. F. Smith, eds., Contemp. Math. 320, AMS, Providence, RI, 2003, pp. 353–366.
[25] K. Yokoyama, *Global existence of classical solutions to systems of wave equations with critical nonlinearity in three space dimensions*, J. Math. Soc. Japan, 52 (2000), pp. 609–632.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 1863–1889

AN INVERSE PROBLEM FOR A NONLINEAR PARABOLIC EQUATION WITH APPLICATIONS IN POPULATION DYNAMICS AND MAGNETICS*

BARBARA KALTENBACHER† AND MICHAEL V. KLIBANOV‡

Abstract. The parabolic equation of this paper is a nonlinear one with the unknown coefficient depending on the derivative of the solution. A uniqueness result is proven by the method of Carleman estimates. The applicability of this result is illustrated for parameter identification problems in population dynamics and magnetics. For the latter application, we provide numerical results using a reconstruction method based on a multiharmonic formulation of the problem.

Key words. parameter identification, uniqueness, Carleman estimates

AMS subject classifications. 35R30, 65M32, 82D40, 92D25

DOI. 10.1137/070685907

1. Introduction.

Nonlinear parabolic PDEs arise in a large variety of applications ranging from combustion theory via environmental pollution, population dynamics, and nonlinear magnetics to the theory of the economic growth. Since the coefficient depending nonlinearly on the solution of the PDE is often not accessible to direct measurements, its determination from boundary measurements is an important task.

In this paper we consider the question of identifiability, i.e., whether this coefficient function can be uniquely determined from the given data. While there exist results on the situation that the coefficient depends on values of the solution (as it is relevant, e.g., in nonlinear heat conduction), here we concentrate on the problem when this coefficient is a function of the derivative of the solution (as typical for nonlinear magnetics; cf., e.g., [2]) or where the equation is in nondivergence form (e.g., in diffusion models for population dynamics; cf. [1]).

To prove our uniqueness theorem, we apply the method of Carleman estimates; see, e.g., [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [18], [19], [20], [21], [22], [23], [24] for this method. The majority of works on this method is concerned with linear equations in which the unknown coefficients depend on spatial variables. Nonlinear parabolic equations were treated by this technique in [8], [18], [20], and Chapter 4 of [21]. Our inverse problem is a problem with the data resulting from a single measurement event. In the case of multiple measurements, uniqueness theorems for inverse problems for nonlinear parabolic equations with unknown coefficients depending on solutions and their first derivatives were proven in [13]. The case of single measurement data for such equations was considered in, e.g., [8], [18], [20], [21], and [24]. However, in these works the unknown coefficient depends on the solution of the original parabolic equation (as well as on some spatial variables in the multidimensional case [20], [21]). The case of the dependence on derivatives of the solution was not considered.

*Received by the editors March 21, 2007; accepted for publication (in revised form) October 31, 2007; published electronically February 22, 2008.
http://www.siam.org/journals/sima/39-6/68590.html
†Department of Mathematics, University of Stuttgart, 70569 Stuttgart, Germany (Barbara.Kaltenbacher@mathematik.uni-stuttgart.de).
‡Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC 28223 (mklibanv@uncc.edu). This author's research was supported by, or in part by, the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant W911NF-05-1-0378.

Another significantly new element here is that the unknown coefficient $k\,(u_x)$ is involved together with its first derivative. This leads to the necessity to prove some new estimates for Volterra-like integrals in which Carleman weight functions (CWFs) are involved; see Lemmata 3 and 4. In the past such estimates were established only for t-dependent integrals [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [18], [19], [20], [21], [22], [23], [24]. Now, however, we also consider the z-dependence (Lemma 3). In addition, such estimates for iterated integrals (Lemma 4) were not previously established.

In the next five paragraphs we briefly outline the idea of the proof of the uniqueness theorem. First, we differentiate the original parabolic PDE with respect to x to obtain a new PDE with respect to the function $v(x,t) = u_x(x,t)$, the x-derivative of the original solution $u(x,t)$. Next, in order to be in a position of applicability of the above-mentioned method of Carleman estimates, we "transform" the function $v(x,t)$ into the spatial variable z using the change of variables (4.6), which is $v\,(w(z,t),t) = z$. Naturally, by the implicit function theorem we need the condition $v_x(x,t) = u_{xx}(x,t) \neq 0$. Hence, for the sake of definiteness, we impose $u_{xx}\,(x,t) \geq \alpha = const. > 0$; see (3.3). Next, we obtain a linear ordinary differential equation (ODE) of second order with respect to the function $\widetilde{k}(z)$, where $\widetilde{k}(z) = k_1(z) - k_2(z)$ is the difference of two possible target coefficients, and we need to prove that this difference equals zero.

Note that the idea of considering ODEs with respect to the unknown coefficients is new, because such ODEs were not considered in the past works on the method of Carleman estimates for inverse problems. Because of the presence of the factor z at $\widetilde{k}''(z)$, we need to impose condition (3.6) in order to avoid the singularities in that equation. We represent the solution of that ODE via Volterra-like integrals. Since the function $k(z)$ is independent of t, we follow one of the ideas of the original work [4] (also, see [19], [21]) differentiating with respect to t and, thus, "eliminating" the function $\widetilde{k}(z)$ from the resulting equation. Next, using the differentiation with respect to z, we obtain an integral differential equation with respect to the function $p\,(z,t) = (w_{1t} - w_{2t})\,(z,t)$, where functions w_1 and w_2 correspond to the coefficients k_1 and k_2, respectively.

However, two major difficulties of the latter equation compared with the above listed previous works on the method of Carleman estimates for inverse problems are the following: (a) it contains not only t-dependent Volterra-like integrals (like the ones in the past) but also z-dependent integrals as well as "mixed" (z,t)-dependent Volterra-like integrals; and (b) second order derivatives p_{zz} are involved as parts of integrands. The element (a) arises because of the solution of the above ODE, thus ultimately because the unknown coefficient is involved in the original PDE together with its derivative. The element (b) is due to the nonlinearity of the original PDE. To overome (a), we need to establish new estimates from the above for those "new" Volterra-like integrals in which CWFs are involved (Lemmata 3 and 4). Note that such estimates cannot be proven in the same way as Carleman estimates are proven for differential operators, because the structure of Volterra-like integral operators is different from the structure of differential operators.

It is not straightforward to deal with (b) because the term with the second z-derivative stands with a small factor $1/\lambda$ in the Carleman estimate (4.50) for the parabolic operator $\partial_t - h_1\partial_z^2$. To overcome (b), we modify one of the ideas of [19]. Namely, we rescale variables (z,t) via the change of variables (4.30), which causes some technical difficulties and rather long formulas, because we need to figure out carefully the appearance of a small factor γ. As a result, this rescaling leads to the appearance of γ in the resulting integral differential inequality (4.43). The latter, as well as Lemmata 2–4, enable us to come up with the "standard" inequality (4.54), after which the proof is similar to previous works.

However, the move from the rescaling (4.30) to (4.54) can be done only if one considers a small part of the rectangle R' in (4.31), rather than the whole rectangle "at once," thus "exhausting" the entire R' via a sequence of small δ-steps. Note that it is unclear at this point whether the method of [11] for an inverse parabolic problem can be adapted to our case to obtain a Lipschitz stability estimate, since it is unclear whether it is possible to obtain analogues of Lemmata 3 and 4 if using the CWF of [12]. It is also unclear whether the method of [12] for hyperbolic equations can be adapted in this case, because of the necessity to exhaust the entire domain via a sequence of small rectangles. Another important difference with the hyperbolic case of [12] is that the Carleman estimate for a hyperbolic operator does not include derivatives involved in the principal part of that operator, unlike the parabolic case (see Lemma 1 for the latter). However, we do need that "inclusion" because some integrals in (4.43) contain the second order derivative p_{zz}, and our estimate should incorporate it in the left-hand side.

The paper is organized as follows. In section 2, we state the inverse problem under consideration, and in section 3 we formulate our uniqueness result. The proof of this theorem is given in section 4. Section 5 describes two applications, namely in population dynamics and in nonlinear magnetics. For the latter, the final section (section 6) provides numerical results obtained with a method based on a multiharmonic formulation.

2. Statement of the problem. The forward problem is

$$u_t = \left(k\left(u_x\right)u_x\right)_x\,, \quad (x,t) \in (0,L)\times(0,T)\,, \tag{2.1}$$

$$u\left(x,0\right) = r(x), \quad x \in (0,L)\,, \tag{2.2}$$

$$u\left(0,t\right) = f_0\left(t\right), \quad u(L,t) = f_1\left(t\right), \quad t \in (0,T)\,, \tag{2.3}$$

or with $v(x,t) = u_x(x,t)$, i.e., $u(x,t) = \int_0^x v(\xi,t)\,d\xi + f_0(t)$,

$$v_t = \left(k\left(v\right)v\right)_{xx}\,, \quad (x,t) \in (0,L)\times(0,T)\,, \tag{2.4}$$

$$v\left(x,0\right) = r'(x), \quad x \in (0,L)\,, \tag{2.5}$$

$$(k(v)v)_x(0,t) = f_0'(t), \quad (k(v)v)_x(L,t) = f_1'(t), \quad t \in (0,T)\,, \tag{2.6}$$

which results from the previous setting by differentiation with respect to the space variable. Naturally, it is assumed that the function $k(z) \in C^1\left(\mathbb{R}\right)$ and

$$k\left(z\right) \geq k_0 = const. > 0 \quad \forall z \in \mathbb{R}. \tag{2.7}$$

It is well known that it is difficult to investigate the question of uniqueness of coefficient inverse problems for parabolic equations unless one assumes that the solution of the forward problem is known at $t = \varepsilon \in (0,T)$. In the latter case, however, one does not need to assume knowledge of the initial condition at $\{t = 0\}$; see, e.g., [5], [11], and sections 3.3.1 and 3.3.2 in [21]. In other words, one needs to assume that knowledge of the initial condition (2.2) is replaced with knowledge of the function $u(x,\varepsilon)$ for an arbitrary $\varepsilon \in (0,T)$; also see section 5 for a physical interpretation. Because of this, we formulate the inverse problem as follows.

Inverse problem. Assume that (2.1) is satisfied for $(x,t) \in (0,L) \times (-c,T)$, where $c \in (0,T]$ is an arbitrary number, and the function $u(x,0) = r(x)$ in (2.2) is known, and assume that boundary conditions $f_0(t)$ and $f_1(t)$ are satisfied for $t \in (-c,T)$. Also, assume that the following two functions $g_0(t)$ and $g_1(t)$ are given:

$$u_x(0,t) = g_0(t), \quad u_x(L,t) = g_1(t), \quad t \in (-c,T). \tag{2.8}$$

Let the interval (a,b) be the range of the function $u_x(x,t)$ for $(x,t) \in (0,L) \times (-c,T)$, i.e., $a \le u_x(x,t) \le b$. Determine the interval (a,b) as well as the unknown coefficient $k(z)$ for $z \in (a,b)$.

3. Uniqueness theorem. The following uniqueness theorem holds.

THEOREM 1. *Let the conditions of the statement of the inverse problem be satisfied. Suppose that there exist two functions $k_1(z)$ and $k_2(z)$ satisfying condition* (2.7) *and such that*

$$zk_i'(z) + k_i(z) \ge const. > 0 \quad \forall z \in \mathbb{R}, \quad i = 1,2. \tag{3.1}$$

Let $u_1(x,t)$ and $u_2(x,t)$ be two corresponding solutions of the forward problem (2.1)–(2.3) *satisfying the same Neumann boundary conditions* (2.8) *and such that*

$$\partial_t^s \partial_x^j u_i \in C([0,L] \times [-c,T]) \quad \text{for } s = 0,1; \quad j = 0,1,2,3; \quad i = 1,2. \tag{3.2}$$

Assume also that in $(0,L) \times (-c,T)$

$$\partial_x^2 u_i \ge \alpha = const. > 0, \quad i = 1,2. \tag{3.3}$$

In addition, let

$$g_0'(t) \ge 0, \quad t \in (-c,T), \tag{3.4}$$

$$g_1(t) \equiv const. = \widetilde{g}, \quad t \in (-c,T). \tag{3.5}$$

Also, assume that there exists a number $\varepsilon \in (0,T)$ such that

$$0 \notin [g_0(\varepsilon), \widetilde{g}] \tag{3.6}$$

and

$$k_1(z) = k_2(z), \quad z \in (g_0(-c), g_0(\varepsilon)). \tag{3.7}$$

Denote $a = g_0(-c), b = \widetilde{g}$. Then the range of both functions u_{1x} and u_{2x} coincides with the interval (a,b) and $k_1(z) = k_2(z)$ for $z \in (a,b)$. Furthermore, it is not necessary to know the boundary condition $f_1(t)$ in (2.3).

Observe that by (2.8), (3.3), and (3.5)

$$g_0(t) \le g_0(T) < \widetilde{g}, t \in [-c,T]. \tag{3.8}$$

Hence, it should be demonstrated in the proof of this theorem that the function $k(z)$ can be uniquely determined for $z \in [g_0(\varepsilon), \widetilde{g}]$. It follows from (3.3) and (3.4) that the range of the function $u_x(x,t)$ is the interval (a,b) indicated in Theorem 1.

Remark.

- As to the boundary condition $f_1(t)$ in (2.3), it is not actually used in the proof. Instead, only the boundary condition $g_1(t)$ is used at $x = L$.

- To explain the condition (3.5), we note that if it is not satisfied, then one should apply the above outlined method (see the introduction) to the domain with curvilinear boundaries $z = g_0(t)$, which is the left boundary, and $z = g_1(t)$, the right boundary. While we actually avoid working with the left boundary assuming (3.7) (see (4.25)–(4.27)), we cannot avoid the right boundary. In the latter case, however, we would need to assume that the function $k(g_1(t))$ is known. So, the condition (3.5) is imposed to avoid the curvilinear right boundary and thus not to use the latter assumption.
- Note that the smoothness condition (3.2) can be guaranteed via imposing certain assumptions on the coefficient $k(z)$, the unknown "initial" condition $u(x,-c)$, and boundary conditions (2.3) (although it is not necessary to actually know the function $u(x,-c)$).
- Conditions (3.3) might also be guaranteed via applying the maximum principle and imposing additional conditions on the functions f_0, f_1, g_0, g_1, but this is outside the scope of this publication.
- Note that (3.1), (3.4)–(3.7) follow sometimes from the physics of the problem; see, e.g., section 5 below.

4. Proof of Theorem 1.

4.1. An integral differential equation. We first obtain an integral differential equation which does not contain the difference $(k_1 - k_2)(z)$. Let $(k(z), u(x,t))$ be one of the pairs $(k_i(z), u_i(x,t))$, $i = 1, 2$. Rewrite (2.1) in the form

$$u_t = [k'(u_x)\, u_x + k(u_x)]\, u_{xx}. \tag{4.1}$$

Denote $v(x,t) = u_x(x,t)$. Differentiating (4.1) with respect to x, we obtain the following equation for the function v:

$$v_t = [k'(v)\, v + k(v)]\, v_{xx} + [k''(v)\, v + 2k'(v)]\, v_x^2, (x,t) \in (0,L) \times (-c,T). \tag{4.2}$$

By (3.1), $k'(v)\, v + k(v) \geq \mathit{const.} > 0$, which implies that (4.1) is of the parabolic type. In addition, (4.1), (2.2), (2.3), (2.8), and (3.5) imply that

$$v(x,0) = r'(x), \tag{4.3}$$

$$v(0,t) = g_0(t), \quad v(L,t) = \widetilde{g} = \mathit{const.}, \tag{4.4}$$

$$v_x(0,t) = \frac{f_0'(t)}{k(g_0(t)) + k'(g_0(t))\, g_0(t)} := s_0(t), \tag{4.5}$$

$$v_x(L,t) = \frac{f_1'(t)}{k(\widetilde{g}) + k'(\widetilde{g})\, \widetilde{g}} := s_1(t).$$

Note that by (2.8), (3.7), and (4.1) the function $s_0(t)$ is the same for both pairs (u_1, k_1) and (u_2, k_2) as long as $t \in (-c, \varepsilon)$. On the other hand, the function $s_1(t)$ would even be unknown if we would assume f_1 to be known, because the numbers $k(\widetilde{g}), k'(\widetilde{g})$ are unknown.

We now "turn" the function $v(x,t)$ into a spatial variable z. To do so, we notice that by (3.3), $v_x(x,t) \geq \alpha = \mathit{const.} > 0$. Hence, we can introduce a new function $w(z,t)$ using the implicit function theorem as

$$v(w(z,t), t) = z. \tag{4.6}$$

By (3.8), $g_0(t) < z < \widetilde{g}$. Hence, the function $w(z,t)$ is defined in the domain D with curvilinear boundaries,

$$D = \{(z,t) : g_0(-c) < g_0(t) < z < \widetilde{g}, t \in (-c,T)\}. \tag{4.7}$$

Using (4.6), express derivatives of the function v via derivatives of the function w,

$$u_{xx}(w(z,t),t) = v_x(w(z,t),t) = \frac{1}{w_z(z,t)}, \tag{4.8}$$

$$v_t(w(z,t),t) = -\frac{w_t(z,t)}{w_z(z,t)},$$

$$v_{xx}(w(z,t),t) = -\frac{w_{zz}(z,t)}{w_z^3(z,t)}.$$

Hence, conditions (4.2)–(4.5) become

$$w_t = \frac{[k'(z)z + k(z)]}{w_z^2} w_{zz} - \frac{k''(z)z + 2k'(z)}{w_z}, \quad (z,t) \in D, \tag{4.9}$$

$$w(z,0) = \widetilde{r}(z), \tag{4.10}$$

$$w(g_0(t),t) = 0, \tag{4.11}$$

$$w_z(g_0(t),t) = \frac{1}{s_0(t)}, \quad t \in (-c,\varepsilon), \tag{4.12}$$

$$w(\widetilde{g},t) = L. \tag{4.13}$$

Here $\widetilde{r}(z)$ is the function inverse to the function $r'(x)$, i.e., $r'(\widetilde{r}(z)) = z$. The interval $t \in (-c,\varepsilon)$ is emphasized in (4.12) because the function $s_0(t)$ is known only on this interval.

Denote $q(z,t) = w_1(z,t) - w_2(z,t), \widetilde{k}(z) = k_1(z) - k_2(z)$, where functions w_1 and w_2 correspond to functions u_1 and u_2, respectively. We need to prove that

$$q(z,t) = 0 \text{ in } D \tag{4.14}$$

and

$$\widetilde{k}(z) = 0 \text{ for } z \in (a,b). \tag{4.15}$$

Using the formula

$$a_1 b_1 - a_2 b_2 = (a_1 - a_2) b_1 + (b_1 - b_2) a_2 \quad \forall a_1, a_2, b_1, b_2 \in \mathbb{R},$$

we obtain from (4.9) the following equation for the function q:

$$\begin{aligned} q_t &= h_1(z,t) q_{zz} + h_2(z,t) q_z \\ &\quad - \left[z\widetilde{k}''(z) + \widetilde{k}'(z) \left(2 - \frac{z}{w_{2z}} w_{2zz} \right) - \frac{\widetilde{k}(z)}{w_{2z}} w_{2zz} \right] \frac{1}{w_{2z}} \quad \text{in } D, \end{aligned} \tag{4.16}$$

where

$$h_1(z,t) = \frac{k_1'(z)z + k_1(z)}{w_{1z}^2(z,t)},$$

$$h_2(z,t) = -\left(k_1'(z)z + k_1(z)\right)\frac{w_{1z}+w_{2z}}{w_{1z}^2 w_{2z}^2} w_{2zz} + \frac{k_1''(z)\,z + 2k_1'(z)}{w_{1z}w_{2z}}.$$

Hence, functions

$$h_1, h_2 \in C^1\left(\overline{D}\right) \tag{4.17}$$

and (see (3.1), (3.3))

$$\beta_1 \le h_1 \le \beta_2, \quad |h_{1z}| + |h_{1t}| \le \beta_3 \quad \text{in } \overline{D}, \tag{4.18}$$

where β_1, β_2, and β_3 are certain positive constants. In addition, conditions (4.10)–(4.13) imply that

$$q(z,0) = 0, \tag{4.19}$$

$$q(g_0(t), t) = 0, \quad t \in (-c, T), \tag{4.20}$$

$$q_z(g_0(t), t) = 0, \quad t \in (-c, \varepsilon), \tag{4.21}$$

$$q(\widetilde{g}, t) = 0. \tag{4.22}$$

By (3.4), $g_0(\varepsilon) \ge g_0(-c)$. Hence, by (3.7),

$$\widetilde{k}(z) = 0 \quad \text{for } z \in (g_0(-c), g_0(\varepsilon)). \tag{4.23}$$

Hence, we obtain from (4.16)

$$q_t = h_1(z,t)\, q_{zz} + h_2(z,t)\, q_z, \quad (z,t) \in \{g_0(t) < z < g_0(\varepsilon), -c < t < \varepsilon\}. \tag{4.24}$$

Hence, (4.20), (4.21), (4.24), and the well-known unique continuation theorem for parabolic equations with lateral Cauchy data (see, e.g., section 2.2.2 in [21] for this theorem) imply that

$$q(z,t) = 0, \quad (z,t) \in \{g_0(t) < z < g_0(\varepsilon), -c < t < \varepsilon\}.$$

Thus, we will now consider (4.16) in the rectangle R,

$$R = \{(z,t) : g_0(\varepsilon) < z < \widetilde{g},\ t \in (-\varepsilon, \varepsilon)\}.$$

Conditions (4.20), (4.21), and (4.23) are replaced with

$$q(g_0(\varepsilon), t) = 0, \quad t \in (-\varepsilon, \varepsilon), \tag{4.25}$$

$$q_z(g_0(\varepsilon), t) = 0, \quad t \in (-\varepsilon, \varepsilon), \tag{4.26}$$

$$\widetilde{k}(g_0(\varepsilon)) = \widetilde{k}'(g_0(\varepsilon)) = 0. \tag{4.27}$$

Since the range of the function $u_x(x,t)$ coincides with the interval (a,b), it is sufficient to derive from (4.16), (4.25)–(4.27) that

$$q(z,t) = 0 \quad \text{in } R \tag{4.28}$$

and

$$\widetilde{k}(z) = 0 \quad \text{for } z \in (g_0(\varepsilon), \widetilde{g}). \tag{4.29}$$

Thus, from now on we will consider (4.16) only in the rectangle R. Note that while two boundary conditions (4.25) and (4.26) are given at the left side of R, only one boundary condition (4.22) is given at its right side.

Since the first term in the right-hand side of the Carleman estimate (4.50) (subsection 4.2) stands with the small factor $1/\lambda$, and the estimate of Lemma 2 (subsection 4.3) also gives this factor, we can apply the Carleman estimate successfully only if a small factor stands at the Volterra t-integrals with the second z-derivative p_{zz}; see (4.43). For this reason we rescale variables as

$$(z, t) \leftrightarrow (\sqrt{\gamma} z^*, \gamma t^*), \tag{4.30}$$

where $\gamma \in (0, 1)$ is a small constant which we will choose later. The rectangle R becomes

$$R' = \left\{ (z^*, t^*) : \frac{g_0(\varepsilon)}{\sqrt{\gamma}} < z^* < \frac{\widetilde{g}}{\sqrt{\gamma}},\ t \in \left(-\frac{\varepsilon}{\gamma}, \frac{\varepsilon}{\gamma} \right) \right\}. \tag{4.31}$$

Then for any function $Y(z, t) \in C^1\left(\overline{R}\right)$

(4.32)

$$|\partial_{z^*} Y(\sqrt{\gamma} z^*, \gamma t^*)| \le \sqrt{\gamma} \|\partial_z Y(z, t)\|_{C(\overline{R})}, \quad |\partial_{t^*} Y(\sqrt{\gamma} z^*, \gamma t^*)| \le \gamma \|\partial_t Y(z, t)\|_{C(\overline{R})}.$$

Since we need to work carefully with the new variables, we do not keep the same notations as for the old variables for them. Also, because of (3.2) and (3.6), there exists a positive constant b such that either $g_0(\varepsilon) < \widetilde{g} \le -b$ or $b \le g_0(\varepsilon) < \widetilde{g}$. Hence, we will assume without loss of generality that $1 \le g_0(\varepsilon) < \widetilde{g}$. Therefore, by (4.32),

$$1 \le \sqrt{\gamma} z^* \le \widetilde{g} \quad \text{in } R'. \tag{4.33}$$

Denote $\widehat{k}(z^*) := \widetilde{k}(\sqrt{\gamma} z^*)$, $s(z^*, t^*) := q(\sqrt{\gamma} z^*, \gamma t^*)$, $\widehat{w}(z^*, t^*) := w(\sqrt{\gamma} z^*, \gamma t^*)$, and $\widehat{h}_i(z^*, t^*) := h_i(\sqrt{\gamma} z^*, \gamma t^*)$, $i = 1, 2$. Below, derivatives of $\widehat{k}(z^*)$ are understood as derivatives with respect to z^*.

Equation (4.16) now becomes

(4.34)

$$\begin{aligned}
& s_{t^*}(z^*, t^*) \\
& = \left\{ \widehat{h}_1 s_{z^* z^*} + \sqrt{\gamma} \widehat{h}_2 s_{z^*} \right\} (z^*, t^*) \\
& \quad - \left[(\sqrt{\gamma} z^*) \widehat{k}''(z^*) + \widehat{k}'(z^*) \left(2\sqrt{\gamma} - \frac{\sqrt{\gamma} z^*}{\widehat{w}_{2z^*}(z^*, t^*)} \widehat{w}_{2z^* z^*}(z^*, t^*) \right) \right. \\
& \qquad \left. - \frac{\sqrt{\gamma} \widehat{k}(z^*)}{\widehat{w}_{2z^*}(z^*, t^*)} \widehat{w}_{2z^* z^*}(z^*, t^*) \right] \frac{1}{w_{2z}\left(\sqrt{\gamma} z^*, \gamma t^*\right)}
\end{aligned}$$

in R'.

Note that it is convenient to keep the z-derivative instead of the z^*-derivative in the multiplier $1/w_{2z}$ in (4.34). Our next goal is to obtain an ODE for the function $\widehat{k}(z^*)$ and solve it via Volterra-like integrals. Resolving (4.34) with respect to $\widehat{k}''(z^*)$, we obtain

$$\widehat{k}''(z^*) = -w_{2z}(\sqrt{\gamma}z^*, \gamma t^*)\left[s_{t^*} - \widehat{h}_1 s_{z^*z^*} - \sqrt{\gamma}\widehat{h}_2 s_{z^*}\right](z^*, t^*) \cdot \frac{1}{\sqrt{\gamma}z^*}$$

$$+ d_1(z^*, t^*)\widehat{k}'(z^*) + d_2(z^*, t^*)\widehat{k}(z^*). \tag{4.35}$$

Using (4.33), we obtain that $0 < \widetilde{g}^{-1} \leq (\sqrt{\gamma}z^*)^{-1} \leq 1$ and functions $d_1,\, d_2 \in C^1(\overline{R}')$. By (4.27),

$$\widehat{k}(z^*) = \int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*} \widehat{k}'(y)\, dy.$$

Hence, using again (4.27), we obtain from (4.35)

$$\widehat{k}'(z^*) = \int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*} d_1(y, t^*)\, \widehat{k}'(y) dy + \int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*} d_2(y, t^*)\, d\tau \int_{g_0(\varepsilon)/\sqrt{\gamma}}^{\tau} \widehat{k}'(y) dy_1$$

$$+ \int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*} w_{2z}(\sqrt{\gamma}y, \gamma t^*)\left[s_{t^*} - \widehat{h}_1 s_{z^*z^*} - \sqrt{\gamma}\widehat{h}_2 s_{z^*}\right](y, t^*)\, \frac{1}{\sqrt{\gamma}y} dy.$$

Considering this as a Volterra integral equation with respect to $\widehat{k}'(z^*)$ and solving it via successive approximations, we obtain

$$\widehat{k}'(z^*) = \int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*} w_{2z}(\sqrt{\gamma}y, \gamma t^*)\left[s_{t^*} - \widehat{h}_1 s_{z^*z^*} - \sqrt{\gamma}\widehat{h}_2 s_{z^*}\right](y, t^*) \frac{1}{\sqrt{\gamma}y} dy$$

$$+ \int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*} G(z^*, y, t^*)\, w_{2z}(\sqrt{\gamma}y, \gamma t^*)\left[s_{t^*} - \widehat{h}_1 s_{z^*z^*} - \sqrt{\gamma}\widehat{h}_2 s_{z^*}\right](y, t^*) dy, \tag{4.36}$$

where the functions $G, G_{z^*}, G_{t^*}, G_{t^*z^*} \in C(\overline{R}')$ and

$$G(z^*, z^*, t^*) = G_{t^*}(z^*, z^*, t^*) = 0. \tag{4.37}$$

Denote $p(z^*, t^*) = s_{t^*}(z^*, t^*)$. Then (4.19) leads to

$$s(z^*, t^*) = \int_0^{t^*} p(z^*, \tau)\, d\tau. \tag{4.38}$$

Denote

$$N(z^*, t^*) = w_{2z}(\sqrt{\gamma}z^*, \gamma t^*). \tag{4.39}$$

Above and below, the function $w_{2z}\left(\sqrt{\gamma}z^*, \gamma t^*\right)$ is understood as follows: First we calculate the derivative $\partial_z w_2(z, t)$. Next we substitute in this derivative $z = \sqrt{\gamma}z^*, t = \gamma t^*$. Hence, by (3.2), (4.3), and (4.8),

$$\frac{1}{\beta} \leq N(z^*, t^*) \leq \frac{1}{\alpha} \tag{4.40}$$

and by (4.32) and (4.39),

$$N_{t^*}\left(z^*,t^*\right)\le\frac{\gamma}{\alpha}, \tag{4.41}$$

where the positive constant β is independent of γ (and so is α, of course; see (3.3)).

Differentiate (4.36) with respect to t^*. Since $\partial_t\widehat{k}'(z^*)\equiv 0$, we obtain using (4.38)

$$\begin{aligned}
&\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}\left\{N\left[p_{t^*}-\widehat{h}_1p_{z^*z^*}-\sqrt{\gamma}\widehat{h}_2p_{z^*}\right]\right\}(y,t^*)\,\frac{dy}{\sqrt{\gamma}y}+\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}\left(N_{t^*}p\right)(y,t^*)\,\frac{dy}{\sqrt{\gamma}y}\\
&-\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}\left[\left(N\widehat{h}_1\right)_{t^*}(y,t^*)\int_0^{t^*}p_{z^*z^*}\left(y,\tau\right)d\tau\right]\frac{dy}{\sqrt{\gamma}y}\\
&-\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}\left[\sqrt{\gamma}\left(N\widehat{h}_2\right)_{t^*}(y,t^*)\int_0^{t^*}p_{z^*}\left(y,\tau\right)d\tau\right]\frac{dy}{\sqrt{\gamma}y}\\
&+\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}G\left(z^*,y,t^*\right)\left\{N\left[p_{t^*}-\widehat{h}_1p_{z^*z^*}-\sqrt{\gamma}\widehat{h}_2p_{z^*}\right]\right\}(y,t^*)\,dy\\
&+\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}G_{t^*}\left(z^*,y,t^*\right)\Bigg\{N\left(y,t^*\right)\Bigg[p\left(y,t^*\right)-\widehat{h}_1\left(y,t^*\right)\int_0^{t^*}p_{z^*z^*}\left(y,\tau\right)d\tau\\
&\qquad\qquad-\sqrt{\gamma}\widehat{h}_2\left(y,t^*\right)\int_0^{t^*}p_{z^*}\left(y,\tau\right)d\tau\Bigg]\Bigg\}\,dy\\
&+\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}G\left(z^*,y,t^*\right)\left(N_{t^*}p\right)(y,t^*)\,dy\\
&-\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}G\left(z^*,y,t^*\right)\left[\left(N\widehat{h}_1\right)_{t^*}(y,t^*)\int_0^{t^*}p_{z^*z^*}\left(y,\tau\right)d\tau\right]dy\\
&-\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}G\left(z^*,y,t^*\right)\left[\sqrt{\gamma}\left(N\widehat{h}_2\right)_{t^*}(y,t^*)\int_0^{t^*}p_{z^*}\left(y,\tau\right)d\tau\right]dy\;=\;0.
\end{aligned}$$

Differentiate this equality with respect to z^* and use (4.37). We obtain

$$
\begin{aligned}
&\left\{N\left[p_{t^*}-\widehat{h}_1 p_{z^*z^*}-\sqrt{\gamma}\widehat{h}_2 p_{z*}\right]\right\}(z^*,t^*)+\left(N_{t^*}p\right)(z^*,t^*)\\
&-\left(N\widehat{h}_1\right)_{t^*}(z^*,t^*)\int_0^{t^*}p_{z^*z^*}(z^*,\tau)\,d\tau-\sqrt{\gamma}\left(N\widehat{h}_2\right)_{t^*}(z^*,t^*)\int_0^{t*}p_{z^*}(z^*,\tau)\,d\tau\\
&+\sqrt{\gamma}z^*\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}G_{z^*t^*}(z^*,y,t^*)\left\{N(y,t^*)\left[p(y,t^*)-\widehat{h}_1(y,t^*)\int_0^{t^*}p_{z^*z^*}(y,\tau)\,d\tau\right.\right.\\
&\qquad\qquad\left.\left.-\sqrt{\gamma}\widehat{h}_2(y,t^*)\int_0^{t^*}p_{z^*}(y,\tau)\,d\tau\right]\right\}dy\\
&+\sqrt{\gamma}z^*\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}G_{z^*}(z^*,y,t^*)\left(N_{t^*}p\right)(y,t^*)\,dy\\
&-\sqrt{\gamma}z^*\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}G_{z^*}(z^*,y,t^*)\left[\left(N\widehat{h}_1\right)_{t^*}(y,t^*)\int_0^{t^*}p_{z^*z^*}(y,\tau)\,d\tau\right]dy\\
&-\sqrt{\gamma}z^*\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}G_{z^*}(z^*,y,t^*)\left[\sqrt{\gamma}\left(N\widehat{h}_2\right)_{t^*}(y,t^*)\int_0^{t^*}p_{z^*}(y,\tau)\,d\tau\right]dy\;=\;0.
\end{aligned}
\tag{4.42}
$$

Here and below, M denotes different positive constants independent of both γ and the function p. Note that by (4.32), (4.33), the right inequality in (4.40), and (4.41),

$$\left\|\left(N\widehat{h}_1\right)_{t^*}\right\|_{C\left(\overline{R}'\right)}+\left\|\left(N\widehat{h}_2\right)_{t^*}\right\|_{C\left(\overline{R}'\right)}+\left\|N_{t^*}\right\|_{C\left(\overline{R}'\right)}\le M\gamma$$

and

$$\left\|G_{z^*t^*}\right\|_{C\left(\overline{R}''\right)}+\left\|G_{t^*}\right\|_{C\left(\overline{R}''\right)}\le M\gamma,$$

where

$$R''=\left\{\frac{g_0(\varepsilon)}{\sqrt{\gamma}}<z^*<\frac{\widetilde{g}}{\sqrt{\gamma}}\right\}\times R'.$$

Hence, dividing both sides of (4.42) by N and using the left inequality in (4.40), we obtain an integral differential equation, from which the integral differential inequality (4.43) follows immediately, where

$$
\begin{aligned}
&\left|p_{t^*}-\widehat{h}_1 p_{z^*z^*}\right|(z^*,t^*)\\
&\le M\left[|p|+|p_{z^*}|\right](z^*,t^*)+M\gamma sgn(t^*)\int_0^{t^*}\left|p_{z^*z^*}(z^*,\tau)\right|d\tau\\
&\quad+M\gamma sgn(t^*)\int_0^{t^*}\left|p_{z^*}(z^*,\tau)\right|d\tau+M\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}\left|p(y,t^*)\right|dy\\
&\quad+M\gamma sgn(t^*)\int_{g_0(\varepsilon)/\sqrt{\gamma}}^{z^*}\int_0^{t^*}\left[|p_{z^*z^*}|+|p_{z^*}|\right](y,\tau)\,d\tau dy\quad\text{ in } R',
\end{aligned}
\tag{4.43}
$$

where

$$sgn(t^*) = \left\{ \begin{array}{c} 1 \text{ if } t^* > 0 \\ -1 \text{ if } t^* < 0 \end{array} \right\}.$$

Hence,

$$sgn(t^*) \int_0^{t^*} |.| \, d\tau = \left| \int_0^{t^*} |.| \, d\tau \right|.$$

Remark. Since we have now singled out the small factor γ, starting from (4.44), below we use old notations for the variables $(z^*, t^*) := (z, t)$ for brevity, keeping in mind, however, that we have not actually returned to old variables.

Since $p = s_{t^*}$, (4.25) and (4.26) imply zero Cauchy data at $\{z = \frac{g_0(\varepsilon)}{\sqrt{\gamma}}\}$,

$$p\left(\frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}}, t\right) = p_z\left(\frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}}, t\right) = 0, \quad t \in \left(-\frac{\varepsilon}{\gamma}, \frac{\varepsilon}{\gamma}\right). \tag{4.44}$$

It is well known that when applying Carleman estimates, it is often more convenient to work with inequalities rather than with equations; see, e.g., [21]. The goal of the next three subsections is to prove that (4.43) and (4.44) imply that $p(z,t) = 0$ in R', which in turn obviously implies (4.28) and (4.29).

4.2. Carleman estimate. Let $\xi = const. \in (0,1)$. Introduce the function $\psi(z,t)$ by

$$\psi\left(z,t\right) = \left(z - \frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}}\right) + t^2 + \xi.$$

Choose an arbitrary constant $\delta > 0$ such that

$$\delta < \min\left[1, \frac{\widetilde{g} - g_0\left(\varepsilon\right)}{\sqrt{\gamma}}, \left(\frac{\varepsilon}{\gamma}\right)^2\right].$$

Denote

$$K_\delta = \left\{(z,t) : \psi\left(z,t\right) < \xi + \delta, z > \frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}}\right\}.$$

Hence, the boundary ∂K_δ of the domain K_δ is

$$\partial K_\delta = \partial_1 K_\delta \cup \partial_2 K_\delta, \tag{4.45}$$

where

$$\partial_1 K_\delta = \left\{(z,t) : z = \frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}}, t^2 < \delta\right\}, \tag{4.46}$$

$$\partial_2 K_\delta = \left\{(z,t) : \left(z - \frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}}\right) + t^2 + \xi = \xi + \delta, z > 0\right\}. \tag{4.47}$$

Hence, by (4.44) and (4.46),

$$p \mid_{\partial_1 K_\delta} = p_z \mid_{\partial_1 K_\delta} = 0. \tag{4.48}$$

By (4.31), $K_\delta \subset R'$. Clearly $K_{\delta'} \subset K_\delta$ for all $\delta' \in (0,\delta)$.

Let $\lambda, \nu > 1$ be two large parameters which we will choose later. Introduce the CWF $\varphi(z,t)$ by

$$\varphi(z,t) = \exp\left(\lambda \psi^{-\nu}\right). \tag{4.49}$$

Lemma 1 (Carleman estimate; see, e.g., section 2.2 in [21] and section 1 of Chapter 4 in [22]). *There exist a sufficiently large number $\nu = \nu_0 = \nu_0(\xi,\delta,\beta_1,\beta_2,\beta_3)$ (see (4.18) for β_1,β_2,β_3) and a positive constant $C = C(\xi,\delta,\beta_1,\beta_2,\beta_3)$, both ν_0 and C independent of the number γ, such that for all values of the parameter $\lambda > 1$ and all functions $u \in C^{2,1}\left(\overline{K}_\delta\right)$ the following pointwise Carleman estimate holds in K_δ:*

$$(u_t - h_1 u_{zz})^2 \varphi^2 \geq \frac{C}{\lambda}\left(u_t^2 + u_{zz}^2\right)\varphi^2 + C\lambda u_z^2 \varphi^2 + C\lambda^3 u^2 \varphi^2 + U_z + V_t, \tag{4.50}$$

where functions U and V satisfy

$$|U| + |V| \leq C\lambda\left[u_t^2 + u_z^2 + \lambda^2 u^2\right]\varphi^2. \tag{4.51}$$

Below, C denotes different positive constants dependent on numbers $\xi,\delta,\beta_1,\beta_2,\beta_3$ but independent of the number γ.

4.3. Estimates of integrals. While (4.50) provides an estimate of the left-hand side of the inequality (4.43) from below, we also need to estimate its right-hand side from above in the presence of the CWF φ. It is not a problem to estimate $M^2\left[|p| + |p_z|\right]\varphi^2$. However, the presence of Volterra-like integrals in (4.43) presents a difficulty. In the above cited previous works [4], [5], [6], [7], [8], [9], [10], [11], [12], [18], [19], [20], [21], [22], [23], [24] only integrals with respect to time t,

$$\int_0^t (.)\, d\tau,$$

were estimated (Lemma 2). A new element of this publication is Lemmata 3 and 4, which are concerned with estimates of integrals

$$\int_{g_0(\varepsilon)}^z (.)\, dy, \quad \int_{g_0(\varepsilon)}^z \int_0^t (.)\, d\tau dy.$$

The z-integration occurs because of the involvement of both the function $\widetilde{k}$ and its derivatives $\widetilde{k}', \widetilde{k}''$ in (4.16), which was not the case in the past.

Lemma 2 (see [21, p. 77]). *The following estimate holds for all real valued functions $f \in L_2(K_\delta)$:*

$$\int_{K_\delta}\left(\int_0^t f(z,\tau)\, d\tau\right)^2 \varphi^2 dz dt \leq \frac{C}{\lambda}\int_{K_\delta} f^2(z,t)\,\varphi^2 dz dt \quad \forall \lambda > 1.$$

We now prove the following lemma.

Lemma 3. *The following estimate holds for all real valued functions $f \in L_2(K_\delta)$:*

$$\int_{K_\delta}\left(\int_{g_0(\varepsilon)}^z f(y,t)\, dy\right)^2 \varphi^2 dz dt \leq \frac{C}{\lambda^2}\int_{K_\delta} f^2(z,t)\,\varphi^2 dz dt \quad \forall \lambda > 1.$$

Proof.

$$\int_{K_\delta}\left(\int_{g_0(\varepsilon)}^{z} f(y,t)\,dy\right)^2\varphi^2 dzdt=\int_{-\sqrt{\delta}}^{\sqrt{\delta}}dt\int_0^{\delta-t^2}\left(\int_{g_0(\varepsilon)}^{z} f(y,t)\,dy\right)^2\varphi^2 dz. \tag{4.52}$$

We have by (4.49)

$$\begin{aligned}
&\int_0^{\delta-t^2}\left(\int_{g_0(\varepsilon)}^{z} f(y,t)\,dy\right)^2\varphi^2 dz\\
&\quad=\int_0^{\delta-t^2}\left(\int_{g_0(\varepsilon)}^{z} f(y,t)\,dy\right)^2\exp\left[2\lambda\left(z+t^2+\xi\right)^{-\nu}\right]dz\\
&\quad=\frac{1}{2\lambda\nu}\int_0^{\delta-t^2}\left(\int_{g_0(\varepsilon)}^{z} f(y,t)\,dy\right)^2\psi^{\nu+1}\frac{d}{dz}\left[-\exp\left(2\lambda\psi^{-\nu}\right)\right]dz\\
&\quad\le\frac{(\xi+\delta)^{\nu+1}}{2\lambda\nu}\int_0^{\delta-t^2}\left(\int_{g_0(\varepsilon)}^{z} f(y,t)\,dy\right)^2\frac{d}{dz}\left[-\exp\left(2\lambda\psi^{-\nu}\right)\right]dz\\
&\quad\le\frac{C}{\lambda}\Bigg\{-\exp\left(2\lambda\left(\xi+\delta\right)^{-\nu}\right)\left(\int_0^{\delta-t^2} f(y,t)\,dy\right)^2\\
&\qquad\qquad+2\int_0^{\delta-t^2}\varphi^2 f(z,t)\int_{g_0(\varepsilon)}^{z} f(y,t)\,dydz\Bigg\}\\
&\quad\le\frac{C}{\lambda}\int_0^{\delta-t^2}\varphi^2 f(z,t)\int_{g_0(\varepsilon)}^{z} f(y,t)\,dydz.
\end{aligned}$$

Hence, we have obtained that

$$\int_0^{\delta-t^2}\left(\int_{g_0(\varepsilon)}^{z} f(y,t)\,dy\right)^2\varphi^2 dz\le\frac{C}{\lambda}\int_0^{\delta-t^2}\varphi^2 f(z,t)\int_{g_0(\varepsilon)}^{z} f(y,t)\,dydz. \tag{4.53}$$

By the Cauchy–Schwarz inequality

$$\begin{aligned}
&\frac{C}{\lambda}\int_0^{\delta-t^2}\varphi^2 f(z,t)\int_{g_0(\varepsilon)}^{z} f(y,t)\,dydz\\
&\le\frac{C}{\lambda}\left[\int_0^{\delta-t^2} f^2(z,t)\,\varphi^2 dz\right]^{1/2}\left[\int_0^{\delta-t^2}\left(\int_{g_0(\varepsilon)}^{z} f(y,t)\,dy\right)^2\varphi^2 dz\right]^{1/2}.
\end{aligned}$$

Substituting the latter into (4.53), we obtain

$$\left[\int_0^{\delta-t^2}\left(\int_{g_0(\varepsilon)}^{z} f(y,t)\,dy\right)^2\varphi^2 dz\right]^{1/2}\le\frac{C}{\lambda}\left[\int_0^{\delta-t^2} f^2(z,t)\,\varphi^2 dz\right]^{1/2}.$$

Squaring both sides of this inequality, we obtain the target estimate of this lemma. $\square$

LEMMA 4. *The following estimate holds for all real valued functions $f \in L_2(K_\delta)$:*

$$\int_{K_\delta}\left(\int_{g_0(\varepsilon)}^{z}\int_0^t f(y,t)\,d\tau dy\right)^2\varphi^2 dzdt \le \frac{C}{\lambda^3}\int_{K_\delta} f^2(z,t)\,\varphi^2 dzdt \quad \forall\lambda>1.$$

Proof. Denote

$$F(y,t)=\int_0^t f(y,t)\,d\tau.$$

Applying sequentially Lemmata 3 and 2, we obtain

$$\begin{aligned}&\int_{K_\delta}\left(\int_{g_0(\varepsilon)}^{z}\int_0^t f(y,t)\,d\tau dy\right)^2\varphi^2 dzdt\\ &=\int_{K_\delta}\left(\int_{g_0(\varepsilon)}^{z}F(y,t)dy\right)^2\varphi^2 dzdt \;\le\; \frac{C}{\lambda^2}\int_{K_\delta}F^2(z,t)\,\varphi^2 dzdt\\ &\le\frac{C}{\lambda^3}\int_{K_\delta} f^2(z,t)\,\varphi^2 dzdt. \qquad\square\end{aligned}$$

4.4. Finishing the proof of Theorem 1. Square both sides of the inequality (4.43), multiply by the function φ^2, integrate over the domain K_δ, and apply (4.50) and (4.51) to the left-hand side using Gauss's formula and (4.45)–(4.48). Also, apply Lemmata 2–4 to the right-hand side. We obtain

$$\begin{aligned}&\frac{C}{\lambda}\int_{K_\delta}\left(p_t^2+p_{zz}^2\right)\varphi^2 dzdt + C\lambda\int_{K_\delta}\left(p_z^2+\lambda^2p^2\right)\varphi^2 dzdt\\ &\le M\frac{\gamma^2}{\lambda}\int_{K_\delta}p_{zz}^2\varphi^2 dzdt+\frac{M}{\lambda^3}\int_{K_\delta}p_{zz}^2\varphi^2 dzdt+M\int_{K_\delta}\left(p_z^2+p^2\right)\varphi^2 dzdt\\ &\quad + C\lambda\exp\left[2\lambda\left(\xi+\delta\right)^{-\nu}\right]\int_{\partial_2K_\delta}\left(p_t^2+p_z^2+\lambda^2p^2\right)dS \quad \forall\lambda>1.\end{aligned}$$

Since constants C and M are independent of γ, we choose γ so small that $M\gamma^2<C/4$. Next, we choose $\lambda_0>1$ so large that $M/\lambda_0^2<C/4$ and $M<C\lambda_0/2$. Then we obtain with a new constant C

$$\begin{aligned}&\frac{1}{\lambda}\int_{K_\delta}\left(p_t^2+p_{zz}^2\right)\varphi^2 dzdt+\lambda\int_{K_\delta}\left(p_z^2+\lambda^2p^2\right)\varphi^2 dzdt\\ &\le C\lambda\exp\left[2\lambda\left(\xi+\delta\right)^{-\nu}\right]\int_{\partial_2K_\delta}\left(p_t^2+p_z^2+\lambda^2p^2\right)dS \quad \forall\lambda\ge\lambda_0.\end{aligned}\tag{4.54}$$

Consider a number $\delta'\in(0,\delta)$. Since $\varphi^2\ge\exp[2\lambda(\xi+\delta')^{-\nu}]$ in $K_{\delta'}$ and $K_{\delta'}\subset K_\delta$, we obtain from (4.54)

$$\begin{aligned}&\exp\left[2\lambda\left(\xi+\delta'\right)^{-\nu}\right]\int_{K_{\delta'}}p^2dzdt\\ &\le\frac{C}{\lambda^2}\exp\left[2\lambda\left(\xi+\delta\right)^{-\nu}\right]\int_{\partial_2K_\delta}\left(p_t^2+p_z^2+\lambda^2p^2\right)dS \quad \forall\lambda\ge\lambda_0.\end{aligned}$$

Dividing this inequality by $\exp[2\lambda(\xi+\delta')^{-\nu}]$, we obtain

$$\begin{aligned}&\int_{K_{\delta'}} p^2 dzdt\\ &\le \frac{C}{\lambda^2}\exp\left\{-2\lambda\left[(\xi+\delta')^{-\nu}-(\xi+\delta)^{-\nu}\right]\right\}\int_{\partial_2 K_\delta}\left(p_t^2+p_z^2+\lambda^2p^2\right)dS \quad \forall\lambda\ge\lambda_0.\end{aligned}$$

Note that $(\xi+\delta')^{-\nu}-(\xi+\delta)^{-\nu}>0$. Hence, letting $\lambda\to\infty$, we obtain

$$\int_{K_{\delta'}} p^2 dzdt = 0.$$

Hence, $p=0$ in $K_{\delta'}$. Since δ' is an arbitrary number from the interval $(0,\delta)$, $p=0$ in K_δ. By (4.38), the function $q=0$ in K_δ. Hence, (4.16) together with (4.27) implies that

$$\widetilde{k}(z)=0 \quad \text{ for } z\in\left(\frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}},\frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}}+\delta\right).$$

Dropping $\widetilde{k}(z)$ in (4.16) for $z\in(\frac{g_0(\varepsilon)}{\sqrt{\gamma}},\frac{g_0(\varepsilon)}{\sqrt{\gamma}}+\delta)$, we obtain a parabolic equation

$$q_t-h_1\left(z,t\right)q_{zz}-h_2\left(z,t\right)q_z=0, \quad (z,t)\in\left(\frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}},\frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}}+\delta\right)\times\left(-\frac{\varepsilon}{\gamma},\frac{\varepsilon}{\gamma}\right)$$

with the Cauchy data at $\{z=\frac{g_0(\varepsilon)}{\sqrt{\gamma}}\}$,

$$q\left(\frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}},t\right)=q_z\left(\frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}},t\right)=0, \quad t\in\left(-\frac{\varepsilon}{\gamma},\frac{\varepsilon}{\gamma}\right).$$

Hence, the above-mentioned unique continuation theorem implies that $q(z,t)=0$ for $(z,t)\in(\frac{g_0(\varepsilon)}{\sqrt{\gamma}},\frac{g_0(\varepsilon)}{\sqrt{\gamma}}+\delta)\times(-\frac{\varepsilon}{\gamma},\frac{\varepsilon}{\gamma})$. Therefore, we can now "shift" the left boundary $\{z=\frac{g_0(\varepsilon)}{\sqrt{\gamma}}\}$ of the rectangle R' to the right by δ via replacing R' with the rectangle

$$R'_\delta=\left(\frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}}+\delta,\frac{g_0\left(\varepsilon\right)}{\sqrt{\gamma}}+2\delta\right)\times\left(-\frac{\varepsilon}{\gamma},\frac{\varepsilon}{\gamma}\right).$$

Next, we can repeat the above process. This way we arrive at target equalities (4.28) and (4.29).

5. Application examples.

5.1. A diffusion model in population dynamics. In this section we discuss the assumptions of Theorem 1 for an example from population dynamics, given in the form (2.4): Aronson in [1] proposes the diffusion model

$$v_t=(\phi(v)v)_{xx} \tag{5.1}$$

in the absence of drift, where $v(x,t)$ is the population density at location x and time t and the diffusion coefficient $\phi(v)$ is the population density dependent limit (as the minimal individual step length tends to zero) of the second moment of the transition probability between different locations. The function ϕ plays the role of k in the setting of Theorem 1. With $u(x,t) = \int_0^x v(\xi,t)\,d\xi + f_0(t)$ and an integration with respect to the space variable this can be rewritten as

$$u_t - (\phi(u_x)u_x)_x = c, \tag{5.2}$$

with c depending on time only, which without loss of generality we can set to zero (as an inspection of the proof of Theorem 1 shows), which yields

$$u_t = (\phi(u_x)u_x)_x \tag{5.3}$$

so that with $k = \phi$ we get the PDE considered in Theorem 1. The boundary and initial data as well as the conditions imposed on them have the following interpretation in the context of this example:

- Conditions (2.7), (3.1) are natural assumptions according to [1]. Condition (3.6) can be guaranteed by imposing nonvanishing population densities at the boundary, which seems to be natural as well.
- $u(x,0) = r(x)$; i.e., $v(x,0) = r'(x)$ means that the population density is known everywhere at time $t = 0$, e.g., from a census at that instance of time.
- $u(0,t) = f_0(t)$ is satisfied by our setting of the integration constants.
- $u_x(0,t) = v(0,t) = g_0(t)$ with $g' \geq 0$: The population density at location $x = 0$ is observed for all time instances and increases with time.
- $u_x(L,t) = v(L,t) = g_1(t) = \tilde{g}$ and $[g_0(\varepsilon), \tilde{g}]$ is the interval on which k can be identified according to Theorem 1: The population density at location $x = L$ is observed for all time instances and is constant in time, where this constant should be as large as possible, e.g., equal to some saturation value.
- $u_{xx} = v_x \geq \alpha > 0$: The population density increases from left to right.

5.2. Nonlinear magnetics. Quasi-static magnetic fields can be described by a subset of Maxwell's equations, more precisely, Ampère's law in the quasi-static case

$$\nabla \times \mathbf{H} = \mathbf{J}$$

and Faraday's law

$$\nabla \times \mathbf{E} = -\mathbf{B}_t\,,$$

combined with the constitutive relations

$$\mathbf{J} = \gamma \mathbf{E} + \mathbf{J}^{imp}\,, \quad \mathbf{H} = \nu \mathbf{B},$$

where $\mathbf{H}$ is the magnetic field intensity, $\mathbf{B}$ the magnetic flux density, $\mathbf{E}$ the electric field, γ the electric conductivity, ν the magnetic reluctivity (i.e., the reciprocal of the magnetic permeability μ), and J_{imp} the impressed current density; see, e.g., [6], [10], [16]. Using the fact that $\mathbf{B}$ is solenoidal,

$$\nabla \cdot \mathbf{B} = 0\,,$$

we can make use of a magnetic vector potential $\mathbf{A}$ satisfying

$$\mathbf{B} = \nabla \times \mathbf{A}\,,$$

which leads to the system of PDEs

$$\gamma \mathbf{A}_t + \nabla \times (\nu \nabla \times \mathbf{A}) = J_{imp} \tag{5.4}$$

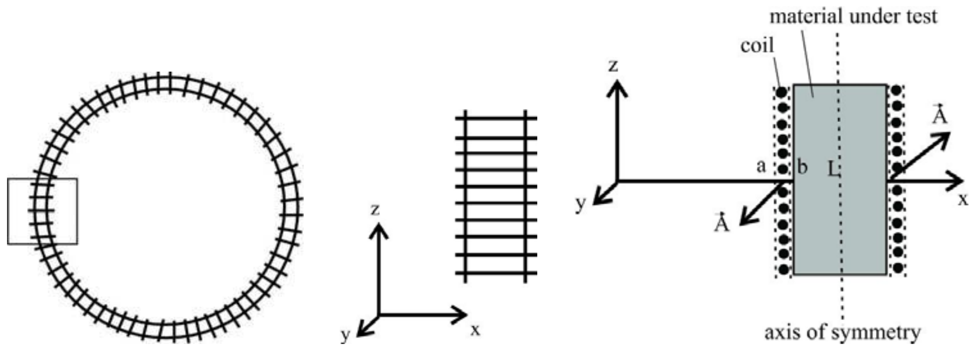


FIG. 5.1. *Schematic of probe with coil (left), quasi-straight detail (middle), and cut along x-z plane (right).*

for the vector field $\mathbf{A}$. In the situation of high magnetic fields, the parameter ν is not constant but depends on the magnetic flux density, i.e.,

$$\mathbf{H} = \nu(\mathbf{B})\mathbf{B}\,. \tag{5.5}$$

By an appropriate experimental setup, this can be reduced to the spatially one-dimensional case: Consider a ring-shaped probe entwined with an excitation/measurement coil according to Figure 5.1 (left) with a large interior radius so that the curvature can be neglected (see Figure 5.1 (middle)), and the magnetic flux density points into the z direction but does not vary in the z direction. Moreover, we consider a cut along the x-z plane (see Figure 5.1 (right)), in which $\mathbf{A}$ must be parallel to the y axis and dependence of $\mathbf{A}$, $\mathbf{B}$ on y can be neglected, so that altogether

$$\mathbf{B}(x, y, z) = (0, 0, B^z(x))^T,$$

or equivalently

$$\mathbf{A}(x, y, z) = (0, A^y(x), 0)^T.$$

With $u(x) := A^y(x)$ the system (5.4) becomes the spatially one-dimensional model problem

$$\gamma u_t - (\nu(u_x)u_x)_x = 0 \quad \text{in } (b, L) \times (0, T);$$

i.e., with $k = \frac{1}{\gamma}\nu$ we arrive at the PDE considered in Theorem 1. An interpretation of the conditions of Theorem 1 in the context of this example can be made as follows:

- Conditions (2.7), (3.1) are natural conditions on the coefficient also in the context of magnetics, unless very special materials are considered. Also, the assumption of knowledge of the coefficient on a short interval $(g_0(-c), g_0(\varepsilon))$ according to (3.7) is not restrictive: Note that to resolve the full range of the B-H curve, one will typically choose g_0 small and $\widetilde{g}$ large, i.e., close to the saturation value. Hence (3.7) means that the coefficient is known for small magnetic fields, which is actually the case, since for moderate amplitudes of the fields, ν is a constant that is either known or can be determined from simpler measurements. It therefore also makes sense to exclude zero magnetic flux density from the range of measurements according to (3.6).

- Neumann boundary data $u_x(0,t) = g_0(t)$, $u_x(L,t) = g_1(t)$ here have the meaning of a magnetic flux density, which can be extracted from measurements of the magnetic flux through two coils positioned on both sides of the material strip. The magnetic flux can be controlled via the impressed current through these coils such that it is monotonically nondecreasing (and small) at the left endpoint as well as constant (and large) at the right endpoint to achieve $g_0' \geq 0$, $g_1' = 0$, and a wide interval (a, b).
- Via the relation $\mathbf{u}_t = \mathbf{E}$ with $\mathbf{E}$ the electric field, one can obtain the initial and boundary data of u by time integration of electric field measurements.

The conventional measurement setup for determining ν consists of just an excitation (and measurement) coil wound around the probe and does not enable us to collect all the data required from the point of view of our uniqueness theorem (especially not electric field measurements inside the probe which would be required for giving initial data for u). In this sense, there is a gap between theory and practice. Still, the model problem considered in this paper gives important insight, since the correct form of the PDE with a coefficient depending on the space derivative of u is considered. For more details on a PDE-based approach for nonlinearity identification in magnetics, we refer the reader to [17].

6. Numerical results. In this section, we briefly outline a numerical method for determining the coefficient k in (2.1) and provide some computational results.

A method for forward magnetic field computations in the nonlinear case that has recently attracted considerable interest is the so-called harmonic balancing finite element method (cf., e.g., [2], [9], [25]). It is based on the idea of making a multiharmonic ansatz in time to capture the higher harmonics physically arising due to nonlinearity. Mathematically speaking, one makes use of a special time discretization based on trigonometric polynomials; i.e., in a complex valued setting, time harmonic functions $t \mapsto \exp(\imath nt)$, where $\imath = \sqrt{-1}$. An idea for using this approach in parameter identification was presented in [15] for a hyperbolic PDE. Here we will derive such a multiharmonic parameter identification method in the original context of magnetics, i.e., a parabolic PDE.

Thinking of the boundary data as extended periodically and sufficiently smoothly to the larger interval $[0, \bar{T}] = [0, T + T_1]$ with some $T_1 > 0$, we make an ansatz

$$u(x,t) \approx u^N(x,t) := \sum_{n=-N}^{N} \exp(\imath n\omega t)\hat{u}_n(x)\,, \tag{6.1}$$

with $\omega = 2\pi/\bar{T}$. (Note that continuation of the boundary values beyond T by the causality of the problem does not influence u for $t < T$.) Inserting into the PDE, we get

$$\sum_{j=-N}^{N} e^{\imath j\omega t} \left(\imath\omega j \hat{u}_j - \left(k \left(\sum_{n=-N}^{N} e^{\imath n\omega t} \hat{u}_{n\,x} \right) \hat{u}_{j\,x} \right)_x \right) = 0\,.$$

Testing this with the functions

$$t \mapsto \frac{1}{\bar{T}} e^{-\imath l\omega t}$$

and using the orthonormality

$$\frac{1}{\bar{T}} \int_0^{\bar{T}} e^{-\imath l\omega t} e^{\imath j\omega t}\, dt = \delta_{lj}, \tag{6.2}$$

where δ_{lj} is the Kronecker symbol, yields

$$\imath\omega l\hat{u}_l - \frac{1}{\bar{T}}\int_0^{\bar{T}} \sum_{j=-N}^{N} e^{\imath(j-l)\omega t}\left(k\left(\sum_{n=-N}^{N} e^{\imath n\omega t}\hat{u}_{n\,x}\right)\hat{u}_{j\,x}\right)_x dt \ = 0 \quad \forall l \in \{-N,\dots,N\}\,. \tag{6.3}$$

To eliminate the time integral, we use an appropriate approximation for k that enables us to take advantage of the orthogonality (6.2). For this purpose we make use of a polynomial ansatz

$$k(z) \approx \sum_{p=0}^{P} \alpha_p z^p\,,$$

which is justified by the fact that in the applications we have in mind, k is typically a smooth function. Since the multinomial theorem yields

$$\left(\sum_{n=-N}^{N} e^{\imath n\omega t}\hat{u}_{n\,x}\right)^p = \sum_{\substack{\mathbf{p}=(p_{-N},\dots,p_N)\in\mathbb{N}_0^{2N+1}\\ \sum_{n=-N}^{N} p_n = p}} \begin{pmatrix} p \\ \mathbf{p}\end{pmatrix} \prod_{m=-N}^{N}\left(e^{\imath m\omega t}\hat{u}_{m\,x}\right)^{p_m}\,,$$

with the multinomial coefficients

$$\begin{pmatrix} p \\ (p_{-N},\dots,p_N)\end{pmatrix} = \frac{p!}{p_{-N}!\cdots p_N!}\,,$$

by (6.2) one arrives at a system of space dependent PDEs

$$\imath\omega l\hat{u}_l - \left(\sum_{n=-N}^{N}\sum_{p=0}^{P}\alpha_p \vec{c}^{\,p}_{l-n}\hat{u}_{n\,x}\right)_x = 0, \quad l \in \{-N,\dots,N\}\,, \tag{6.4}$$

where

$$\vec{c}^{\,p}_s = \sum_{\mathbf{p}\in\mathcal{I}(p,s)} \begin{pmatrix} p \\ \mathbf{p}\end{pmatrix} \prod_{m=-N}^{N}(\hat{u}_{m\,x})^{p_m}\,,$$

$$\mathcal{I}(p,s) = \left\{\mathbf{p}=(p_{-N},\dots,p_N)\subseteq\mathbb{N}_0^{2N+1} \;\middle|\; \sum_{n=-N}^{N} p_n = p \;\wedge\; \sum_{n=-N}^{N} np_n = s\right\};$$

i.e., the coefficients $\vec{c}^{\,p}_{l-n}$ depend nonlinearly on $(\hat{u}_{-N\,x},\dots,\hat{u}_{N\,x})$. The boundary conditions and measurements become

$$\hat{u}_{l\,x}(0) = \hat{g}_{0\,l},\ \hat{u}_{l\,x}(L) = \hat{g}_{1\,l}, \quad l \in \{-N,\dots,N\}, \tag{6.5}$$

$$\hat{u}_l(0) = \hat{f}_{0\,l},\ \hat{u}_l(L) = \hat{f}_{1\,l}, \quad l \in \{-N,\dots,N\}, \tag{6.6}$$

with

$$\hat{f}_{i\,l} = \frac{1}{\bar{T}}\int_0^{\bar{T}} \exp(-\imath l\omega\tau) f_i(\tau)\ d\tau\,,$$

and analogously for g_i, $i = 0, 1$. Note that we consider the Neumann data as boundary conditions (and the Dirichlet data as measurement) in our numerical tests below in order to be able to directly prescribe g_0 monotonically increasing on the subinterval $[0, T]$ and g_1 constant. To avoid singularity of the system due to the elliptic Neumann problem for index $l = 0$, we make the additional normalization assumption $\hat{u}_0 = \frac{1}{\bar{T}} \int_0^{\bar{T}} u^N(\cdot, t)\, dt = 0$. Physically, this corresponds to the fact that in experiments only *higher* harmonics (i.e., multiples $n\omega$ with $|n| \geq 1$ of the basic frequency ω) appear. From a mathematical point of view this normalization corresponds to a $\bar{T}$ periodicity assumption on the antiderivative of u^N, while the multiharmonic ansatz implies periodicity of u^N itself. To keep compatibility, we add this assumption of a vanishing time integral also to the conditions imposed on the periodic extension of the boundary values f_0, f_1, g_0, g_1.

If k is real valued, we expect to get a real valued solution and therefore additionally stipulate

$$\hat{u}_{-n} = \overline{\hat{u}_n}\,. \tag{6.7}$$

We refer the reader to [15] for more details in the context of a hyperbolic model problem and to [2] for a well-posedness proof of the forward problem in this form under the conditions we have made here for the coefficient function k. The inverse problem of reconstructing k from boundary measurements of u in this kind of discretization amounts to determining the vector of polynomial coefficients $\underline{\alpha} = (\alpha_0, \ldots, \alpha_p)$ from boundary data of $\hat{u}_l$, $l = 0, \ldots, N$; cf. [15]. Written as a system of equations

$$F_N^P(\underline{\alpha}) = y_N\,, \tag{6.8}$$

where $F_N^P : \underline{\alpha} \mapsto (\hat{u}_n(0), \hat{u}_n(L))_{n=0}^N$ with $(\hat{u}_n)_{n=0}^N$ solving (6.4), (6.5), (6.7), this can be viewed as the projected version of an operator equation

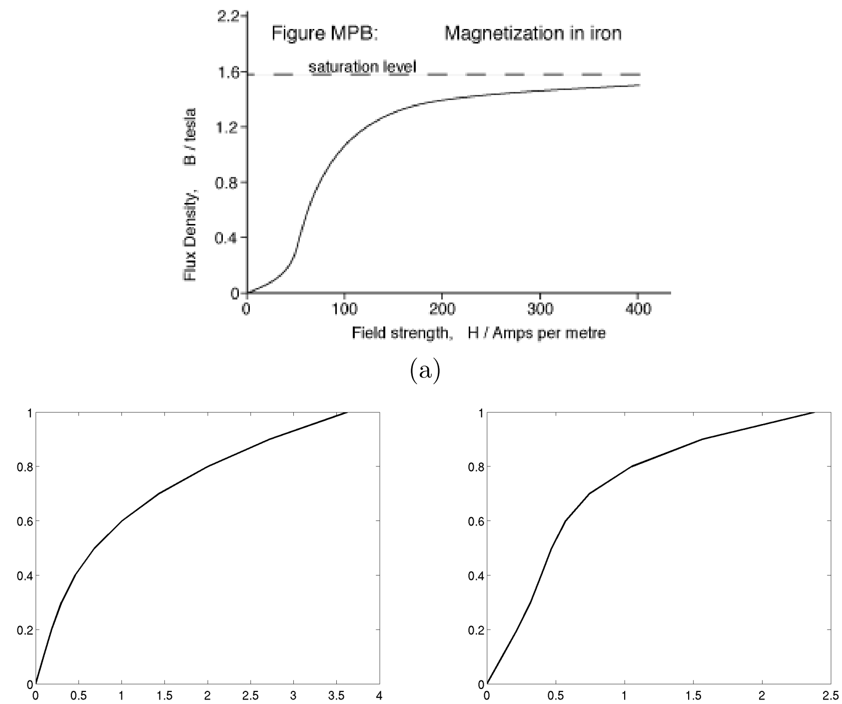
$$F^P(\underline{\alpha}) = y \tag{6.9}$$

resulting from replacing N by ∞ in (6.1), i.e., taking the full expansion of u into account. More precisely, our special time discretization corresponds to an L^2 projection in data space to the subspace spanned by the first $2N + 1$ trigonometric polynomials. These L^2 projections converge pointwise to the identity due to the fact that $\{t \mapsto \frac{1}{\sqrt{\bar{T}}} e^{\imath j \omega t} \mid j \in \mathbb{Z}\}$ forms a complete orthonormal system in $L^2_{\mathbb{C}}(0, \bar{T})$. Keeping in mind the fact that the original infinite-dimensional inverse problem of recovering k is ill-posed in the sense of instability, here we rely on the regularizing effect of discretization. As a matter of fact, we can combine the approximation error estimates from [2] with Theorem 1 in [14] to conclude that (6.8) is a regularization method for the solution of (6.9) with N acting as a regularization parameter. Since the convergence and stability properties of this method hold independently of the polynomial degree P, this can be regarded as a regularization method for the original problem of determining k from boundary measurements of u.

We can apply Newton's method to (6.8),

$$\underline{\alpha}^{m+1} = \underline{\alpha}^m + \theta \underline{\beta} \quad \text{with } F_N^{P'}(\underline{\alpha}^m)\underline{\beta} = y - F_N^P(\underline{\alpha}^m),$$

to iteratively recover $\underline{\alpha}$. Here θ is an appropriately chosen stepsize that guarantees strictly monotone decrease of the residual. To fully discretize this method it remains



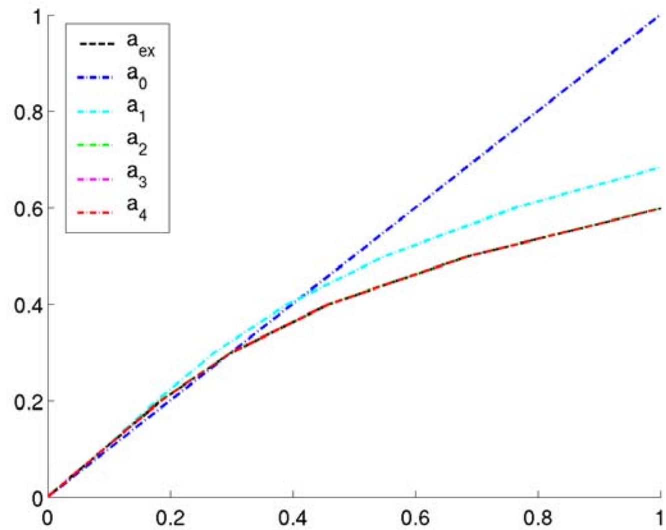
(a) (b) (c)

FIG. 6.1. *Typical B-H curve for iron ((a), schematic) and two synthetic test examples ((b), (c)).*

to define a space discretization of the space dependent functions $\hat{u}_l$, which we do by using finite elements on a uniform grid.

The application we have in mind here is nonlinear magnetics (cf. subsection 5.2), where the nonlinearity is usually considered in terms of the so-called B-H curves, i.e., a plot of the magnetic flux density $\mathbf{B}$ over the magnetic field intensity $\mathbf{H}$. By the relation (5.5), in our context this translates into plotting the function a, where $a^{-1}(z) = k(z)z$. A typical B-H curve for iron is displayed in Figure 6.1(a), where the ratio between the slope at $B = 0$ and the slope at the saturation level $B = B_{sat}$ amounts to the relative reluctivity $\nu_{rel} \sim \frac{1}{5000}$. For testing the proposed method, we constructed two test problems with similar behavior. Upon renormalization to $B_{sat}^{norm} = 1$, $\nu_{rel}^{norm} = \frac{1}{10}$, the values $H_{sat}^{norm} = 3.6$ and $H_{sat}^{norm} = 2.4$ as they appear in the synthetic test examples in Figure 6.1(b), (c) indeed turn out to be in a realistic range.

For the first test example (i.e., Figure 6.1(b)), Figure 6.2 shows the iteration history of the proposed method for exact and for randomly perturbed data with a noise level of 1 percent, as compared to the true curve a_{ex} in black. In our inverse computations, we used $N = 5$ higher harmonics. To avoid an inverse crime, we generated the data with a larger number N in the multiharmonic ansatz. Figure 6.3 shows the development of the residuals during the Newton iteration in the exact and noisy case, respectively. In Figures 6.4 and 6.5 we give another example of such a B-H curve, namely the one corresponding to Figure 6.1(c). This example exhibits a changing sign of curvature as typical for some materials; see, e.g., Figure 6.1(a). The proposed method is able to also satisfactorily reproduce this more complicated behavior. Note that in each of these tests, the starting value for k is just a constant



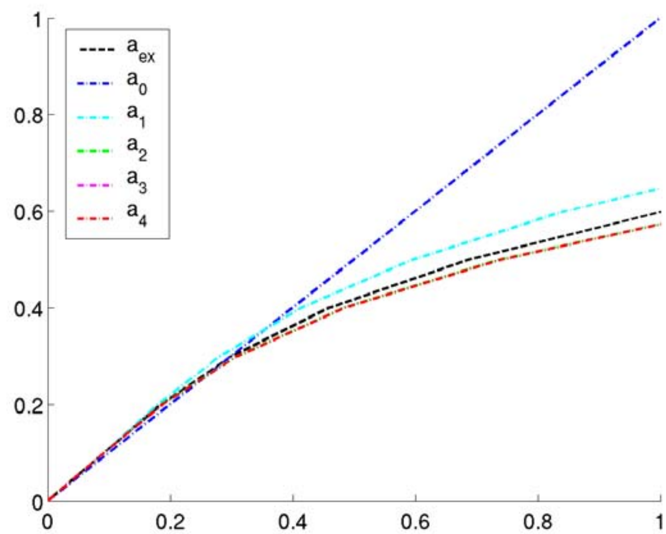


FIG. 6.2. *Starting value* a_0 *and Newton iterations* $\mathrm{a}_1, \mathrm{a}_2, \mathrm{a}_3, \mathrm{a}_4$ *of* a*, where* $\mathrm{a}^{-1}(z) = k(z)z$*, compared to exact curve* a_{ex} *for the example in Figure* 6.1(b); *top: exact data, bottom: noisy data (*1 *percent noise).*

function, corresponding to the fact that in practice the constant reluctivity coefficient for small magnetic fields is typically known.

Acknowledgment. The authors are grateful to anonymous referees for their valuable comments, which helped to improve the quality of presentation.

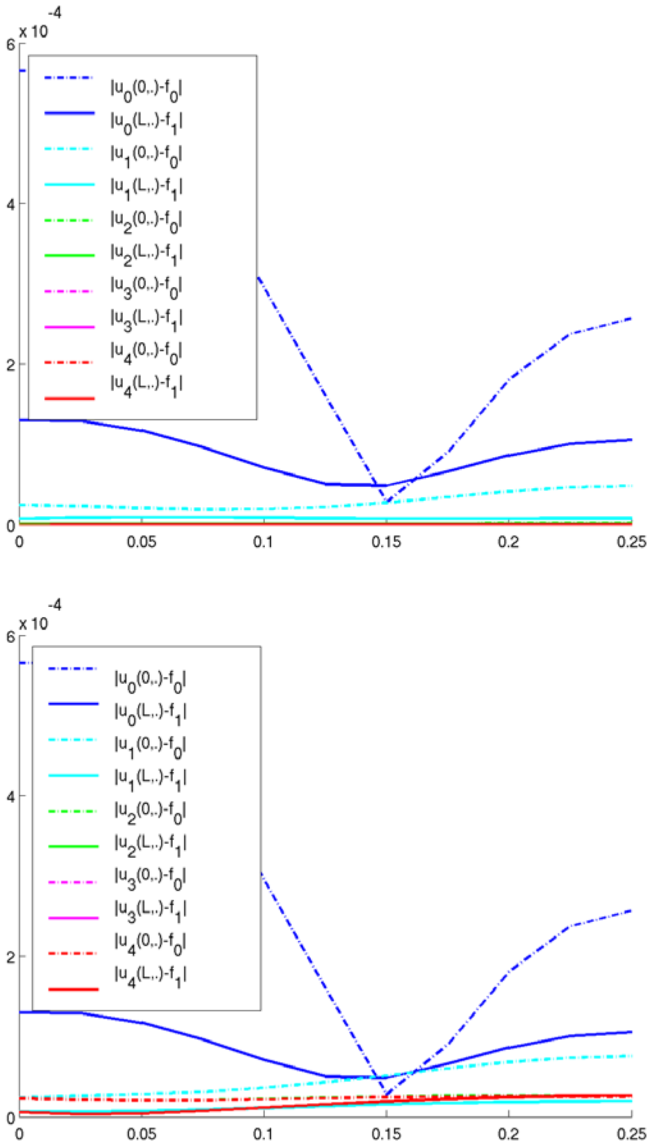


FIG. 6.3. *Starting value and Newton iterations of the residual corresponding to Figure* 6.2; *top: exact data, bottom: noisy data (*1 *percent noise).*

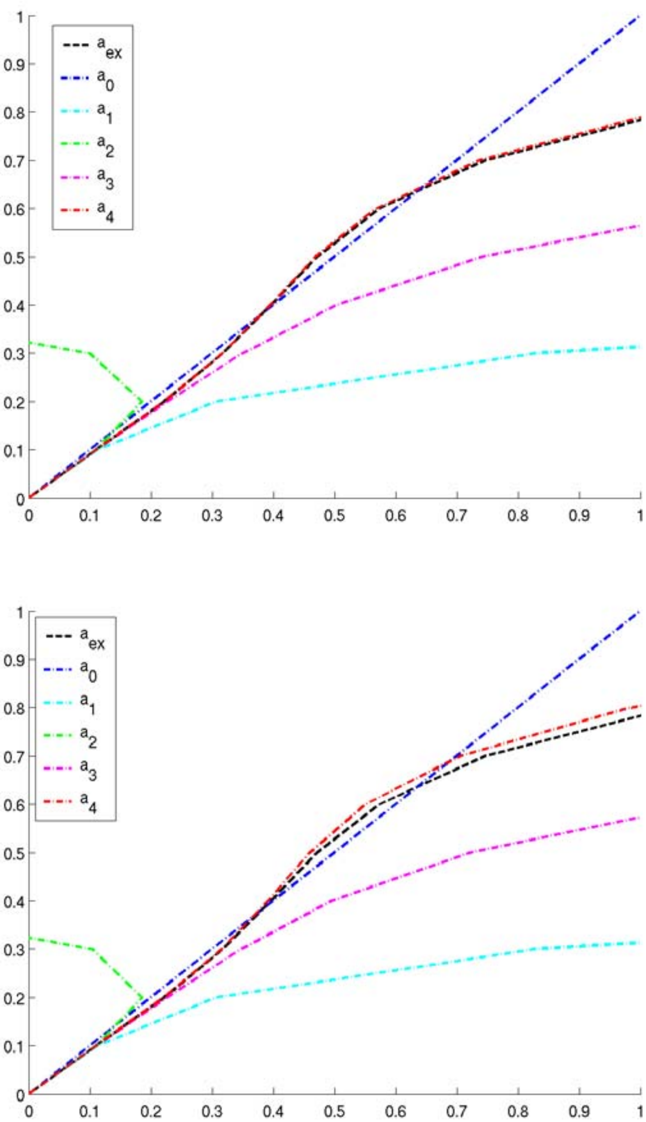
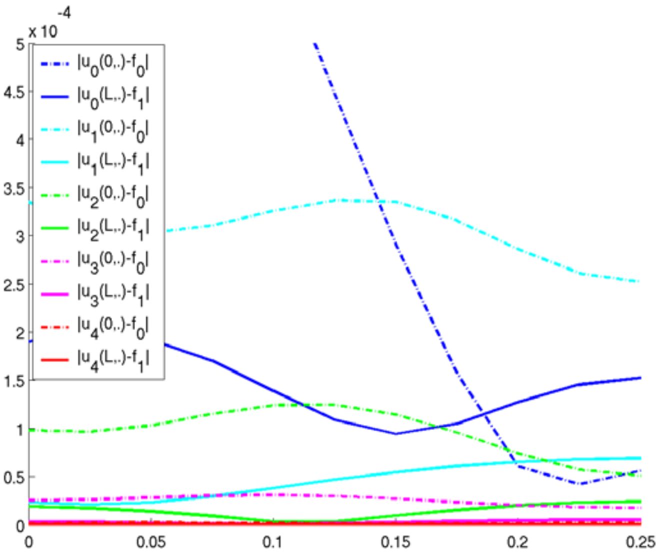


FIG. 6.4. *Starting value* a_0 *and Newton iterations* a_1, a_2, a_3, a_4 *of* a*, where* $a^{-1}(z) = k(z)z$*, compared to exact curve* a_{ex} *for the example in Figure* 6.1(c)*; top: exact data, bottom: noisy data (*0.1 *percent noise).*



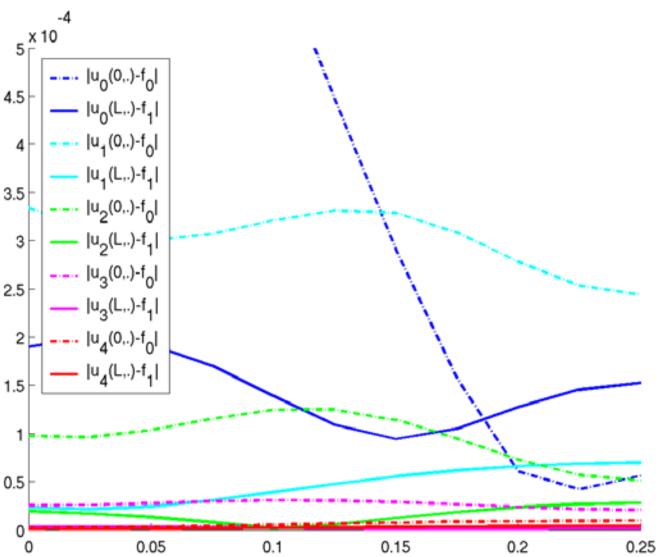


FIG. 6.5. *Starting value and Newton iterations of the residual corresponding to Figure* 6.4; *top: exact data, bottom: noisy data (*0.1 *percent noise).*

REFERENCES

[1] D. G. Aronson, *The role of diffusion in mathematical population biology: Skellam revisited*, in Mathematics in Biology and Medicine, S. Levine, ed., Lecture Notes in Biomath. 57, Springer, Berlin, 1985, pp. 2–6.

[2] F. Bachinger, U. Langer, and J Schöberl, *Numerical analysis of nonlinear multiharmonic eddy current problems*, Numer. Math., 100 (2005), pp. 593–616.

[3] M. Bellassoued, *Global logarithmic stability in inverse hyperbolic problem by arbitrary boundary observation*, Inverse Problems, 20 (2004), pp. 1033–1052.

[4] A. L. Buhgeim and M. V. Klibanov, *Uniqueness in the large of a class of multidimensional inverse problems*, Soviet Math. Dokl., 17 (1981), pp. 244–247.

[5] M. Cristofol, P. Gaitan, and H. Ramoul, *Inverse problems for a 2×2 reaction-diffusion system using a Carleman estimate with one observation*, Inverse Problems, 22 (2006), pp. 1561–1573.

[6] N. A. Demerdash and D. H. Gillot, *A new approach for determination of eddy current and flux penetration in nonlinear ferromagnetic materials*, IEEE Trans. Magnet., 10 (1974), pp. 682–685.

[7] A. Doubova and A. Osses, *Rotated weights in global Carleman estimates applied to an inverse problem for the wave equation*, Inverse Problems, 22 (2006), pp. 265–296.

[8] H. Egger, H. W. Engl, and M. V. Klibanov, *Global uniqueness and Holder stability for recovering a nonlinear source term in a parabolic equation*, Inverse Problems, 21 (2005), pp. 271–290.

[9] J. Gyselinck, P. Dular, C. Geuzaine, and W. Legros, *Harmonic-balance finite-element modeling of electromagnetic devices: A novel approach*, IEEE Trans. Magnet., 38 (2002), pp. 521–524.

[10] N. Ida and J. P. Bastos, *Electromagnetics and Calculation of Fields*, Springer, New York, 1992.

[11] O. Yu. Imanuvilov and M. Yamamoto, *Lipschitz stability in inverse parabolic problems by the Carleman estimate*, Inverse Problems, 14 (1998), pp. 1229–1245.

[12] O. Yu. Imanuvilov and M. Yamamoto, *Global uniqueness and stability in determining coefficients of wave equations*, Comm. Partial Differential Equations, 26 (2001), pp. 1409–1425.

[13] V. Isakov, *Uniqueness of recovery of some quasilinear partial differential equations*, Comm. Partial Differential Equations, 26 (2001), pp. 1947–1973.

[14] B. Kaltenbacher, *Towards global convergence for strongly nonlinear ill-posed problems via a regularizing multilevel approach*, Numer. Funct. Anal. Optim., 27 (2006), pp. 637–665.

[15] B. Kaltenbacher, *Determination of parameters in nonlinear hyperbolic PDEs via a multiharmonic formulation, used in piezoelectric material characterization*, Math. Models Methods Appl. Sci., 16 (2006), pp. 869–895.

[16] M. Kaltenbacher, *Numerical Simulation of Mechatronic Sensors and Actuators*, 2nd ed., Springer, Berlin, Heidelberg, New York, 2007.

[17] B. Kaltenbacher, M. Kaltenbacher, and S. Reitzinger, *Identification of nonlinear B-H curves based on magnetic field computations and using multigrid preconditioners for ill-posed problems*, European J. Appl. Math., 14 (2003), pp. 15–38.

[18] M. V. Klibanov, *A class of inverse problems for nonlinear parabolic equations*, Siberian Math. J., 27 (1987), pp. 697–707.

[19] M. V. Klibanov, *Inverse problems and Carleman estimates*, Inverse Problems, 8 (1992), pp. 575–596.

[20] M. V. Klibanov, *Global uniqueness for a multidimensional inverse problem for a nonlinear parabolic equation*, Inverse Problems, 20 (2004), pp. 1003–1032.

[21] M. V. Klibanov and A. Timonov, *Carleman Estimates for Coefficient Inverse Problems and Numerical Applications*, VSP, Utrecht, The Netherlands, 2004.

[22] M. M. Lavrent'ev, V. G. Romanov, and S. P. Shishatskii, *Ill-Posed Problems of Mathematical Physics and Analysis*, AMS, Providence, RI, 1986.

[23] S. Li, *An inverse problem for Maxwell's equations in bi-isotropic media*, SIAM J. Math. Anal., 37 (2005), pp. 1027–1043.

[24] M. Pilant and W. Rundell, *An inverse problem for a nonlinear parabolic equation*, Comm. Partial Differential Equations, 11 (1986), pp. 445–457.

[25] S. Yamada and K. Bessho, *Harmonic field calculation by the combination of finite element analysis and harmonic balance method*, IEEE Trans. Magnet., 24 (1988), pp. 2588–2590.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 1890–1920

LOW REGULARITY LOCAL WELL-POSEDNESS OF THE DERIVATIVE NONLINEAR SCHRÖDINGER EQUATION WITH PERIODIC INITIAL DATA*

AXEL GRÜNROCK† AND SEBASTIAN HERR‡

Abstract. The Cauchy problem for the derivative nonlinear Schrödinger equation with periodic boundary condition is considered. Local well-posedness for data u_0 in the space $\widehat{H}^s_r(\mathbb{T})$, defined by the norms $\|u_0\|_{\widehat{H}^s_r(\mathbb{T})} = \|\langle\xi\rangle^s \widehat{u}_0\|_{\ell^{r'}_\xi}$, is shown in the parameter range $s \geq \frac{1}{2}$, $2 > r > \frac{4}{3}$. The proof is based on an adaptation of the gauge transform to the periodic setting and an appropriate variant of the Fourier restriction norm method.

Key words. local well-posedness, derivative nonlinear Schrödinger equation, periodic functions, sharp multilinear estimates, gauge transformation, generalized Fourier restriction norm method

AMS subject classification. 35Q55

DOI. 10.1137/070689139

1. Introduction and main result. The Cauchy problem for the derivative nonlinear Schrödinger equation

$$\text{(DNLS)} \qquad \begin{aligned} i\partial_t u + \partial_x^2 u &= i\partial_x(|u|^2 u), \\ u(0,x) &= u_0(x) \end{aligned}$$

describes the propagation of nonlinear waves arising in plasma physics, nonlinear optics, and other disciplines. For initial data u_0 in the classical Sobolev spaces $H^s(\mathbb{R})$ of functions defined on the real line, (DNLS) is known to be locally well-posed for $s \geq \frac{1}{2}$. This was shown by Takaoka in [25], where he improved the earlier $H^1(\mathbb{R})$-result of Hayashi and Ozawa [16, 17] and Hayashi [15]. His method of proof combines the gauge transform already used by Hayashi and Ozawa with Bourgain's Fourier restriction norm method. Global well-posedness in $H^s(\mathbb{R})$ for $s > \frac{1}{2}$ is derived by Colliander et al. in [7, 8]. A counterexample of Biagioni and Linares [2] shows the optimality of Takaoka's local result on the $H^s(\mathbb{R})$-scale of data spaces: For $s < \frac{1}{2}$ the Cauchy problem (DNLS) is ill-posed in the C^0-uniform sense, although the standard scaling argument suggests only local ill-posedness for $s < 0$. This gap of $1/2$ derivative between the scaling prediction and Takaoka's result can be closed by leaving the $H^s(\mathbb{R})$-scale and considering data in the spaces $\widehat{H}^s_r(\mathbb{R})$ defined by the norms

$$\|u_0\|_{\widehat{H}^s_r(\mathbb{R})} = \|\langle\xi\rangle^s \widehat{u}_0\|_{L^{r'}_\xi}, \qquad \langle\xi\rangle = (1+\xi^2)^{\frac{1}{2}}, \qquad \frac{1}{r} + \frac{1}{r'} = 1.$$

We remark that these spaces coincide with $B_{r',k}$ (with weight $k(\xi) = \langle\xi\rangle^s$), introduced by Hörmander; cf. [20, section 10.1]. The idea to consider them as data spaces for nonlinear Schrödinger equations goes back to the work of Cazenave, Vega, and

*Received by the editors April 23, 2007; accepted for publication (in revised form) November 6, 2007; published electronically February 22, 2008.
http://www.siam.org/journals/sima/39-6/68913.html

†Bergische Universität Wuppertal, Fachbereich C: Mathematik/Naturwissenschaften, Gaußstraße 20, 42097 Wuppertal, Germany (axel.gruenrock@math.uni-wuppertal.de).
‡Rheinische Friedrich-Wilhelms-Universität Bonn, Mathematisches Institut, Beringstraße 1, 53115 Bonn, Germany (herr@math.uni-bonn.de). This author's research was initially supported by German Research Foundation (DFG) grant KO 1307/5-3.

Vilela [4], where corresponding weak norms are used. Yet another alternative class of data spaces has been considered by Vargas and Vega in [26].

Concerning the (DNLS) equation on the real line, it was shown by the first author in [13] that local well-posedness holds for data in $\widehat{H}^s_r(\mathbb{R})$, provided $s \geq \frac{1}{2}$ and $2 \geq r > 1$. This generalization of Takaoka's result almost reaches the case $(s,r) = (\frac{1}{2}, 1)$, which is critical with respect to the scaling symmetry. The proof uses the gauge transform again and an appropriate variant of the Fourier restriction norm method, which was developed in [12]. Furthermore, it relies heavily on certain smoothing properties of the Schrödinger group, expressed in terms of bi- and trilinear estimates for free solutions.

On the other hand, it could be shown by the second author in [18] that Takaoka's result concerning the real line can be carried over to the periodic case *with the same lower bound* $s \geq \frac{1}{2}$ on the Sobolev regularity. This is remarkable, since there is a number of nonlinear Schrödinger and Korteweg–de Vries-type equations, which are—due to a lack of smoothing properties—strictly worse behaved in the periodic setting than in the continuous case. To prove the result concerning the one-dimensional torus, the gauge transform had to be adjusted to the periodic case; see [18, section 2]. The transformed equation is then treated by the Fourier restriction norm method. Here, the L^4 Strichartz estimate [27, 3] turned out to be a central tool in the derivation of the nonlinear estimates.

Now it is natural to ask for a synthesis of the two last-mentioned results, i.e., to consider the Cauchy problem (DNLS) with u_0 in the following two parameter scale of data spaces.

DEFINITION 1.1. *Let $s \in \mathbb{R}$, $1 \leq r \leq \infty$, and $\frac{1}{r} + \frac{1}{r'} = 1$. Define $\widehat{H}^s_r(\mathbb{T})$ as the completion of all trigonometric polynomials with respect to the norm*

$$\|f\|_{\widehat{H}^s_r} := \|\widehat{J^s f}\|_{\ell^{r'}_\xi}, \tag{1}$$

where J^s is the Bessel potential operator of order $-s$ given by $\widehat{J^s f}(\xi) = \langle \xi \rangle^s \widehat{f}(\xi)$.

Remark 1. In contrast to the nonperiodic case the continuous embedding $\widehat{H}^s_q(\mathbb{T}) \subset \widehat{H}^s_r(\mathbb{T})$ holds true for any $1 \leq r \leq q \leq \infty$. Moreover, we have $H^s_2(\mathbb{T}) = \widehat{H}^s_2(\mathbb{T})$ and, more generally, $H^s_r(\mathbb{T}) \subset \widehat{H}^s_r(\mathbb{T})$ for $1 \leq r \leq 2$ by the Hausdorff–Young inequality, where $H^s_r(\mathbb{T})$ denotes the Bessel potential space of all u such that $J^s u \in L^r(\mathbb{T})$. If $r = 2$, we will usually omit the index r.

The main result of this paper is local well-posedness of (DNLS) in these data spaces in the parameter range[1] $s \geq \frac{1}{2}$ and $2 > r > \frac{4}{3}$. More precisely, the following theorem will be shown.

THEOREM 1.2. *Let $\frac{4}{3} < q \leq r \leq 2$. For every*

$$u_0 \in B_R := \left\{ u_0 \in \widehat{H}^{\frac{1}{2}}_r(\mathbb{T}) \mid \|u_0\|_{\widehat{H}^{\frac{1}{2}}_q} < R \right\}$$

and $T \lesssim R^{-2q'-}$ there exists a solution $u \in C([-T,T], \widehat{H}^{\frac{1}{2}}_r(\mathbb{T}))$ of the Cauchy problem (DNLS). *This solution is the unique limit of smooth solutions, and the map*

$$\left(B_R, \|\cdot\|_{\widehat{H}^{\frac{1}{2}}_r}\right) \longrightarrow C([-T,T], \widehat{H}^{\frac{1}{2}}_r(\mathbb{T})): \quad u_0 \mapsto u$$

[1] For the sake of clearness of the exposition we state the main results only for the most interesting case $s = \frac{1}{2}$, but standard arguments also provide persistence of higher regularity.

is continuous but not locally uniformly continuous. However, on subsets of B_R with fixed L^2-norm it is locally Lipschitz continuous.

Remark 2.

1. The uniqueness statement in the theorem above can be sharpened; see Remark 5.

2. Our methods rely on the L^2 conservation law, but not on the complete integrability of (DNLS) (see [21]), and also apply to nonlinearities with (say) additional polynomial terms of type $|u|^k u$.

3. Solution always means solution of the corresponding integral equation

$$u(t) = e^{it\partial_x^2} u_0 + i \int_0^t e^{i(t-t')\partial_x^2} \partial_x (|u|^2 u)(t') dt', \quad t \in (-T, T).$$

4. On any level of regularity the flow map is not uniformly continuous. This is precisely due to the influence of a translation; see Lemma 6.3. However, our arguments imply that the map $u_0 \mapsto u(t, x - \frac{t}{\pi}\|u\|_{L^2}^2)$ is real analytic for any $s \geq \frac{1}{2}$; see also [19, Theorem 3.1.5]. In view of the counterexamples in [18, Theorem 5.3], and [19, Theorem 3.1.5] for the corresponding modified equation, which are essentially of the same kind as the one already given in [25, Proposition 3.3], we cannot expect any positive result with smooth dependence (modulo the translation specified above) for $s < \frac{1}{2}$. Observe that the examples concerning the periodic case are monochromatic waves and thus do not distinguish between an ℓ^2_ξ- and an $\ell^{r'}_\xi$-norm. Concerning the second parameter r, we must leave open the question whether or not there is local well-posedness for $r \leq \frac{4}{3}$. Nonetheless, we will show below that our result is optimal within the framework we use.

Before we turn to details, let us point out that in the periodic case almost nothing is known about Cauchy problems with data in the $\widehat{H}^s_r(\mathbb{T})$ spaces. The only result we are aware of is due to Christ [6, 5], who considers the following modification of the cubic nonlinear Schrödinger equation on the one-dimensional torus:

$$\text{(NLS*)} \qquad i\partial_t u + \partial_x^2 u = \left(|u|^2 - 2 \fint_0^{2\pi} |u|^2 dx \right) u,$$

with initial condition $u(0) = u_0 \in \widehat{H}^s_r(\mathbb{T})$. He shows that for $s \geq 0$ and $r > 1$ the solution map

$$S : H^\sigma(\mathbb{T}) \longrightarrow C([0,\infty), H^\sigma(\mathbb{T})) \cap C^1([0,\infty), H^{\sigma-2}(\mathbb{T}))$$

(σ sufficiently large) "extends by continuity to a uniformly continuous mapping from the ball centered at 0 of [arbitrary] radius R in" $\widehat{H}^s_r(\mathbb{T})$ to $C([0,\tau], \widehat{H}^s_r(\mathbb{T}))$, where τ depends on R; see [6, Theorem 6.1.1].[2] This result is shown by a *new method of solution*, which is developed in [6]; a summary of this method is given in section 6.1.5 of that paper. The positive result in [6] is supplemented in [5] by a statement of nonuniqueness: For $2 > r > 1$ there exists a nonvanishing weak solution[3] $u \in C([0,1], \widehat{H}^0_r(\mathbb{T}))$ of (NLS*) with initial value $u_0 \equiv 0$; see [5, Theorem 2.3].

Using the function spaces $X^{s,b}_r$, defined by the norms

$$\|u\|_{X^{s,b}_r} = \|\langle \xi \rangle^s \langle \tau + \xi^2 \rangle^b \widehat{u}\|_{\ell^{r'}_\xi L^{r'}_\tau},$$

[2] In [6] the data spaces are denoted as $\mathcal{F}L^{s,p}(\mathbb{T})$, which corresponds to $\widehat{H}^s_{p'}(\mathbb{T})$ in our terms.

[3] A precise definition of a weak solution is given in [5, section 2.1].

where $\frac{1}{r} + \frac{1}{r'} = 1$ (cf. [12, section 2]), we can actually show local *well-posedness* of the initial value problem associated with (NLS*) with data $u_0 \in \widehat{H}^0_r(\mathbb{T})$, $2 > r > 1$, thus giving an alternative proof (based on the contraction mapping principle) of Christ's result from [6]. The argument also provides uniqueness of the solution in the restriction norm space based on $X^{0,b}_r$; see [12, formulas (2.37) and (2.38)]. It was the starting point for our investigations concerning the (DNLS) equation and already exhibits some of the main arguments; so let us sketch this proof.

We define the trilinear operator C_1 by its partial Fourier transform (in the space variable only)

$$\widehat{C_1(u_1, u_2, u_3)}(\xi) = (2\pi)^{-1} \sum_{\substack{\xi = \xi_1 + \xi_2 + \xi_3 \\ \xi_1 \neq \xi,\, \xi_2 \neq \xi}} \widehat{u}_1(\xi_1)\widehat{u}_2(\xi_2)\widehat{\overline{u}}_3(\xi_3),$$

so that the (partial) Fourier transform of the nonlinearity in (NLS*) becomes

$$\widehat{C_1(u, u, u)}(\xi) - (2\pi)^{-1}\widehat{u}^2(\xi)\widehat{\overline{u}}(-\xi).$$

By [12, Theorem 2.3] it is sufficient to estimate the latter appropriately in $X^{0,b}_r$-norms. Here, the second contribution turns out to be harmless; cf. the end of the proof of Theorem 2.4 below. So matters essentially reduce to show the following estimate.

PROPOSITION 1.3. *Let $r > 1$, $\varepsilon > 0$, and $b > \frac{1}{r}$. Then*

$$\|C_1(u_1, u_2, u_3)\|_{X^{0,-\varepsilon}_r} \lesssim \prod_{i=1}^{3} \|u_i\|_{X^{0,b}_r}.$$

Proof. Choosing $f_i \in \ell^{r'}_\xi L^{r'}_\tau$ such that $\|f_i\|_{\ell^{r'}_\xi L^{r'}_\tau} = \|u_i\|_{X^{0,b}_r}$ the above estimate can be rewritten as

$$\left\| \langle\sigma_0\rangle^{-\varepsilon} \sum_{\substack{\xi = \xi_1 + \xi_2 + \xi_3 \\ \xi_1 \neq \xi,\, \xi_2 \neq \xi}} \int_{\tau = \tau_1 + \tau_2 + \tau_3} \prod_{i=1}^{3} \frac{f_i(\xi_i, \tau_i)}{\langle\sigma_i\rangle^b} \, d\tau_1 d\tau_2 \right\|_{\ell^{r'}_\xi L^{r'}_\tau} \lesssim \prod_{i=1}^{3} \|f_i\|_{\ell^{r'}_\xi L^{r'}_\tau}, \tag{2}$$

where $\sigma_0 = \tau + \xi^2$, $\sigma_i = \tau_i + \xi_i^2$ $(i = 1, 2)$, and $\sigma_3 = \tau_3 - \xi_3^2$. By Hölder's inequality and Fubini's theorem, (2) can be deduced from

$$\sup_{\xi,\tau} \langle\sigma_0\rangle^{-r\varepsilon} \sum_{\substack{\xi = \xi_1 + \xi_2 + \xi_3 \\ \xi_1 \neq \xi,\, \xi_2 \neq \xi}} \int_{\tau = \tau_1 + \tau_2 + \tau_3} \prod_{i=1}^{3} \langle\sigma_i\rangle^{-rb} \, d\tau_1 d\tau_2 < \infty. \tag{3}$$

Using the resonance relation

$$2|\xi_1\xi_2 + \xi\xi_3| = 2|\xi - \xi_1||\xi - \xi_2| \leq \sum_{i=0}^{3} \langle\sigma_i\rangle \leq \prod_{i=0}^{3} \langle\sigma_i\rangle \tag{4}$$

the left-hand side of (3) is bounded by

$$\begin{aligned} &\sum_{\substack{\xi = \xi_1 + \xi_2 + \xi_3 \\ \xi_1 \neq \xi,\, \xi_2 \neq \xi}} \langle\xi - \xi_1\rangle^{0-}\langle\xi - \xi_2\rangle^{0-} \int_{\tau = \tau_1 + \tau_2 + \tau_3} d\tau_1 d\tau_2 \prod_{i=1}^{3} \langle\sigma_i\rangle^{-1-} \\ \lesssim &\sum_{\substack{\xi = \xi_1 + \xi_2 + \xi_3 \\ \xi_1 \neq \xi,\, \xi_2 \neq \xi}} \langle\xi - \xi_1\rangle^{0-}\langle\xi - \xi_2\rangle^{0-}\langle\tau + \xi^2 - 2(\xi - \xi_1)(\xi - \xi_2)\rangle^{-1-}, \end{aligned}$$

where in the last step we have used Lemma 4.1 twice. Setting $\mathbb{Z}^* = \mathbb{Z}\setminus\{0\}$, $n_i = \xi - \xi_i$ for $i = 1, 2$, and $r = n_1 n_2$ the last sum can be rewritten as

$$\sum_{r\in\mathbb{Z}^*} \langle \tau + \xi^2 - 2r\rangle^{-1-} \langle r\rangle^{0-} \sum_{\substack{n_1,n_2\in\mathbb{Z}^*\\ r=n_1n_2}} 1,$$

which is bounded by a constant independent of ξ and τ, since the number of divisors of $r \in \mathbb{N}$ can be estimated by $c_\varepsilon r^\varepsilon$ for any positive ε. □

Three aspects of the preceding are worth being emphasized in view of our investigations here.

Necessity of cancellations and correction terms. For $r < 2$ the above argument breaks down completely without the restrictions $\xi \neq \xi_1$ and $\xi \neq \xi_2$ in the sum over the Fourier coefficients. As was pointed out already by Christ, this cancellation comes from the correction term $2 \fint_0^{2\pi} |u|^2 dx u$ subtracted in the nonlinearity, and for any other coefficient in front of this term one cannot obtain continuous dependence; see [6, last sentence of section 6.1.3 and the remark before (6.2.6)]. A very similar cancellation turns out to be fundamental in our analysis of the (DNLS) equation, but here *the corresponding correction term comes from the gauge transform* in its periodic variant, which is discussed in Remark 4 in section 6. In fact, the main contribution to the cubic part of the transformed equation is given by $T^*(u, u, u)$, where

$$T^*(\widehat{u_1, u_2}, u_3)(\xi) = (2\pi)^{-1} \sum_{\substack{\xi=\xi_1+\xi_2+\xi_3\\ \xi_1\neq\xi,\, \xi_2\neq\xi}} \widehat{u}_1(\xi_1)\widehat{u}_2(\xi_2) i\xi_3 \widehat{\overline{u}}_3(\xi_3).$$

Again, our argument would not work without the restrictions $\xi \neq \xi_1$, $\xi \neq \xi_2$.

Modification of the norms. If we try to estimate the term $T^*(u_1, u_2, u_3)$ in an $X_r^{s,b}$-norm in a similar manner as in the proof of Proposition 1.3, we have to get control over a whole derivative, that is, on the Fourier side, over the factor ξ_3. The complete absence of smoothing effects (gaining derivatives) in the periodic case forces us to get this control from the resonance relation (4) only, which is the same for (DNLS) as for (NLS*). This means that we have to choose the b-parameters equal to $-\frac{1}{2}$ on the left-hand side and to $+\frac{1}{2}$ on the right-hand side of the estimate. Now, the necessity to cancel the ξ_3-factor and the resonance relation lead to the consideration of eight cases—some of them being symmetric—depending on which of the σ's is maximal and on whether or not $|\xi\xi_3| \lesssim |\xi_1\xi_2|$; see the table in the proof of Theorem 2.4 below. Picking out the (relatively harmless) subcase, where $|\xi\xi_3| \lesssim |\xi_1\xi_2|$ and σ_0 is maximal, so that $\prod_{i=1}^3 \langle\sigma_i\rangle^{\frac{1}{6}} \leq \langle\sigma_0\rangle^{\frac{1}{2}}$, we are in the situation of the above proof, with a half derivative on each factor (as desired) but with a b-parameter on the right of at most $\frac{2}{3}$, which means that we end up with the nonoptimal restriction $r > \frac{3}{2}$. This leads us to *introduce a fourth parameter* p in the $X_r^{s,b}$-norms, which is the Hölder exponent concerning the τ-integration and may differ from r; see Definition 2.1 below. In our application here we choose $p = 2$, thus going back to some extent to the meanwhile classical $X^{s,b}$-spaces.

Number of divisor estimate. The number of divisor argument at the end of the proof of Proposition 1.3 has already been used in Christ's work and can be seen as a substitute for Bourgain's L^6 Strichartz estimate for the periodic case, which itself was shown by the aid of this argument; see [3, Proposition 2.36]. We will need a refined version thereof, which is shown by elementary geometric considerations in section 3. Here, we use arguments similar to those of De Silva et al. [10, section 4].

Concerning the organization of the paper the following should be added: In section 2 we introduce the relevant function spaces and state all the nonlinear estimates needed as well as a sharpness result. The crucial trilinear estimates and a counterexample are derived in section 4, which is very much in the spirit of [22]. Section 5 deals with the quintilinear estimate. In both cases we have made some effort to extract the correct lifespan from the nonlinear estimates and to obtain persistence of higher regularity.[4] By this we mean that the lifespan of a solution with $\widehat{H}^{\frac{1}{2}}_r(\mathbb{T})$-data depends only on the smaller $\widehat{H}^{\frac{1}{2}}_q(\mathbb{T})$-norm of the initial value, where $2 \geq r > q > \frac{4}{3}$. Finally, in section 7, the contraction mapping principle is invoked to prove local well-posedness for the transformed equation (49); see Theorem 7.2. Our main result, Theorem 1.2, is then a consequence of Lemma 6.4 on the gauge transform.

We close this section by fixing some notational conventions.

- The Fourier transform with respect to the space variable (periodic) is

$$\mathcal{F}_x f(\xi) = \widehat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} f(x) e^{-ix\xi} dx \quad (f \in L^1(\mathbb{T})).$$

- The Fourier transform with respect to the time variable (nonperiodic) is

$$\mathcal{F}_t f(\tau) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(t) e^{-it\tau} dt \quad (f \in L^1(\mathbb{R})).$$

- The Fourier transform with respect to time and space variables is $\mathcal{F} = \mathcal{F}_t \mathcal{F}_x$.
- For the mean value integral we write

$$⨍_0^{2\pi} f(x) dx = \frac{1}{2\pi} \int_0^{2\pi} f(x) dx \quad (f \in L^1(\mathbb{T})).$$

- Let $a \in \mathbb{R}$. The expressions $a\pm$ denote numbers $a \pm \varepsilon$ for sufficiently small $\varepsilon > 0$ (in the case $a = \infty$ the expression $a-$ denotes a sufficiently large real number). The implicit constraint on ε may involve the corresponding expressions in the same line and the lines before in a natural way, and eventually ε is chosen small enough such that it also fulfills the constraints imposed in what follows.
- For a given set of parameters (typically a subset of $\varepsilon, \delta, \nu, p, q, r, s$) the statement $A \lesssim B$ means that there exists a constant $C > 0$ which depends only on these parameters such that $A \leq CB$. This is equivalent to $B \gtrsim A$. We may write $A \ll B$ if it is possible to choose $0 < C < \frac{1}{4}$.
- For all parameters $1 \leq p \leq \infty$ the number $1 \leq p' \leq \infty$ is defined to be the dual parameter satisfying $\frac{1}{p} + \frac{1}{p'} = 1$.

2. Function spaces and main estimates. Let $\mathcal{S}(\mathbb{R} \times \mathbb{T})$ be the linear space of all C^∞-functions $f : \mathbb{R}^2 \to \mathbb{C}$ such that

$$f(t,x) = f(t, x + 2\pi), \qquad \sup_{(t,x)\in\mathbb{R}^2} |t^\alpha \partial_t^\beta \partial_x^\gamma f(t,x)| < \infty, \quad \alpha, \beta, \gamma \in \mathbb{N}_0.$$

[4]We refer the reader to [24, Theorem 2, part V] for the corresponding notion if data in the H^s-scale are considered. In the H^s-case this property usually is a simple consequence of the convolution constraint; see again [24, Remark 2 below Theorem 3]. We cannot see that a similar argument should work in our setting.

DEFINITION 2.1. *Let $s, b \in \mathbb{R}$, $1 \le r, p \le \infty$, and $\frac{1}{r} + \frac{1}{r'} = 1 = \frac{1}{p} + \frac{1}{p'}$. Define the space $X^{s,b}_{r,p}$ as the completion of $\mathcal{S}(\mathbb{R} \times \mathbb{T})$ with respect to the norm*

$$\|u\|_{X^{s,b}_{r,p}} = \|\langle \tau + \xi^2 \rangle^b \langle \xi \rangle^s \mathcal{F} u\|_{\ell^{r'}_\xi L^{p'}_\tau}. \tag{5}$$

In the case where $r = p = 2$ we write $X^{s,b}_{r,p} = X^{s,b}$ as usual.

LEMMA 2.2. *Let $s, b_1, b_2 \in \mathbb{R}$, $1 \le r \le \infty$, and $b_1 > b_2 + \frac{1}{2}$. The following embeddings are continuous:*

$$X^{s,b_1}_{r,2} \subset X^{s,b_2}_{r,\infty}, \tag{6}$$

$$X^{s,0}_{r,\infty} \subset C(\mathbb{R}, \widehat{H}^s_r(\mathbb{T})). \tag{7}$$

Proof. The first embedding is proved by the Cauchy–Schwarz inequality with respect to the L^1_τ-norm.

The second embedding follows from $\mathcal{F}_t^{-1} L^1(\mathbb{R}) \subset L^\infty(\mathbb{R})$. ◻

DEFINITION 2.3. *Let $s \in \mathbb{R}$ and $1 \le r \le \infty$. We define*

$$Z^s_r := X^{s,\frac{1}{2}}_{r,2} \cap X^{s,0}_{r,\infty}$$

and for $0 < T \le 1$ the restriction space $Z^s_r(T)$ of all $v = w\,|_{[-T,T]}$ for some $w \in Z^s_r$ with norm

$$\|v\|_{Z^{\frac{1}{2}}_r(T)} := \inf \left\{ \|w\|_{Z^{\frac{1}{2}}_r} | w \in Z^{\frac{1}{2}}_r : w\,|_{[-T,T]} = v \right\}.$$

A main ingredient for the proof of Theorem 1.2 is an estimate on the trilinear operator (suppressing the t dependence):

$$\begin{aligned} \widehat{T(u_1, u_2, u_3)}(\xi) = (2\pi)^{-1} &\sum_{\substack{\xi = \xi_1 + \xi_2 + \xi_3 \\ \xi_1 \ne \xi,\, \xi_2 \ne \xi}} \widehat{u}_1(\xi_1) \widehat{u}_2(\xi_2) i\xi_3 \widehat{\overline{u}}_3(\xi_3) \\ &+ (2\pi)^{-1} \widehat{u}_1(\xi) \widehat{u}_2(\xi) i\xi \widehat{\overline{u}}_3(-\xi). \end{aligned} \tag{8}$$

THEOREM 2.4. *Let $\frac{4}{3} < q \le r \le 2$ and $0 \le \delta < \frac{1}{q'}$. Then*

$$\|T(u_1, u_2, u_3)\|_{X^{\frac{1}{2},-\frac{1}{2}}_{r,2}} \lesssim T^\delta \|u_1\|_{X^{\frac{1}{2},\frac{1}{2}}_{q,2}} \|u_2\|_{X^{\frac{1}{2},\frac{1}{2}}_{q,2}} \|u_3\|_{X^{\frac{1}{2},\frac{1}{2}}_{r,2}} \tag{9}$$

if $\operatorname{supp}(u_i) \subset \{(t, x) \mid |t| \le T\}$, $0 < T \le 1$.

Additionally, we will need the following estimate on $T(u_1, u_2, u_3)$.

THEOREM 2.5. *Let $\frac{4}{3} < q \le r \le 2$ and $0 \le \delta < \frac{1}{q'}$. Then*

$$\|T(u_1, u_2, u_3)\|_{X^{\frac{1}{2},-1}_{r,\infty}} \lesssim T^\delta \|u_1\|_{X^{\frac{1}{2},\frac{1}{2}}_{q,2}} \|u_2\|_{X^{\frac{1}{2},\frac{1}{2}}_{q,2}} \|u_3\|_{X^{\frac{1}{2},\frac{1}{2}}_{r,2}} \tag{10}$$

if $\operatorname{supp}(u_i) \subset \{(t, x) \mid |t| \le T\}$, $0 < T \le 1$.

The above estimates are sharp with respect to the lower threshold on r within the full scale of spaces $X^{\frac{1}{2},\frac{1}{2}}_{r,p}$. Note that in particular the estimates fail to hold in the endpoint case $r = \frac{4}{3}$.

Remark 3. For all $b \le 0$, $1 \le r \le \frac{4}{3}$, and $1 \le p, q \le \infty$ the estimate

$$\|T(u_1, u_2, u_3)\|_{X^{\frac{1}{2},b}_{r,p}} \lesssim \prod_{i=1}^{3} \|u_i\|_{X^{\frac{1}{2},\frac{1}{2}}_{r,q}} \tag{11}$$

is false.

We also consider the quintilinear expression defined as

$$\widehat{Q(u_1,\dots,u_5)}(\xi) = (2\pi)^{-2} \sum_{**} \widehat{u_1}(\xi_1)\widehat{\overline{u_2}}(\xi_2)\widehat{u_3}(\xi_3)\widehat{\overline{u_4}}(\xi_4)\widehat{u_5}(\xi_5), \tag{12}$$

where we suppressed the t dependence and $**$ is shorthand for summation over the subset of $\mathbb{Z}^5$ given by the restrictions

$$\xi = \xi_1 + \cdots + \xi_5; \quad \xi_1 + \cdots + \xi_4 \neq 0; \quad \xi_1 + \xi_2 \neq 0; \quad \xi_3 + \xi_4 \neq 0.$$

THEOREM 2.6. *Let* $\frac{4}{3} < q \leq r \leq 2$ *and* $b > \frac{1}{6} + \frac{1}{3q}$. *Then*

$$\|u_1\overline{u}_2u_3\overline{u}_4u_5\|_{X^{\frac{1}{2},-b}_{r,2}} \lesssim \sum_{k=1}^{5} \|u_k\|_{X^{\frac{1}{2},b}_{r,2}} \prod_{\substack{1\leq i\leq 5\\ i\neq k}} \|u_i\|_{X^{\frac{1}{2},b}_{q,2}}. \tag{13}$$

Additionally assume that for $0 < T \leq 1$ *we have* $\operatorname{supp}(u_i) \subset \{(t,x) \mid |t| \leq T\}$ *and* $0 \leq \delta < \frac{2}{q'}$. *Then*

$$\|u_1\overline{u}_2u_3\overline{u}_4u_5\|_{X^{\frac{1}{2},-\frac{1}{2}}_{r,2} \cap X^{\frac{1}{2},-1}_{r,\infty}} \lesssim T^{\delta} \sum_{k=1}^{5} \|u_k\|_{X^{\frac{1}{2},\frac{1}{2}}_{r,2}} \prod_{\substack{1\leq i\leq 5\\ i\neq k}} \|u_i\|_{X^{\frac{1}{2},\frac{1}{2}}_{q,2}} \tag{14}$$

and

$$\|Q(u_1,u_2,u_3,u_4,u_5)\|_{X^{\frac{1}{2},-\frac{1}{2}}_{r,2} \cap X^{\frac{1}{2},-1}_{r,\infty}} \lesssim T^{\delta} \sum_{k=1}^{5} \|u_k\|_{X^{\frac{1}{2},\frac{1}{2}}_{r,2}} \prod_{\substack{1\leq i\leq 5\\ i\neq k}} \|u_i\|_{X^{\frac{1}{2},\frac{1}{2}}_{q,2}}. \tag{15}$$

3. Number of divisor estimates and consequences. The next lemma contains estimates on the number of divisors of a given natural number r. Part 1 is well known; see Hardy and Wright [14, Theorem 315]. The approach used to prove part 2 of Lemma 3.1 is motivated by [10, Lemma 4.4].

LEMMA 3.1.

1. *Let* $\varepsilon > 0$. *There exists* $c_\varepsilon > 0$ *such that for all* $r \in \mathbb{N}$

$$\#\left\{(n_1,n_2) \in \mathbb{N}^2 \mid n_1n_2 = r\right\} \leq c_\varepsilon r^{\varepsilon}. \tag{16}$$

2. *For all* $r \in \mathbb{N}$

$$\#\left\{(n_1,n_2) \in \mathbb{N}^2 \mid n_1n_2 = r, \quad 3|n_1 - n_2| \leq r^{\frac{1}{6}}\right\} \leq 2. \tag{17}$$

Proof of part 2. Let $r \in \mathbb{N}$. Assume that there are three lattice points contained in the above set. Then these points form a triangle of area $\mu \geq \frac{1}{2}$; see Figure 1. This triangle is located

(i) in the strip

$$S = \left\{(x_1,x_2) \in \mathbb{R}^2 \mid \sqrt{r} - \delta \leq x_1 \leq \sqrt{r} + \delta\right\},$$

where $\delta = \frac{1}{3}r^{\frac{1}{6}}$, because $|n_1 - \sqrt{r}| \leq |n_1 - n_2| \leq \delta$;

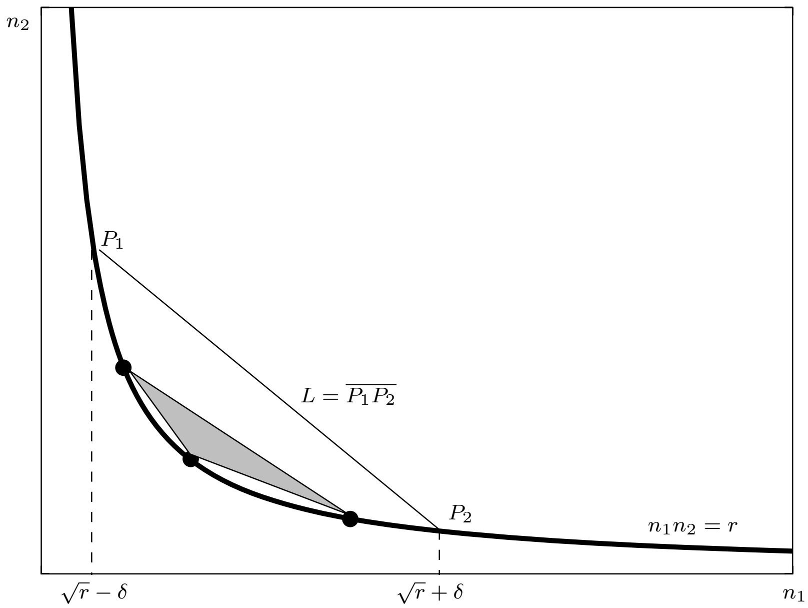


FIG. 1. *The notional triangle from the proof of Lemma* 3.1.

(ii) below the line $L = \overline{P_1P_2}$ which connects the points

$$P_1 = \left(\sqrt{r}-\delta, \frac{r}{\sqrt{r}-\delta}\right), \qquad P_2 = \left(\sqrt{r}+\delta, \frac{r}{\sqrt{r}+\delta}\right);$$

(iii) above the hyperbola

$$H = \left\{(x_1, x_2) \in (0,\infty)^2 \mid x_1x_2 = r\right\}$$

because the function $x \mapsto \frac{r}{x}$ is convex for $x > 0$.

Now, the area μ of the triangle is bounded from above by the area of the region in the strip S below L and above H; hence

$$\begin{aligned}
\mu &\le r\delta\left(\frac{1}{\sqrt{r}-\delta}+\frac{1}{\sqrt{r}+\delta}\right) - \int_{\sqrt{r}-\delta}^{\sqrt{r}+\delta} \frac{r}{x}\,dx \\
&= r\left(\frac{2\sqrt{r}\delta}{r-\delta^2} - \ln\left(1+\frac{2\delta}{\sqrt{r}-\delta}\right)\right) \\
&\le r\left(\frac{2\sqrt{r}\delta}{r-\delta^2} - \frac{2\delta}{\sqrt{r}-\delta} + \frac{2\delta^2}{(\sqrt{r}-\delta)^2}\right) = \frac{4r\delta^3}{(\sqrt{r}-\delta)^2(\sqrt{r}+\delta)} \\
&\le \frac{4\sqrt{r}\delta^3}{(\sqrt{r}-\delta)^2} \le \frac{4}{27}\frac{r}{(\sqrt{r}-\delta)^2} \le \frac{1}{3},
\end{aligned}$$

which contradicts $\mu \ge \frac{1}{2}$. $\square$

Now, we use Lemma 3.1 to prove the following.

COROLLARY 3.2. *Fix* $\varepsilon > 0$.

1. *There exists* $C_\varepsilon > 0$ *such that for all* $\xi \in \mathbb{Z}$ *and* $a \in \mathbb{R}$

$$\sum_{\substack{\xi_1,\xi_2\in\mathbb{Z} \\ \xi_1,\xi_2\neq\xi}} \langle \xi-\xi_1\rangle^{-\varepsilon}\langle \xi-\xi_2\rangle^{-\varepsilon}\langle a+2(\xi-\xi_1)(\xi-\xi_2)\rangle^{-1-\varepsilon} \le C_\varepsilon. \tag{18}$$

2. *There exists $C_\varepsilon > 0$ such that for all $\xi_1 \in \mathbb{Z}$ and $a \in \mathbb{R}$*

$$\sum_{\substack{\xi,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \langle \xi-\xi_1\rangle^{-\varepsilon}\langle \xi-\xi_2\rangle^{-\varepsilon}\langle a+2(\xi-\xi_1)(\xi-\xi_2)\rangle^{-1-\varepsilon} \le C_\varepsilon. \tag{19}$$

3. *There exists $C_\varepsilon > 0$ such that for all $\xi \in \mathbb{Z}$ and $a \in \mathbb{R}$*

$$\sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \langle \xi_1\rangle^{-\varepsilon}\langle \xi_2\rangle^{-\varepsilon}\langle a+2(\xi-\xi_1)(\xi-\xi_2)\rangle^{-1-\varepsilon} \le C_\varepsilon. \tag{20}$$

4. *There exists $C_\varepsilon > 0$ such that for all $\xi_1 \in \mathbb{Z}$ and $a \in \mathbb{R}$*

$$\langle \xi_1\rangle^{-\varepsilon}\sum_{\substack{\xi,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \langle \xi_2\rangle^{-\varepsilon}\langle a+2(\xi-\xi_1)(\xi-\xi_2)\rangle^{-1-\varepsilon} \le C_\varepsilon. \tag{21}$$

Proof. The first and second part follow from the standard number of divisors estimate (16) as follows: By the change of variables $n_1 = \xi-\xi_1$, $n_2 = \xi-\xi_2$ the sums in (18) and (19) are equal to

$$\sum_{n_1,n_2\in\mathbb{Z}^*} \langle n_1\rangle^{-\varepsilon}\langle n_2\rangle^{-\varepsilon}\langle a+2n_1n_2\rangle^{-1-\varepsilon}.$$

This can be written as

$$\begin{aligned}
&\sum_{r\in\mathbb{Z}^*}\sum_{\substack{n_1,n_2\in\mathbb{Z}^*\\ n_1n_2=r}} \langle n_1\rangle^{-\varepsilon}\langle n_2\rangle^{-\varepsilon}\langle a+2n_1n_2\rangle^{-1-\varepsilon}\\
&\le \sum_{r\in\mathbb{Z}^*}\langle a+2r\rangle^{-1-\varepsilon}\langle r\rangle^{-\varepsilon}\#\left\{(n_1,n_2)\in(\mathbb{Z}^*)^2 \mid n_1n_2=r\right\}\\
&\le c_\varepsilon\sum_{r\in\mathbb{Z}^*}\langle a+2r\rangle^{-1-\varepsilon}
\end{aligned}$$

for some $c_\varepsilon > 1$. We write $a = 2b+\delta$, $b\in\mathbb{Z}$, $\delta\in[0,2)$, and

$$\sum_{r\in\mathbb{Z}^*}\langle a+2r\rangle^{-1-\varepsilon} \le \sum_{r\in\mathbb{Z}}\langle r+\delta\rangle^{-1-\varepsilon} \le 3+2\sum_{r\in\mathbb{N}}\langle r\rangle^{-1-\varepsilon} =: s_\varepsilon.$$

Now, the estimates (18) and (19) hold with $C_\varepsilon := s_\varepsilon c_\varepsilon$. In order to show formula (20) of the third part we use the same change of variables as above and obtain

$$\sum_{r\in\mathbb{Z}^*}\langle a+2r\rangle^{-1-\varepsilon}\sum_{\substack{n_1,n_2\in\mathbb{Z}^*\\ n_1n_2=r}}\langle \xi-n_1\rangle^{-\varepsilon}\langle \xi-n_2\rangle^{-\varepsilon}.$$

Let $M(r) = \{(n_1,n_2)\in(\mathbb{Z}^*)^2 \mid n_1n_2 = r\}$. Now, we split the inner sum into two parts. Let

$$M_1(r) = \left\{(n_1,n_2)\in M(r) \mid 6|\xi-n_1| \ge |r|^{\frac16} \text{ or } 6|\xi-n_2| \ge |r|^{\frac16}\right\}$$

and

$$M_2(r) = \left\{(n_1,n_2)\in M(r) \mid 6|\xi-n_1| \le |r|^{\frac16} \text{ and } 6|\xi-n_2| \le |r|^{\frac16}\right\}.$$

Obviously we have $M(r) = M_1(r) \cup M_2(r)$. By part 1 of Lemma 3.1 there exists $c_\varepsilon > 1$ such that

$$\# M_1(r) \leq 2\,\# \left\{(n_1, n_2) \in \mathbb{N}^2 \mid n_1 n_2 = r\right\} \leq c_\varepsilon |r|^{\frac{\varepsilon}{6}},$$

and it follows that

$$\sum_{(n_1,n_2)\in M_1(r)} \langle \xi - n_1\rangle^{-\varepsilon}\langle \xi - n_2\rangle^{-\varepsilon} \leq 6^\varepsilon |r|^{-\frac{\varepsilon}{6}} \# M_1(r) \leq 6^\varepsilon c_\varepsilon.$$

For $(n_1, n_2) \in M_2(r)$ it holds that $3|n_1 - n_2| \leq |r|^{\frac{1}{6}}$. An application of part 2 of Lemma 3.1 shows that

$$\sum_{(n_1,n_2)\in M_2(r)} \langle \xi - n_1\rangle^{-\varepsilon}\langle \xi - n_2\rangle^{-\varepsilon} \leq \# M_2(r) \leq 4.$$

Therefore, we see that

$$\sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \langle \xi_1\rangle^{-\varepsilon}\langle \xi_2\rangle^{-\varepsilon}\langle a + 2(\xi-\xi_1)(\xi-\xi_2)\rangle^{-1-\varepsilon} \leq (6^\varepsilon c_\varepsilon + 4)\sum_{r\in\mathbb{Z}^*}\langle a + 2r\rangle^{-1-\varepsilon},$$

and the third part is proved with constant $C_\varepsilon = s_\varepsilon(6^\varepsilon c_\varepsilon + 4)$. Concerning the fourth part we proceed similarly. After changing variables

$$\begin{aligned}
&\langle \xi_1\rangle^{-\varepsilon} \sum_{\substack{\xi,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \langle \xi_2\rangle^{-\varepsilon}\langle a + 2(\xi-\xi_1)(\xi-\xi_2)\rangle^{-1-\varepsilon}\\
&= \langle \xi_1\rangle^{-\varepsilon}\sum_{r\in\mathbb{Z}^*}\langle a+2r\rangle^{-1-\varepsilon}\sum_{\substack{n_1,n_2\in\mathbb{Z}^*\\ n_1 n_2 = r}}\langle \xi_1 + n_1 - n_2\rangle^{-\varepsilon}
\end{aligned}$$

we consider two subregions of summation. In the case where $|r|^{\frac{1}{6}} \leq 6|\xi_1|$ or $|r|^{\frac{1}{6}} \leq 6|\xi_1 + n_1 - n_2|$ we apply estimate (16) from the first part of Lemma 3.1, while in the remaining case it holds that $3|n_1 - n_2| \leq |r|^{\frac{1}{6}}$, and we utilize estimate (17) from the second part of Lemma 3.1. $\square$

4. The proof of the trilinear estimates. In this section we prove Theorems 2.4 and 2.5. We will frequently use the following well-known tool; see, e.g., [11, Lemma 4.2].

Lemma 4.1. *Let $0 \leq \alpha \leq \beta$ such that $\alpha + \beta > 1$ and $\varepsilon > 0$. Then*

$$\int_{\mathbb{R}} \langle s-a\rangle^{-\alpha}\langle s-b\rangle^{-\beta} ds \lesssim \langle a-b\rangle^{-\gamma}, \quad \gamma = \begin{cases}\alpha+\beta-1, & \beta<1,\\ \alpha-\varepsilon, & \beta=1,\\ \alpha, & \beta>1.\end{cases}$$

We write $T = T^* + T^{**}$, where

$$\begin{aligned}
T^*\widehat{(u_1,u_2},u_3)(\xi) &= (2\pi)^{-1}\sum_{\substack{\xi=\xi_1+\xi_2+\xi_3\\ \xi_1\neq\xi,\ \xi_2\neq\xi}} \widehat{u}_1(\xi_1)\widehat{u}_2(\xi_2) i\xi_3 \widehat{\overline{u}}_3(\xi_3),\\
T^{**}\widehat{(u_1,u_2},u_3)(\xi) &= (2\pi)^{-1}\widehat{u}_1(\xi)\widehat{u}_2(\xi) i\xi\widehat{\overline{u}}_3(-\xi).
\end{aligned}$$

Proof of Theorem 2.4. To fix notation, let $\sigma_0 = \tau + \xi^2$, $\sigma_j = \tau_j + \xi_j^2$, $j = 1, 2$, and $\sigma_3 = \tau_3 - \xi_3^2$. Throughout the proof the quantities ξ_3, τ_3 are defined as $\xi_3 = \xi - \xi_1 - \xi_2$ and $\tau_3 = \tau - \tau_1 - \tau_2$, respectively. Let us denote $\mu = (\tau, \xi)$, $\mu_i = (\tau_i, \xi_i)$, $i = 1, 2, 3$, for brevity. By the definition of the norms we may assume that $\widehat{u}_j \geq 0$. Then

$$\|T(u_1, u_2, u_3)\|_{X^{\frac{1}{2},-\frac{1}{2}}_{r,2}} \leq \|T^*(u_1, u_2, u_3)\|_{X^{\frac{1}{2},-\frac{1}{2}}_{r,2}} + \|T^{**}(u_1, u_2, u_3)\|_{X^{\frac{1}{2},-\frac{1}{2}}_{r,2}},$$

and we consider the contribution from T^* first: Let m be given by

$$m(\mu, \mu_1, \mu_2) = \frac{\langle \xi \rangle^{\frac{1}{2}} i\xi_3}{\prod_{j=1}^3 \langle \xi_j \rangle^{\frac{1}{2}} \prod_{j=0}^3 \langle \sigma_j \rangle^{\frac{1}{2}}}.$$

Estimate (9) for the T^* contribution is equivalent to

$$\left\| \sum_{\substack{\xi_1,\xi_2 \in \mathbb{Z} \\ \xi_1 \neq \xi,\, \xi_2 \neq \xi}} \int m(\mu, \mu_1, \mu_2) f_1(\mu_1) f_2(\mu_2) f_3(\mu_3) d\tau_1 d\tau_2 \right\|_{\ell^{r'}_\xi L^2_\tau} \tag{22}$$
$$\lesssim T^\delta \|f_1\|_{\ell^{q'}_\xi L^2_\tau} \|f_2\|_{\ell^{q'}_\xi L^2_\tau} \|f_3\|_{\ell^{r'}_\xi L^2_\tau},$$

where we may assume that $f_3(\tau_3, 0) = 0$. The resonance relation

$$\sigma_0 - \sigma_1 - \sigma_2 - \sigma_3 = 2(\xi - \xi_1)(\xi - \xi_2) = 2(\xi_1\xi_2 + \xi\xi_3) \tag{23}$$

holds true; cf. [25, 13, 18]. Let us first consider the subregion where $\langle \xi_1 \rangle \langle \xi_2 \rangle \ll \langle \xi \rangle \langle \xi_3 \rangle$. Then

$$\langle \xi \rangle^{\frac{1}{2}} \langle \xi_3 \rangle^{\frac{1}{2}} \lesssim \sum_{k=0}^3 \langle \sigma_k \rangle^{\frac{1}{2}}, \tag{24}$$

and in this subregion we control $|m|$ by the sum of all

$$m_{k,1}(\mu, \mu_1, \mu_2) = \frac{1}{\langle \xi_1 \rangle^{\frac{1}{2}} \langle \xi_2 \rangle^{\frac{1}{2}} \prod_{j=0,\, j\neq k}^3 \langle \sigma_j \rangle^{\frac{1}{2}}}$$

for $k = 0, \ldots, 3$. Second, in the subregion where $\langle \xi \rangle \langle \xi_3 \rangle \lesssim \langle \xi_1 \rangle \langle \xi_2 \rangle$ (note that $\xi_1 \neq \xi$, $\xi_2 \neq \xi$ within the domain of summation) it holds that

$$\langle \xi - \xi_1 \rangle^{\frac{1}{2}} \langle \xi - \xi_2 \rangle^{\frac{1}{2}} \lesssim \sum_{k=0}^3 \langle \sigma_k \rangle^{\frac{1}{2}}, \tag{25}$$

and in this subregion we control $|m|$ by the sum of all

$$m_{k,2}(\mu, \mu_1, \mu_2) = \frac{1}{\langle \xi - \xi_1 \rangle^{\frac{1}{2}} \langle \xi - \xi_2 \rangle^{\frac{1}{2}} \prod_{j=0,\, j\neq k}^3 \langle \sigma_j \rangle^{\frac{1}{2}}}$$

for $k = 0, \ldots, 3$. According to these multipliers we subdivide the proof into several cases summarized in Table 1. For technical reasons, we will prove the slightly stronger

Table 1

Summary of cases (with a preview of the lower bound on q for each subcase obtained by our arguments below).

	$\sigma_0 = \max$	$\sigma_1 = \max$	$\sigma_2 = \max$	$\sigma_3 = \max$
$\langle\xi_1\rangle\langle\xi_2\rangle \ll \langle\xi\rangle\langle\xi_3\rangle$	Case 0.1: $q > 1$	Case 1.1: $q > 4/3$	Case 2.1: $q > 4/3$	Case 3.1: $q > 4/3$
$\langle\xi\rangle\langle\xi_3\rangle \lesssim \langle\xi_1\rangle\langle\xi_2\rangle$	Case 0.2: $q > 1$	Case 1.2: $q > 4/3$	Case 2.2: $q > 4/3$	Case 3.2: $q > 4/3$

estimates

$$\left\| \sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1\neq\xi,\,\xi_2\neq\xi}} \int m_{k,j,\nu}(\mu,\mu_1,\mu_2) f_1(\mu_1) f_2(\mu_2) f_3(\mu_3) d\tau_1 d\tau_2 \right\|_{\ell^{r'}_\xi L^2_\tau} \tag{26}$$
$$\lesssim \|f_1\|_{\ell^{q'}_\xi L^2_\tau} \|f_2\|_{\ell^{q'}_\xi L^2_\tau} \|f_3\|_{\ell^{r'}_\xi L^2_\tau}$$

for any $0 \le \nu < \frac{1}{3q'}$, $k = 0,\dots,3$, and $j = 1,2$, where

$$m_{k,1,\nu}(\mu,\mu_1,\mu_2) = \frac{1}{\langle\xi_1\rangle^{\frac12}\langle\xi_2\rangle^{\frac12}\prod_{j=0,\,j\neq k}^{3}\langle\sigma_j\rangle^{\frac12-\nu}},$$
$$m_{k,2,\nu}(\mu,\mu_1,\mu_2) = \frac{1}{\langle\xi-\xi_1\rangle^{\frac12}\langle\xi-\xi_2\rangle^{\frac12}\prod_{j=0,\,j\neq k}^{3}\langle\sigma_j\rangle^{\frac12-\nu}}$$

for $k = 0,\dots,3$. Clearly, (26) implies (22) with $\delta = 3\nu$ because

$$\left\|\langle\sigma_j\rangle^{-\nu} f_j\right\|_{\ell^{p'}_\xi L^2_\tau} \lesssim T^\nu \|f_j\|_{\ell^{p'}_\xi L^2_\tau}, \quad 1 \le p \le \infty. \tag{27}$$

Case 0.1. We consider the contribution

$$t_{0,1} := \left\| \sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \int \frac{f_1(\mu_1)}{\langle\xi_1\rangle^{\frac12}\langle\sigma_1\rangle^{\frac12-\nu}} \frac{f_2(\mu_2)}{\langle\xi_2\rangle^{\frac12}\langle\sigma_2\rangle^{\frac12-\nu}} \frac{f_3(\mu_3)}{\langle\sigma_3\rangle^{\frac12-\nu}} d\tau_1 d\tau_2 \right\|_{\ell^{r'}_\xi L^2_\tau}$$
$$\lesssim \left\| \sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \langle\xi_1\rangle^{-\frac12}\langle\xi_2\rangle^{-\frac12} I(\mu,\mu_1,\mu_2) \left(\int f_1^2 f_2^2 f_3^2 d\tau_1 d\tau_2\right)^{\frac12} \right\|_{\ell^{r'}_\xi L^2_\tau},$$

where

$$I(\mu,\mu_1,\mu_2) := \left(\int \frac{d\tau_1 d\tau_2}{(\langle\sigma_1\rangle\langle\sigma_2\rangle\langle\sigma_3\rangle)^{1-2\nu}}\right)^{\frac12} \lesssim \langle\sigma^{(0)}_{res}\rangle^{\frac{1}{q'}-\frac12-}$$

with $\sigma^{(0)}_{res} = \tau + \xi^2 - 2(\xi-\xi_1)(\xi-\xi_2)$ by two applications of Lemma 4.1. Hölder's inequality in ξ_1,ξ_2 leads to[5]

$$t_{0,1} \lesssim \left\| \Sigma_{0,1}(\mu) \left(\sum_{\xi_1,\xi_2\in\mathbb{Z}} \left(\int \frac{f_1^2(\mu_1)}{\langle\xi_1\rangle^{1-}} \frac{f_2^2(\mu_2)}{\langle\xi_2\rangle^{1-}} f_3^2(\mu_3) d\tau_1 d\tau_2 \right)^{\frac{\varrho}{2}} \right)^{\frac{1}{\varrho}} \right\|_{\ell^{r'}_\xi L^2_\tau},$$

[5]Note that here the exponents of $\langle\xi_1\rangle$ and $\langle\xi_2\rangle$ may be chosen arbitrarily close to 1, independently of the implicit constraints on ϱ and ϱ'.

where

$$\Sigma_{0,1}(\mu) := \left(\sum_{\substack{\xi_1,\xi_2\in\mathbb{Z} \\ \xi_1,\xi_2\neq\xi}} \langle\xi_1\rangle^{0-}\langle\xi_2\rangle^{0-}\langle\sigma_{res}^{(0)}\rangle^{-1-} \right)^{\frac{1}{\varrho'}}$$

for $\varrho = \frac{2q'}{q'+2}+$ and $\varrho' = \frac{2q'}{q'-2}-$. The sum $\Sigma_{0,1}(\mu)$ is uniformly bounded due to Corollary 3.2, estimate (20). Hence,

$$t_{0,1} \lesssim \left\| \left(\sum_{\xi_1,\xi_2\in\mathbb{Z}} \frac{\|f_1(\cdot,\xi_1)\|_{L^2}^{\varrho}}{\langle\xi_1\rangle^{\frac{\varrho}{2}-}} \frac{\|f_2(\cdot,\xi_2)\|_{L^2}^{\varrho}}{\langle\xi_2\rangle^{\frac{\varrho}{2}-}} \|f_3(\cdot,\xi_3)\|_{L^2}^{\varrho} \right)^{\frac{1}{\varrho}} \right\|_{\ell_\xi^{r'}}$$

by Minkowski's inequality because $\varrho \le 2$. Now, we apply Hölder's inequality to obtain

$$t_{0,1} \lesssim \left\| \left(\sum_{\xi_1,\xi_2\in\mathbb{Z}} \frac{\|f_1(\cdot,\xi_1)\|_{L^2}^{r'}}{\langle\xi_1\rangle^{1-\frac{r'}{q'}+}} \frac{\|f_2(\cdot,\xi_2)\|_{L^2}^{r'}}{\langle\xi_2\rangle^{1-\frac{r'}{q'}+}} \|f_3(\cdot,\xi_3)\|_{L^2}^{r'} \right)^{\frac{1}{r'}} \right\|_{\ell_\xi^{r'}}. \tag{28}$$

Hölder's inequality shows that

$$\left(\sum_{\xi_i\in\mathbb{Z}} \|f_i(\cdot,\xi_i)\|_{L^2}^{r'} \langle\xi_i\rangle^{\frac{r'}{q'}-1-} \right)^{\frac{1}{r'}} \lesssim \|f_i\|_{\ell_\xi^{q'}L_\tau^2}, \quad i=1,2.$$

Hence, Fubini's theorem provides

$$t_{0,1} \lesssim \|f_1\|_{\ell_\xi^{q'}L_\tau^2}\|f_2\|_{\ell_\xi^{q'}L_\tau^2}\|f_3\|_{\ell_\xi^{r'}L_\tau^2}$$

for any $1 < q \le r \le 2$, as desired.

Case 0.2. We consider the contribution

$$t_{0,2} := \left\| \sum_{\substack{\xi_1,\xi_2\in\mathbb{Z} \\ \xi_1,\xi_2\neq\xi}} \int \frac{f_1(\mu_1)}{\langle\xi-\xi_1\rangle^{\frac{1}{2}}\langle\sigma_1\rangle^{\frac{1}{2}-\nu}} \frac{f_2(\mu_2)}{\langle\xi-\xi_2\rangle^{\frac{1}{2}}\langle\sigma_2\rangle^{\frac{1}{2}-\nu}} \frac{f_3(\mu_3)}{\langle\sigma_3\rangle^{\frac{1}{2}-\nu}} d\tau_1 d\tau_2 \right\|_{\ell_\xi^{r'}L_\tau^2}.$$

By replacing the weight $\langle\xi_1\rangle^{-\frac{1}{2}}\langle\xi_2\rangle^{-\frac{1}{2}}$ by $\langle\xi-\xi_1\rangle^{-\frac{1}{2}}\langle\xi-\xi_2\rangle^{-\frac{1}{2}}$ in the expression $t_{0,1}$, the same arguments as in the previous case lead to

$$t_{0,2} \lesssim \left\| \left(\sum_{\xi_1,\xi_2\in\mathbb{Z}} \frac{\|f_1(\cdot,\xi_1)\|_{L^2}^{r'}}{\langle\xi-\xi_1\rangle^{1-\frac{r'}{q'}+}} \frac{\|f_2(\cdot,\xi_2)\|_{L^2}^{r'}}{\langle\xi-\xi_2\rangle^{1-\frac{r'}{q'}+}} \|f_3(\cdot,\xi_3)\|_{L^2}^{r'} \right)^{\frac{1}{r'}} \right\|_{\ell_\xi^{r'}}$$

instead of (28), where we used Corollary 3.2, estimate (18) to bound the sum

$$\Sigma_{0,2}(\mu) := \left(\sum_{\substack{\xi_1,\xi_2\in\mathbb{Z} \\ \xi_1,\xi_2\neq\xi}} \langle\xi-\xi_1\rangle^{0-}\langle\xi-\xi_2\rangle^{0-}\langle\sigma_{res}^{(0)}\rangle^{-1-} \right)^{\frac{1}{\varrho'}}.$$

By the change of variables $\xi \mapsto \xi - \xi_1 - \xi_2$ we obtain

$$t_{0,2} \lesssim \left(\sum_{\xi,\xi_1,\xi_2 \in \mathbb{Z}} \frac{\|f_1(\cdot,\xi_1)\|_{L^2}^{r'}}{\langle \xi+\xi_1\rangle^{1-\frac{r'}{q'}+}} \frac{\|f_2(\cdot,\xi_2)\|_{L^2}^{r'}}{\langle \xi+\xi_2\rangle^{1-\frac{r'}{q'}+}} \|f_3(\cdot,\xi)\|_{L^2}^{r'} \right)^{\frac{1}{r'}}.$$

Now, we sum first in ξ_1, ξ_2 and use

$$\sup_{\xi \in \mathbb{Z}} \left(\sum_{\xi_i \in \mathbb{Z}} \|f_i(\cdot,\xi_i)\|_{L^2}^{r'} \langle \xi+\xi_i\rangle^{\frac{r'}{q'}-1-} \right)^{\frac{1}{r'}} \lesssim \|f_i\|_{\ell_\xi^{q'} L_\tau^2}, \quad i=1,2, \tag{29}$$

to obtain

$$t_{0,2} \lesssim \|f_1\|_{\ell_\xi^{q'} L_\tau^2} \|f_2\|_{\ell_\xi^{q'} L_\tau^2} \|f_3\|_{\ell_\xi^{r'} L_\tau^2}$$

similarly as above.

Case 1.1. From now on we have to restrict ourselves to $2 \le q' < 4$, which is mainly due to various applications of Young's inequality. We use duality and consider for $\varphi \in \ell_\xi^r L_\tau^2$ the quantity $t_{1,1}$ defined by

$$\begin{aligned} &\sum_{\xi \in \mathbb{Z}} \int \frac{\varphi(\mu)}{\langle \sigma_0\rangle^{\frac12-\nu}} \sum_{\substack{\xi_1,\xi_2 \in \mathbb{Z} \\ \xi_1,\xi_2 \neq \xi}} \int \frac{f_1(\mu_1)}{\langle \xi_1\rangle^{\frac12}} \frac{f_2(\mu_2)}{\langle \xi_2\rangle^{\frac12}\langle \sigma_2\rangle^{\frac12-\nu}} \frac{f_3(\mu_3)}{\langle \sigma_3\rangle^{\frac12-\nu}} d\tau_1 d\tau_2 d\tau \\ =& \sum_{\xi_1 \in \mathbb{Z}} \int f_1(\mu_1) \sum_{\substack{\xi,\xi_2 \in \mathbb{Z} \\ \xi_1,\xi_2 \neq \xi}} \int \frac{\varphi(\mu)}{\langle \sigma_0\rangle^{\frac12-\nu}} \frac{f_2(\mu_2)}{\langle \xi_2\rangle^{\frac12}\langle \sigma_2\rangle^{\frac12-\nu}} \frac{f_3(\mu_3)}{\langle \xi_1\rangle^{\frac12}\langle \sigma_3\rangle^{\frac12-\nu}} d\tau d\tau_2 d\tau_1. \end{aligned}$$

The Cauchy–Schwarz inequality in τ, τ_2 and two applications of Lemma 4.1 show that

$$t_{1,1} \lesssim \sum_{\xi_1 \in \mathbb{Z}} \int f_1(\mu_1) \sum_{\substack{\xi,\xi_2 \in \mathbb{Z} \\ \xi_1,\xi_2 \neq \xi}} \langle \sigma_{res}^{(1)}\rangle^{\frac{1}{q'}-\frac12-} \left(\int \frac{\varphi^2(\mu) f_2^2(\mu_2) f_3^2(\mu_3)}{\langle \xi_1\rangle\langle \xi_2\rangle} d\tau d\tau_2 \right)^{\frac12} d\tau_1,$$

where with $\sigma_{res}^{(1)} = \tau_1 + \xi_1^2 + 2(\xi-\xi_1)(\xi-\xi_2)$ Hölder's inequality in ξ, ξ_2 leads to

$$t_{1,1} \lesssim \sum_{\xi_1 \in \mathbb{Z}} \int f_1(\mu_1) \Sigma_{1,1}(\mu_1) \left(\sum_{\xi,\xi_2 \in \mathbb{Z}} \left(\int \frac{\varphi^2(\mu) f_2^2(\mu_2) f_3^2(\mu_3)}{\langle \xi_1\rangle^{1-}\langle \xi_2\rangle^{1-}} d\tau d\tau_2 \right)^{\frac{\varrho}{2}} \right)^{\frac{1}{\varrho}} d\tau_1$$

for $\varrho = \frac{2q'}{q'+2}+$ and $\varrho' = \frac{2q'}{q'-2}-$, where

$$\Sigma_{1,1}(\mu_1) := \left(\langle \xi_1\rangle^{0-} \sum_{\substack{\xi,\xi_2 \in \mathbb{Z} \\ \xi_1,\xi_2 \neq \xi}} \langle \xi_2\rangle^{0-} \langle \sigma_{res}^{(1)}\rangle^{-1-} \right)^{\frac{1}{\varrho'}}.$$

This is bounded by Corollary 3.2, estimate (21). The Cauchy–Schwarz inequality in τ_1 and Minkowski's inequality provide

$$t_{1,1} \lesssim \sum_{\xi_1 \in \mathbb{Z}} \|f_1(\cdot,\xi_1)\|_{L_\tau^2} \left(\sum_{\xi,\xi_2} \frac{\|\varphi(\cdot,\xi)\|_{L_\tau^2}^{\varrho} \|f_2(\cdot,\xi_2)\|_{L_\tau^2}^{\varrho} \|f_3(\cdot,\xi_3)\|_{L_\tau^2}^{\varrho}}{\langle \xi_1\rangle^{\frac{\varrho}{2}-}\langle \xi_2\rangle^{\frac{\varrho}{2}-}} \right)^{\frac{1}{\varrho}}. \tag{30}$$

Now, we use Hölder's inequality in ξ_1 to obtain

$$t_{1,1} \lesssim \|f_1\|_{\ell^{q'}_\xi L^2_\tau} \left\| \left(\sum_{\xi,\xi_2} \frac{\|f_2(\cdot,\xi_2)\|^\varrho_{L^2_\tau} \|f_3(\cdot,\xi_3)\|^\varrho_{L^2_\tau} \|\varphi(\cdot,\xi)\|^\varrho_{L^2_\tau}}{\langle \xi_2 \rangle^{\frac{\varrho}{2}-}} \right)^{\frac{1}{\varrho}} \right\|_{\ell^{\varrho'}_{\xi_1}} . \tag{31}$$

By the change of variables $\xi_2 \mapsto -\xi_2$, $\xi \mapsto \xi$ the second factor equals

$$\left\| \sum_{\xi,\xi_2} \frac{\|\tilde{f}_2(\cdot,\xi_2)\|^\varrho_{L^2_\tau}}{\langle \xi_2 \rangle^{\frac{\varrho}{2}-}} \|\varphi(\cdot,\xi)\|^\varrho_{L^2_\tau} \|\tilde{f}_3(\cdot,\xi_1-\xi_2-\xi)\|^\varrho_{L^2_\tau} \right\|^{\frac{1}{\varrho}}_{\ell^{\frac{\varrho'}{\varrho}}_{\xi_1}},$$

where $\tilde{f}_j = f_j(-\cdot,-\cdot)$, $j=2,3$. This convolution is bounded by

$$\left\| \frac{\tilde{f}_2}{\langle \xi_2 \rangle^{\frac{1}{2}-}} \right\|_{\ell^\varrho_{\xi_2} L^2_\tau} \|\varphi\|_{\ell^r_\xi L^2_\tau} \|\tilde{f}_3\|_{\ell^{r'}_\xi L^2_\tau}$$

due to Young's inequality, because

$$2 + \frac{\varrho}{\varrho'} = 1 + \frac{\varrho}{r} + \frac{\varrho}{r'}.$$

Another application of Hölder's inequality with respect to f_2 yields

$$t_{1,1} \lesssim \|\varphi\|_{\ell^r_\xi L^2_\tau} \|f_1\|_{\ell^{q'}_\xi L^2_\tau} \|f_2\|_{\ell^{q'}_\xi L^2_\tau} \|f_3\|_{\ell^{r'}_\xi L^2_\tau}$$

for $4/3 < q \le r \le 2$.

Case 1.2. For the contribution $t_{1,2}$ the same approach as above leads to

$$t_{1,2} \lesssim \sum_{\xi_1 \in \mathbb{Z}} \|f_1(\cdot,\xi_1)\|_{L^2_\tau} \left(\sum_{\xi,\xi_2} \frac{\|\varphi(\cdot,\xi)\|^\varrho_{L^2_\tau} \|f_2(\cdot,\xi_2)\|^\varrho_{L^2_\tau} \|f_3(\cdot,\xi_3)\|^\varrho_{L^2_\tau}}{\langle \xi-\xi_1 \rangle^{\frac{\varrho}{2}-} \langle \xi-\xi_2 \rangle^{\frac{\varrho}{2}-}} \right)^{\frac{1}{\varrho}}$$

instead of (30) by replacing $\langle \xi_1 \rangle$, $\langle \xi_2 \rangle$ in $t_{1,1}$ by $\langle \xi-\xi_1 \rangle$, $\langle \xi-\xi_2 \rangle$, respectively. The only difference is the use of Corollary 3.2, estimate (19) to bound the sum

$$\Sigma_{1,2}(\mu_1) := \left(\sum_{\substack{\xi,\xi_2 \in \mathbb{Z} \\ \xi_1,\xi_2 \neq \xi}} \langle \xi-\xi_1 \rangle^{0-} \langle \xi-\xi_2 \rangle^{0-} \langle \sigma^{(1)}_{res} \rangle^{-1-} \right)^{\frac{1}{\varrho'}} .$$

Hölder's inequality in ξ_1 and then in ξ,ξ_2 provides

$$\begin{aligned} t_{1,2} &\lesssim \|f_1\|_{\ell^{q'}_\xi L^2_\tau} \left\| \left(\sum_{\xi,\xi_2} \frac{\|f_2(\cdot,\xi_2)\|^\varrho_{L^2_\tau} \|f_3(\cdot,\xi_3)\|^\varrho_{L^2_\tau} \|\varphi(\cdot,\xi)\|^\varrho_{L^2_\tau}}{\langle \xi-\xi_1 \rangle^{\frac{\varrho}{2}-} \langle \xi-\xi_2 \rangle^{\frac{\varrho}{2}-}} \right)^{\frac{1}{\varrho}} \right\|_{\ell^q_{\xi_1}} \\ &\lesssim \|f_1\|_{\ell^{q'}_\xi L^2_\tau} \left(\sum_{\xi,\xi_1,\xi_2 \in \mathbb{Z}} \frac{\|\varphi(\cdot,\xi)\|^q_{L^2}}{\langle \xi-\xi_1 \rangle^{1-\frac{q}{q'}+}} \frac{\|f_2(\cdot,\xi_2)\|^q_{L^2}}{\langle \xi-\xi_2 \rangle^{1-\frac{q}{q'}+}} \|f_3(\cdot,\xi_3)\|^q_{L^2} \right)^{\frac{1}{q}} . \end{aligned}$$

Now, Hölder's inequality in ξ and the change of variables $\xi_1 \mapsto \xi - \xi_1$ give

$$\left(\sum_{\xi,\xi_1,\xi_2 \in \mathbb{Z}} \frac{\|\varphi(\cdot,\xi)\|_{L^2}^q}{\langle \xi - \xi_1 \rangle^{1-\frac{q}{q'}+}} \frac{\|f_2(\cdot,\xi_2)\|_{L^2}^q}{\langle \xi - \xi_2 \rangle^{1-\frac{q}{q'}+}} \|f_3(\cdot,\xi_3)\|_{L^2}^q \right)^{\frac{1}{q}}$$
$$\lesssim \|\varphi\|_{\ell_\xi^r L_\tau^2} \left\| \sum_{\xi_2 \in \mathbb{Z}} \frac{\|f_2(\cdot,\xi_2)\|_{L^2}^q}{\langle \xi - \xi_2 \rangle^{1-\frac{q}{q'}+}} \sum_{\xi_1 \in \mathbb{Z}} \frac{\|f_3(\cdot,\xi_1-\xi_2)\|_{L^2}^q}{\langle \xi_1 \rangle^{1-\frac{q}{q'}+}} \right\|_{\ell_\xi^{\frac{r}{r-q}}}^{\frac{1}{q}}.$$

Let us define

$$\psi(\xi_2) = \sum_{\xi_1 \in \mathbb{Z}} \frac{\|f_3(\cdot,\xi_1-\xi_2)\|_{L^2}^q}{\langle \xi_1 \rangle^{1-\frac{q}{q'}+}}.$$

Young's inequality shows that

$$\|\psi\|_{\ell^{\frac{r}{r-q}}} \lesssim \|f_3\|_{\ell_\xi^{r'} L_\tau^2}^q$$

and therefore

$$\left\| \sum_{\xi_2 \in \mathbb{Z}} \frac{\|f_2(\cdot,\xi_2)\|_{L^2}^q}{\langle \xi - \xi_2 \rangle^{1-\frac{q}{q'}+}} \psi(\xi_2) \right\|_{\ell_\xi^{\frac{r}{r-q}}}^{\frac{1}{q}} \lesssim \|f_3\|_{\ell_\xi^{r'} L_\tau^2} \left\| \sum_{\xi_2 \in \mathbb{Z}} \frac{\|f_2(\cdot,\xi_2)\|_{L^2}^r}{\langle \xi - \xi_2 \rangle^{\frac{r}{q}-\frac{r}{q'}+}} \right\|_{\ell_\xi^{\frac{q}{r-q}}}^{\frac{1}{r}}$$

by Hölder's inequality in ξ_2. Due to the fact that

$$1 + \frac{r-q}{q} = \frac{r}{q'} + \frac{r}{q} - \frac{r}{q'},$$

Young's inequality shows that

$$t_{1,2} \lesssim \|\varphi\|_{\ell_\xi^r L_\tau^2} \|f_1\|_{\ell_\xi^{q'} L_\tau^2} \|f_2\|_{\ell_\xi^{q'} L_\tau^2} \|f_3\|_{\ell_\xi^{r'} L_\tau^2}$$

for all $4/3 < q \leq r \leq 2$.

Case 2.1. To control the contribution from $m_{2,1,\nu}$ we exchange the roles of f_1 and f_2, and the arguments from Case 1.1 apply.

Case 2.2. To control the contribution from $m_{2,2,\nu}$ we exchange the roles of f_1 and f_2, and the arguments from Case 1.2 apply.

Case 3.1. Fix $2 \leq q' < 4$ and $0 \leq \nu < \frac{1}{3q'}$. By the change of variables

$$\mu_1 \mapsto -\mu_1,\ \mu_2 \mapsto -\mu_2,\ \mu \mapsto \mu - \mu_1 - \mu_2 \tag{32}$$

we obtain for the contribution $t_{3,1}$ the identity

$$\sum_{\xi \in \mathbb{Z}} \int \frac{\varphi(\mu)}{\langle \sigma_0 \rangle^{\frac{1}{2}-\nu}} \sum_{\substack{\xi_1,\xi_2 \in \mathbb{Z} \\ \xi_1,\xi_2 \neq \xi}} \iint \frac{f_1(\mu_1)}{\langle \xi_1 \rangle^{\frac{1}{2}} \langle \sigma_1 \rangle^{\frac{1}{2}-\nu}} \frac{f_2(\mu_2)}{\langle \xi_2 \rangle^{\frac{1}{2}} \langle \sigma_2 \rangle^{\frac{1}{2}-\nu}} f_3(\mu_3) d\tau_1 d\tau_2 d\tau$$
$$= \sum_{\xi \in \mathbb{Z}} \int f_3(\mu) \sum_{\substack{\xi_1,\xi_2 \in \mathbb{Z} \\ \xi_1,\xi_2 \neq \xi}} \iint \frac{\tilde{f}_1(\mu_1)}{\langle \xi_1 \rangle^{\frac{1}{2}} \langle \tilde{\sigma}_1 \rangle^{\frac{1}{2}-\nu}} \frac{\tilde{f}_2(\mu_2)}{\langle \xi_2 \rangle^{\frac{1}{2}} \langle \tilde{\sigma}_2 \rangle^{\frac{1}{2}-\nu}} \frac{\varphi(\mu_3)}{\langle \tilde{\sigma}_3 \rangle^{\frac{1}{2}-\nu}} d\tau_1 d\tau_2 d\tau,$$

where $\tilde{\sigma}_1 = \tau_1 - \xi_1^2$, $\tilde{\sigma}_2 = \tau_2 - \xi_2^2$, $\tilde{\sigma}_3 = \tau - \tau_1 - \tau_2 + (\xi - \xi_1 - \xi_2)^2$, and $\tilde{f}_j = f_j(-\cdot, -\cdot)$, $j = 1, 2$. Using the Cauchy–Schwarz inequality in τ_1, τ_2 and Lemma 4.1, the quantity $t_{3,1}$ is bounded by

$$\sum_{\xi\in\mathbb{Z}} \int f_3(\mu) \sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \langle \sigma_{res}^{(3)} \rangle^{\frac{1}{q'}-\frac{1}{2}-} \left(\iint \frac{\tilde{f}_1^2(\mu_1)\tilde{f}_2^2(\mu_2)\varphi^2(\mu_3)}{\langle \xi_1\rangle\langle\xi_2\rangle} d\tau_1 d\tau_2 \right)^{\frac{1}{2}} d\tau,$$

where $\sigma_{res}^{(3)} = \tau - \xi^2 + 2(\xi - \xi_1)(\xi - \xi_2)$. Hölder's inequality leads to

$$\sum_{\xi\in\mathbb{Z}} \int f_3(\mu)\Sigma_{3,1}(\mu) \left(\sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \left(\iint \frac{\tilde{f}_1^2(\mu_1)\tilde{f}_2^2(\mu_2)\varphi^2(\mu_3)}{\langle \xi_1\rangle^{1-}\langle\xi_2\rangle^{1-}} d\tau_1 d\tau_2 \right)^{\frac{\varrho}{2}} \right)^{\frac{1}{\varrho}} d\tau$$

as an upper bound for $t_{3,1}$ with

$$\Sigma_{3,1}(\mu) = \left(\sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \langle\xi_1\rangle^{0-}\langle\xi_2\rangle^{0-}\langle\sigma_{res}^{(3)}\rangle^{-1-} \right)^{\frac{1}{\varrho'}}$$

which is uniformly bounded by Corollary 3.2, estimate (20). By the Cauchy–Schwarz inequality in τ and Minkowski's inequality $t_{3,1}$ is dominated by

$$\sum_{\xi\in\mathbb{Z}} \|f_3(\cdot,\xi)\|_{L^2_\tau} \left(\sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \frac{\|\tilde{f}_1(\cdot,\xi_1)\|_{L^2_\tau}^{\varrho} \|\tilde{f}_2(\cdot,\xi_2)\|_{L^2_\tau}^{\varrho} \|\varphi(\cdot,\xi_3)\|_{L^2_\tau}^{\varrho}}{\langle\xi_1\rangle^{\frac{\varrho}{2}-}\langle\xi_2\rangle^{\frac{\varrho}{2}-}} \right)^{\frac{1}{\varrho}}.$$

Now, we recall that $r \geq \varrho$ for all $4/3 < q \leq r \leq 2$ and apply Hölder's inequality and Fubini's theorem to obtain

$$\begin{aligned} t_{3,1} &\lesssim \|f_3\|_{\ell^{r'}_\xi L^2_\tau} \left\| \left(\sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \frac{\|\tilde{f}_1(\cdot,\xi_1)\|_{L^2_\tau}^{\varrho} \|\tilde{f}_2(\cdot,\xi_2)\|_{L^2_\tau}^{\varrho} \|\varphi(\cdot,\xi_3)\|_{L^2_\tau}^{\varrho}}{\langle\xi_1\rangle^{\frac{\varrho}{2}-}\langle\xi_2\rangle^{\frac{\varrho}{2}-}} \right)^{\frac{1}{\varrho}} \right\|_{\ell^r_\xi} \\ &\lesssim \|f_3\|_{\ell^{r'}_\xi L^2_\tau} \left(\sum_{\substack{\xi,\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \frac{\|\tilde{f}_1(\cdot,\xi_1)\|_{L^2_\tau}^{r} \|\tilde{f}_2(\cdot,\xi_2)\|_{L^2_\tau}^{r} \|\varphi(\cdot,\xi_3)\|_{L^2_\tau}^{r}}{\langle\xi_1\rangle^{1-\frac{r}{q'}+}\langle\xi_2\rangle^{1-\frac{r}{q'}+}} \right)^{\frac{1}{r}}. \end{aligned} \tag{33}$$

Again, Hölder's inequality shows that

$$\left(\sum_{\xi_i\in\mathbb{Z}} \|\tilde{f}_i(\cdot,\xi_i)\|_{L^2}^r \langle\xi_i\rangle^{\frac{r}{q'}-1-} \right)^{\frac{1}{r}} \lesssim \|f_i\|_{\ell^{q'}_\xi L^2_\tau}, \quad i = 1, 2.$$

Hence,

$$t_{3,1} \lesssim \|f_1\|_{\ell^{q'}_\xi L^2_\tau} \|f_2\|_{\ell^{q'}_\xi L^2_\tau} \|f_3\|_{\ell^{r'}_\xi L^2_\tau} \|\varphi\|_{\ell^r_\xi L^2_\tau}$$

for $4/3 < q \le r \le 2$, as desired.

Case 3.2. To obtain the contribution $t_{3,2}$ we replace $\langle \xi_1 \rangle$ and $\langle \xi_2 \rangle$ in $t_{3,1}$ by $\langle \xi - \xi_1 \rangle$ and $\langle \xi - \xi_2 \rangle$, respectively. The change of variables (32) transforms $\langle \xi - \xi_1 \rangle$ into $\langle \xi - \xi_2 \rangle$ and vice versa, and we follow the arguments above to obtain

$$t_{3,2} \lesssim \|f_3\|_{\ell_\xi^{r'} L_\tau^2} \left(\sum_{\substack{\xi,\xi_1,\xi_2 \in \mathbb{Z} \\ \xi_1,\xi_2 \neq \xi}} \frac{\|\tilde{f}_1(\cdot,\xi_1)\|_{L_\tau^2}^r \|\tilde{f}_2(\cdot,\xi_2)\|_{L_\tau^2}^r \|\varphi(\cdot,\xi_3)\|_{L_\tau^2}^r}{\langle \xi - \xi_1 \rangle^{1-\frac{r}{q'}+} \langle \xi - \xi_2 \rangle^{1-\frac{r}{q'}+}} \right)^{\frac{1}{r}}$$

instead of (33), with the only exception that

$$\Sigma_{3,2}(\mu) := \left(\sum_{\substack{\xi_1,\xi_2 \in \mathbb{Z} \\ \xi_1,\xi_2 \neq \xi}} \langle \xi - \xi_1 \rangle^{0-} \langle \xi - \xi_2 \rangle^{0-} \langle \sigma_{res}^{(3)} \rangle^{-1-} \right)^{\frac{1}{\varrho'}}$$

is controlled by Corollary 3.2, estimate (18). Similar to Case 0.2, the change of variables $\xi \mapsto \xi - \xi_1 - \xi_2$ and Hölder's inequality

$$\sup_{\xi \in \mathbb{Z}} \left(\sum_{\xi_i \in \mathbb{Z}} \|f_i(\cdot,\xi_i)\|_{L^2}^r \langle \xi + \xi_i \rangle^{\frac{r}{q'}-1-} \right)^{\frac{1}{r}} \lesssim \|f_i\|_{\ell_\xi^{q'} L_\tau^2}, \quad i = 1,2,$$

yield

$$t_{3,2} \lesssim \|f_1\|_{\ell_\xi^{q'} L_\tau^2} \|f_2\|_{\ell_\xi^{q'} L_\tau^2} \|f_3\|_{\ell_\xi^{r'} L_\tau^2} \|\varphi\|_{\ell_\xi^{r} L_\tau^2},$$

and the estimate for T^* is done. Finally, we consider the harmless contribution from T^{**} and show the much stronger estimate

$$\|T^{**}(u_1,u_2,u_3)\|_{X_{r,2}^{\frac{1}{2},0}} \lesssim \|u_1\|_{X_{1,2}^{\frac{1}{2},\frac{1}{3}+}} \|u_2\|_{X_{1,2}^{\frac{1}{2},\frac{1}{3}+}} \|u_3\|_{X_{r,2}^{\frac{1}{2},\frac{1}{3}+}}, \tag{34}$$

which immediately yields the desired estimate by trivial embeddings and (27). Indeed, Young's and Hölder's inequalities provide

$$\begin{aligned} &\left\| \langle \xi \rangle^{\frac{1}{2}} \xi \int \frac{f_1(\tau_1,\xi)}{\langle \xi \rangle^{\frac{1}{2}} \langle \tau_1 + \xi^2 \rangle^{\frac{1}{3}+}} \frac{f_2(\tau_2,\xi)}{\langle \xi \rangle^{\frac{1}{2}} \langle \tau_2 + \xi^2 \rangle^{\frac{1}{3}+}} \frac{f_3(\tau_3,-\xi)}{\langle \xi \rangle^{\frac{1}{2}} \langle \tau_3 - \xi^2 \rangle^{\frac{1}{3}+}} d\tau_1 d\tau_2 \right\|_{\ell_\xi^{r'} L_\tau^2} \\ \lesssim &\left\| \left\| \frac{f_1(\tau_1,\xi)}{\langle \tau_1 + \xi^2 \rangle^{\frac{1}{3}+}} \right\|_{L_{\tau_1}^{\frac{6}{5}}} \left\| \frac{f_2(\tau_2,\xi)}{\langle \tau_2 + \xi^2 \rangle^{\frac{1}{3}+}} \right\|_{L_{\tau_2}^{\frac{6}{5}}} \left\| \frac{f_3(\tau_3,-\xi)}{\langle \tau_3 - \xi^2 \rangle^{\frac{1}{3}+}} \right\|_{L_{\tau_3}^{\frac{6}{5}}} \right\|_{\ell_\xi^{r'}} \\ \lesssim &\|f_1\|_{\ell_\xi^\infty L_\tau^2} \|f_2\|_{\ell_\xi^\infty L_\tau^2} \|f_3\|_{\ell_\xi^{r'} L_\tau^2}. \end{aligned}$$

This concludes the proof of Theorem 2.4. ☐

Proof of Theorem 2.5. Due to the emdedding (6) the estimate for T^{**} is already covered by (34). With the same notation as above, the estimate (10) for the contribution T^* is equivalent to

$$\begin{aligned} &\left\| \sum_{\substack{\xi_1,\xi_2 \in \mathbb{Z} \\ \xi_1 \neq \xi,\, \xi_2 \neq \xi}} \int n(\mu,\mu_1,\mu_2) f_1(\mu_1) f_2(\mu_2) f_3(\mu_3) d\tau_1 d\tau_2 \right\|_{\ell_\xi^{r'} L_\tau^1} \\ &\lesssim T^\delta \|f_1\|_{\ell_\xi^{q'} L_\tau^2} \|f_2\|_{\ell_\xi^{q'} L_\tau^2} \|f_3\|_{\ell_\xi^{r'} L_\tau^2}, \end{aligned} \tag{35}$$

where $n = \langle\sigma_0\rangle^{-\frac{1}{2}} m$. We decompose $n_{k,j} = \langle\sigma_0\rangle^{-\frac{1}{2}} m_{k,j}$ as above. Again, due to the embedding (6) the stronger estimate (26) already proves estimate (35) for n replaced by $n_{k,j,\nu}$ with $k = 1,2,3$, $j = 1,2$, corresponding to Cases 1–3 above. Hence, it is enough to consider the case $k = 0$, where $\langle\sigma_0\rangle$ is the maximal modulation.

Case 0.1. Let us fix $1 < q \le r \le 2$ and $0 \le \delta < \frac{1}{q'}$. We proceed similarly to the Case 0.1 in the proof of Theorem 2.4:

$$\widetilde{t_{0,1}} := \left\| \langle\sigma_0\rangle^{-\frac{1}{2}} \sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \int \frac{f_1(\mu_1)}{\langle\xi_1\rangle^{\frac{1}{2}}\langle\sigma_1\rangle^{\frac{1}{2}-\nu}} \frac{f_2(\mu_2)}{\langle\xi_2\rangle^{\frac{1}{2}}\langle\sigma_2\rangle^{\frac{1}{2}-\nu}} \frac{f_3(\mu_3)}{\langle\sigma_3\rangle^{\frac{1}{2}-\nu}} d\tau_1 d\tau_2 \right\|_{\ell^{r'}_\xi L^1_\tau}$$
$$\lesssim \left\| \sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \int \frac{g_1(\mu_1)}{\langle\xi_1\rangle^{\frac{1}{2}}\langle\sigma_1\rangle^{\frac{1}{2}-\nu-}} \frac{g_2(\mu_2)}{\langle\xi_2\rangle^{\frac{1}{2}}\langle\sigma_2\rangle^{\frac{1}{2}-\nu-}} \frac{g_3(\mu_3)}{\langle\sigma_3\rangle^{\frac{1}{2}-\nu-}} d\tau_1 d\tau_2 \right\|_{\ell^{r'}_\xi L^p_\tau}$$

for any $p = 2-$, where we define $g_j = \langle\sigma_j\rangle^{0-} f_j$ such that

$$\|g_j\|_{\ell^{q'}_\xi L^p_\tau} \lesssim \|f_j\|_{\ell^{q'}_\xi L^2_\tau}. \tag{36}$$

Now, by Hölder's inequality and two applications of Lemma 4.1 we get

$$\widetilde{t_{0,1}} \lesssim \left\| \sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \langle\xi_1\rangle^{-\frac{1}{2}}\langle\xi_2\rangle^{-\frac{1}{2}}\langle\sigma^{(0)}_{res}\rangle^{\frac{1}{q'}-\frac{1}{2}-} \left(\int g_1^p g_2^p g_3^p d\tau_1 d\tau_2\right)^{\frac{1}{p}} \right\|_{\ell^{r'}_\xi L^p_\tau}$$

with $\sigma^{(0)}_{res} = \tau + \xi^2 - 2(\xi-\xi_1)(\xi-\xi_2)$. Hölder's inequality in ξ_1, ξ_2 leads to

$$\widetilde{t_{0,1}} \lesssim \left\| \widetilde{\Sigma_{0,1}}(\mu) \left(\sum_{\xi_1,\xi_2\in\mathbb{Z}} \left(\int \frac{g_1^p(\mu_1)}{\langle\xi_1\rangle^{1-}} \frac{g_2^p(\mu_2)}{\langle\xi_2\rangle^{1-}} g_3^p(\mu_3) d\tau_1 d\tau_2 \right)^{\frac{\varrho}{p}} \right)^{\frac{1}{\varrho}} \right\|_{\ell^{r'}_\xi L^p_\tau},$$

where

$$\widetilde{\Sigma_{0,1}}(\mu) := \left(\sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \langle\xi_1\rangle^{0-}\langle\xi_2\rangle^{0-}\langle\sigma^{(0)}_{res}\rangle^{-1-} \right)^{\frac{1}{\varrho'}}$$

for $\varrho = \frac{2q'}{q'+2}+$ and $\varrho' = \frac{2q'}{q'-2}-$. The sum $\widetilde{\Sigma_{0,1}}(\mu)$ is uniformly bounded due to Corollary 3.2, estimate (20). Hence,

$$\widetilde{t_{0,1}} \lesssim \left\| \left(\sum_{\xi_1,\xi_2\in\mathbb{Z}} \frac{\|g_1(\cdot,\xi_1)\|^\varrho_{L^p}}{\langle\xi_1\rangle^{\frac{\varrho}{2}-}} \frac{\|g_2(\cdot,\xi_2)\|^\varrho_{L^p}}{\langle\xi_2\rangle^{\frac{\varrho}{2}-}} \|g_3(\cdot,\xi_3)\|^\varrho_{L^p} \right)^{\frac{1}{\varrho}} \right\|_{\ell^{r'}_\xi}$$

by Minkowski's inequality because $\varrho \le p$. Now, we apply Hölder's inequality and Fubini's theorem as in Case 0.1 of the proof of Theorem 2.4 and obtain

$$\widetilde{t_{0,1}} \lesssim \|g_1\|_{\ell^{q'}_\xi L^p_\tau}\|g_2\|_{\ell^{q'}_\xi L^p_\tau}\|g_3\|_{\ell^{r'}_\xi L^p_\tau}$$

for any $1 < q \le r \le 2$. Finally, (36) proves the desired estimate.

Case 0.2. We consider $\widetilde{t_{0,2}}$ defined as

$$\left\| \langle\sigma_0\rangle^{-\frac{1}{2}} \sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \int \frac{f_1(\mu_1)}{\langle\xi-\xi_1\rangle^{\frac{1}{2}}\langle\sigma_1\rangle^{\frac{1}{2}-\nu}} \frac{f_2(\mu_2)}{\langle\xi-\xi_2\rangle^{\frac{1}{2}}\langle\sigma_2\rangle^{\frac{1}{2}-\nu}} \frac{f_3(\mu_3)}{\langle\sigma_3\rangle^{\frac{1}{2}-\nu}} d\tau_1 d\tau_2 \right\|_{\ell^{r'}_\xi L^1_\tau}.$$

The same arguments as in the previous case lead to

$$\widetilde{t_{0,2}} \lesssim \left\| \left(\sum_{\xi_1,\xi_2\in\mathbb{Z}} \frac{\|g_1(\cdot,\xi_1)\|^{\varrho}_{L^p}}{\langle\xi-\xi_1\rangle^{\frac{\varrho}{2}-}} \frac{\|g_2(\cdot,\xi_2)\|^{\varrho}_{L^p}}{\langle\xi-\xi_2\rangle^{\frac{\varrho}{2}-}} \|g_3(\cdot,\xi_3)\|^{\varrho}_{L^p} \right)^{\frac{1}{\varrho}} \right\|_{\ell^{r'}_\xi},$$

where we used Corollary 3.2, estimate (18) to bound the sum

$$\widetilde{\Sigma_{0,2}}(\mu) := \left(\sum_{\substack{\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1,\xi_2\neq\xi}} \langle\xi-\xi_1\rangle^{0-}\langle\xi-\xi_2\rangle^{0-}\langle\sigma^{(0)}_{res}\rangle^{-1-} \right)^{\frac{1}{\varrho'}}.$$

By the change of variables $\xi \mapsto \xi-\xi_1-\xi_2$ we obtain

$$\widetilde{t_{0,2}} \lesssim \left(\sum_{\xi,\xi_1,\xi_2\in\mathbb{Z}} \frac{\|g_1(\cdot,\xi_1)\|^{r'}_{L^p}}{\langle\xi+\xi_1\rangle^{1-\frac{r'}{q'}+}} \frac{\|g_2(\cdot,\xi_2)\|^{r'}_{L^p}}{\langle\xi+\xi_2\rangle^{1-\frac{r'}{q'}+}} \|g_3(\cdot,\xi)\|^{r'}_{L^p} \right)^{\frac{1}{r'}}.$$

Now, we sum first in ξ_1,ξ_2 and use the analogue of (29) for $\|g_i\|_{L^p_\tau}$ $(i=1,2)$ to obtain

$$\widetilde{t_{0,2}} \lesssim \|g_1\|_{\ell^{q'}_\xi L^p_\tau}\|g_2\|_{\ell^{q'}_\xi L^p_\tau}\|g_3\|_{\ell^{r'}_\xi L^p_\tau}$$

and recall the property (36) of g_j. ◻

Proof of Remark 3. Assume that the estimate (11) is valid for some $b \le 0$, $1 \le r \le \frac{4}{3}$, and $1 \le p,q \le \infty$. Then for all $f_i \in \ell^{r'}_\xi L^{q'}_\tau$ $(1 \le i \le 3)$ and $f_0 \in \ell^r_\xi L^p_\tau$ we have

$$\sum_{\substack{\xi,\xi_1,\xi_2\in\mathbb{Z}\\ \xi_1\neq\xi,\ \xi_2\neq\xi}} \int \frac{\langle\xi\rangle^{\frac{1}{2}} f_0(\mu) f_1(\mu_1) f_2(\mu_2)|\xi_3| f_3(\mu_3)}{\langle\sigma_0\rangle^{-b}\langle\xi_1\rangle^{\frac{1}{2}}\langle\sigma_1\rangle^{\frac{1}{2}}\langle\xi_2\rangle^{\frac{1}{2}}\langle\sigma_2\rangle^{\frac{1}{2}}\langle\xi_3\rangle^{\frac{1}{2}}\langle\sigma_3\rangle^{\frac{1}{2}}} d\tau d\tau_1 d\tau_2 < \infty. \tag{37}$$

We choose

$$\begin{aligned}
f_0(0,\tau) &= \chi(\tau) \text{ and } f_0(\xi,\tau) = 0 \quad \text{for } \xi \neq 0,\\
f_2(1,\tau_2) &= \chi(\tau_2) \text{ and } f_2(\xi_2,\tau_2) = 0 \quad \text{for } \xi_2 \neq 1,\\
f_1(0,\tau_1) &= 0 \text{ and } f_1(\xi_1,\tau_1) = \frac{\chi(\tau_1+(\xi_1+1)^2)}{\langle\xi_1\rangle^{\frac{1}{4}}\ln^{\frac{1}{3}}(\langle\xi_1\rangle)} \quad \text{for } \xi_1 \neq 0,\\
f_3(0,\tau_3) &= 0 \text{ and } f_3(\xi_3,\tau_3) = \frac{\chi(\tau_3-\xi_3^2)}{\langle\xi_3\rangle^{\frac{1}{4}}\ln^{\frac{1}{3}}(\langle\xi_3\rangle)} \quad \text{for } \xi_3 \neq 0,
\end{aligned}$$

where χ denotes the characteristic function of $[-1,1]$. Then $f_0 \in \ell^r_\xi L^p_\tau$ and $f_2 \in \ell^{r'}_\xi L^{q'}_\tau$ for all $1 \le r,p,q \le \infty$ and $f_1, f_3 \in \ell^{r'}_\xi L^{q'}_\tau$ for all $r' \ge 4$ and $1 \le q \le \infty$. Let $I(\xi_1)$ be defined as

$$
\begin{aligned}
&\int \chi(\tau)\chi(\tau_1 + (\xi_1+1)^2)\chi(\tau_2)\chi(\tau - \tau_1 - \tau_2 - (\xi_1+1)^2)d\tau d\tau_1 d\tau_2 \\
\gtrsim &\int \chi(\tau)\chi(\tau_1 + (\xi_1+1)^2)\chi(\tau - \tau_1 - (\xi_1+1)^2)d\tau_1 d\tau \\
\gtrsim &\int \chi(\tau_1 + (\xi_1+1)^2)d\tau_1 \gtrsim 1.
\end{aligned}
$$

Due to $\langle\sigma_1\rangle^{\frac12} \lesssim \langle\xi_1\rangle^{\frac12}$, the left-hand side of (37) becomes

$$
\sum_{|\xi_1|\ge 1} \frac{I(\xi_1)}{\langle\xi_1\rangle \ln^{\frac23}(\langle\xi_1\rangle)} \gtrsim \sum_{|\xi_1|\ge 1} \frac{1}{\langle\xi_1\rangle \ln^{\frac23}(\langle\xi_1\rangle)} = \infty,
$$

which contradicts (37). ☐

5. The proof of the quintilinear estimate. Before we start with the proof of Theorem 2.6 we show the following trilinear refinement of the L^6 Strichartz-type estimate; see [3, Proposition 2.36]. The major point is that for one of the factors the loss of ε derivatives can be avoided. In fact, this refinement also follows by carefully using the decomposition arguments and the Galilean transformation in [3, section 5]. However, we decided to present a proof based on the representation $\|u\|^3_{L^6_{xt}} = \|u^2\overline{u}\|_{L^2_{xt}}$ which we learned from [23], in combination with the estimates from section 3. Similar arguments were already used in [10, Proposition 4.6 and its proof].

LEMMA 5.1. *For $\frac13 < b < \frac12$ and $s > 3(\frac12 - b)$ the estimate*

$$
\|u_1 u_2 \overline{u}_3\|_{L^2_{xt}} \lesssim \|u_1\|_{X^{s,b}} \|u_2\|_{X^{s,b}} \|u_3\|_{X^{0,b}} \tag{38}
$$

holds true.

Proof. We rewrite $u_1 u_2 \overline{u}_3 = C_1(u_1,u_2,u_3) + C_2(u_1,u_2,u_3)$ for

$$
\begin{aligned}
\widehat{C_1(u_1,u_2,u_3)}(\xi) &= (2\pi)^{-1} \sum_{\substack{\xi=\xi_1+\xi_2+\xi_3 \\ \xi_1\neq\xi,\, \xi_2\neq\xi}} \widehat{u}_1(\xi_1)\widehat{u}_2(\xi_2)\widehat{\overline{u}}_3(\xi_3), \\
\widehat{C_2(u_1,u_2,u_3)}(\xi) &= (2\pi)^{-1} \sum_{\substack{\xi=\xi_1+\xi_2+\xi_3 \\ \xi_1=\xi \text{ or } \xi_2=\xi}} \widehat{u}_1(\xi_1)\widehat{u}_2(\xi_2)\widehat{\overline{u}}_3(\xi_3),
\end{aligned}
$$

where we suppressed the t dependence. By Plancherel's identity we observe that

$$
C_2(u_1,u_2,u_3) = u_1 \fint_0^{2\pi} u_2\overline{u}_3 dy + u_2 \fint_0^{2\pi} u_1 \overline{u}_3 dy - u_1 * u_2 * \overline{u}_3,
$$

where $*$ denotes convolution with respect to normalized Lebesgue measure on $[0,2\pi]$. Clearly,

$$
\|C_2(u_1,u_2,u_3)\|_{L^2_t L^2_x} \lesssim \prod_{1\le k\le 3} \|u_k\|_{L^6_t L^2_x} \lesssim \prod_{1\le k\le 3} \|u_k\|_{X^{0,\frac13}}
$$

by Sobolev estimates in the time variable. For the convolution term we also used Young's inequality. So it remains to prove (38) with $\|u_1 u_2 \overline{u}_3\|_{L^2_{xt}}$ on the left-hand side replaced by $\|C_1(u_1,u_2,u_3)\|_{L^2_{xt}}$. Now by the Cauchy–Schwarz inequality and Fubini's theorem (cf. the arguments in the previous section), matters reduce to show that

$$\sup_{\xi,\tau} \Sigma(\xi,\tau) < \infty,$$

where $\Sigma(\xi,\tau)$ is defined as

$$\sum_{\substack{\xi=\xi_1+\xi_2+\xi_3\\ \xi_1\neq\xi,\,\xi_2\neq\xi}} \langle\xi_1\rangle^{-2s}\langle\xi_2\rangle^{-2s}\int \langle\tau_1+\xi_1^2\rangle^{-2b}\langle\tau_2+\xi_2^2\rangle^{-2b}\langle\tau_3-\xi_3^2\rangle^{-2b}d\tau_1 d\tau_2.$$

Using Lemma 4.1 twice, we see that

$$\begin{aligned}\Sigma(\xi,\tau) &\lesssim \sum_{\substack{\xi=\xi_1+\xi_2+\xi_3\\ \xi_1\neq\xi,\,\xi_2\neq\xi}} \langle\xi_1\rangle^{-2s}\langle\xi_2\rangle^{-2s}\langle\tau+\xi^2-2(\xi-\xi_1)(\xi-\xi_2)\rangle^{2-6b}\\ &\lesssim \left(\sum_{\substack{\xi=\xi_1+\xi_2+\xi_3\\ \xi_1\neq\xi,\,\xi_2\neq\xi}} \langle\xi_1\rangle^{0-}\langle\xi_2\rangle^{0-}\langle\tau+\xi^2-2(\xi-\xi_1)(\xi-\xi_2)\rangle^{\frac{2-6b}{1-2s}-}\right)^{(1-2s)+}\end{aligned}$$

by Hölder's inequality. Since by assumption $\frac{2-6b}{1-2s}<-1$, a final application of Corollary 3.2, part 3, completes the proof. $\square$

In the L^2_{xt}-norm on the left-hand side of (38) we may, of course, replace any single factor by its complex conjugate. Especially we have

$$\|u_1u_2\overline{u}_3\|_{L^2_{xt}} \lesssim \|u_1\|_{X^{0,b}}\|u_2\|_{X^{s,b}}\|u_3\|_{X^{s,b}}. \tag{39}$$

Fixing u_2 and u_3 and considering the linear operator

$$X^{0,b}\to L^2_{xt}: \quad u_1\mapsto u_1u_2\overline{u}_3$$

we obtain by duality the estimate

$$\|v\overline{u}_2u_3\|_{X^{0,-b}} \lesssim \|v\|_{L^2_{xt}}\|u_2\|_{X^{s,b}}\|u_3\|_{X^{s,b}}. \tag{40}$$

Choosing $v=u_1\overline{u}_4u_5$ and applying (39) (and (38), respectively) again, we have shown the following quintilinear estimate.

Corollary 5.2. *Set $i=1$ or $i=4$. For $\frac{1}{3}<b_0<\frac{1}{2}$ and $s_0>3(\frac{1}{2}-b_0)$ the estimate*

$$\|u_1\overline{u}_2u_3\overline{u}_4u_5\|_{X^{0,-b_0}} \lesssim \|u_i\|_{X^{0,b_0}}\prod_{\substack{k=1\\k\neq i}}^{5}\|u_k\|_{X^{s_0,b_0}} \tag{41}$$

is valid.

In order to prove Theorem 2.6 we shall rely on the interpolation properties of our scale of spaces obtained by the complex method.

Lemma 5.3. *Let $s_i,b_i\in\mathbb{R}$, $1<r_i,p_i<\infty$ for $i=1,2$. Then*

$$(X^{s_0,b_0}_{r_0,p_0},X^{s_1,b_1}_{r_1,p_1})_{[\theta]} = X^{s,b}_{r,p} \qquad (0<\theta<1),$$

where

$$s = (1-\theta)s_0 + \theta s_1, \quad b = (1-\theta)b_0 + \theta b_1,$$
$$\frac{1}{r} = \frac{1-\theta}{r_0} + \frac{\theta}{r_1}, \quad \frac{1}{p} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}.$$

Proof. The map

$$\mathcal{F} \circ e^{-it\partial_x^2} : X^{s,b}_{r,p} \to \ell^{r'}_\xi(\langle\xi\rangle^s; L^{p'}_\tau(\langle\tau\rangle^b))$$

is an isometric isomorphism. Here, the image space is the space of sequences in $\ell^{r'}_\xi$ with weight $\langle\xi\rangle^s$, taking values in $L^{p'}_\tau$ with weight $\langle\tau\rangle^b$ (with the natural norm). Now, arguing as in [1, Theorem 5.6.3] (replacing 2^k by k) and using [1, Theorem 5.5.3] the claim follows. ∎

Proof of Theorem 2.6. By Sobolev-type embeddings and Young's inequality, we see that for $r_1, q_1 > 1$, $s_1 > \frac{1}{q_1}$, $b_1 > \frac{1}{3}$, and an auxiliary exponent p with $b_1 + \frac{1}{2} > \frac{1}{p} > 5(\frac{1}{2} - b_1)$

$$\begin{aligned} \|u_1\overline{u}_2u_3\overline{u}_4u_5\|_{X^{0,-b_1}_{r_1,2}} &\lesssim \|u_1\overline{u}_2u_3\overline{u}_4u_5\|_{X^{0,0}_{r_1,p}} \\ &\leq \|u_1\|_{X^{0,0}_{r_1,5p}} \prod_{i=2}^{5} \|u_i\|_{X^{0,0}_{\infty,5p}} \\ &\lesssim \|u_1\|_{X^{0,b_1}_{r_1,2}} \prod_{i=2}^{5} \|u_i\|_{X^{s_1,b_1}_{q_1,2}}. \end{aligned} \tag{42}$$

Now fix $\frac{4}{3} < q \leq r \leq 2$ and $b > \frac{1}{6} + \frac{1}{3q}$. We will use complex multilinear[6] interpolation [1, Theorem 4.4.1] between (41) and (42) with interpolation parameter $\theta = \frac{1}{2}$. To do so, we choose

$$s_0 = \frac{3}{2} - \frac{2}{q} - \varepsilon, \quad b_0 = \frac{2}{3q} + \varepsilon$$

in the endpoint (41) and

$$\frac{1}{r_1} = \frac{2}{r} - \frac{1}{2}, \quad \frac{1}{q_1} = \frac{2}{q} - \frac{1}{2}, \quad s_1 = \frac{2}{q} - \frac{1}{2} + \varepsilon, \quad b_1 = \frac{1}{3} + \varepsilon$$

in the endpoint (42), where $\varepsilon := b - \frac{1}{6} - \frac{1}{3q} > 0$, and we use Lemma 5.3. As a result,

$$\|u_1\overline{u}_2u_3\overline{u}_4u_5\|_{X^{0,-b}_{r,2}} \lesssim \|u_1\|_{X^{0,b}_{r,2}} \prod_{i=2}^{5} \|u_i\|_{X^{\frac{1}{2},b}_{q,2}},$$

where in the last line of (42) as well as in the last expression we may exchange u_1 and u_4. Now our first claim (13) follows from $\langle\xi\rangle \leq \sum_{1\leq i\leq 5}\langle\xi_i\rangle$. We use (6) of Lemma 2.2, that is, $X^{\frac{1}{2},-\frac{1}{2}+}_{r,2} \subset X^{\frac{1}{2},-1}_{r,\infty}$, and apply (27) with $\nu = \frac{1}{3q'}-$ to all six norms appearing, which gives a factor $T^{\frac{2}{q'}-}$. Finally, (15) follows from (14), and the proof of Theorem 2.6 is complete. ∎

---

[6] Due to the fact that $\|\langle\xi\rangle^s\langle\tau+\xi^2\rangle^b\widehat{\overline{u}}\|_{\ell^r_\xi L^p_\tau} = \|\langle\xi\rangle^s\langle\tau-\xi^2\rangle^b\widehat{u}\|_{\ell^r_\xi L^p_\tau}$ it is equivalent to remove the complex conjugate on the second and fourth functions at the expense of changing signs in the weight of the respective norm. Of course, the analogue of Lemma 5.3 for the corresponding spaces is valid. With this modification the estimate becomes complex linear in each factor.

6. The gauge transform. In this subsection we study a nonlinear transformation which turned out to be a key ingredient to the well-posedness theory of (DNLS). This type of transformation for (DNLS) was already used by Hayashi and Ozawa [16, 17] and Hayashi [15], and later by Takaoka [25], in the nonperiodic case and then adapted to the periodic setting by the second author in [18, 19]. Let us define for $u \in C([-T,T], L^2(\mathbb{T}))$

$$\mathcal{G}_0 u := \exp\left(-i\mathcal{I}u\right) u, \tag{43}$$

where

$$\mathcal{I}u(t,x) := \fint_0^{2\pi} \int_\theta^x \left(|u(t,y)|^2 - \fint_0^{2\pi} |u(t,z)|^2 dz \right) dy d\theta \tag{44}$$

is the unique primitive of

$$x \mapsto |u(t,x)|^2 - \fint_0^{2\pi} |u(t,z)|^2 dz$$

with vanishing mean value.

Before we study the mapping properties of this transformation let us recall the Sobolev multiplication law in our setting.

LEMMA 6.1. *Let $1 < r < \infty$, $0 \le s \le s_1$, and $s_1 > \frac{1}{r}$. Then*

$$\|u_1 u_2\|_{\widehat{H}^s_r} \lesssim \|u_1\|_{\widehat{H}^{s_1}_r} \|u_2\|_{\widehat{H}^s_r}. \tag{45}$$

In particular, for $1 < r < \infty$, $s > \frac{1}{r}$ we have

$$\|u_1 u_2\|_{\widehat{H}^s_r} \lesssim \|u_1\|_{\widehat{H}^s_r} \|u_2\|_{\widehat{H}^s_r}. \tag{46}$$

Proof. We have $\langle \xi \rangle^s \lesssim \langle \xi - \xi_1 \rangle^s + \langle \xi_1 \rangle^{s_1} \langle \xi - \xi_1 \rangle^{s - s_1}$, and by Young's inequality

$$\begin{aligned}
\|u_1 u_2\|_{\widehat{H}^s_r} &\lesssim \|\widehat{u}_1\|_{\ell^1} \|\langle \xi \rangle^s \widehat{u}_2\|_{\ell^{r'}} + \|\langle \xi \rangle^{s_1} \widehat{u}_1\|_{\ell^{r'}} \|\langle \xi \rangle^{s-s_1} \widehat{u}_2\|_{\ell^1} \\
&\lesssim \|u_1\|_{\widehat{H}^{s_1}_r} \|u_2\|_{\widehat{H}^s_r}
\end{aligned}$$

because $\|\langle \xi \rangle^{-s_1} \widehat{u}_i\|_{\ell^1} \lesssim \|\widehat{u}_i\|_{\ell^{r'}}$, $i = 1, 2$. ∎

LEMMA 6.2. *Let $1 < r \le 2$ and $s > \frac{1}{r} - \frac{1}{2}$ or $r = 2$ and $s \ge 0$. Then*

$$\mathcal{G}_0 : C([-T,T], \widehat{H}^s_r(\mathbb{T})) \to C([-T,T], \widehat{H}^s_r(\mathbb{T}))$$

is a locally bi-Lipschitz homeomorphism with inverse $\mathcal{G}_0^{-1} u = \exp\left(i\mathcal{I}u\right) u$.

Proof. We will transfer the ideas from the proof of [18, Lemma 2.3] (where the claim is shown in the L^2 setting) to the case $1 < r < 2$ and $s > \frac{1}{r} - \frac{1}{2}$. Obviously, it suffices to prove

$$\begin{aligned}
&\|(\exp(\pm i\mathcal{I}f) - \exp(\pm i\mathcal{I}g))\, h\|_{\widehat{H}^s_r} \\
&\lesssim \exp(c\|f\|^2_{\widehat{H}^s_r} + c\|g\|^2_{\widehat{H}^s_r}) \|f - g\|_{\widehat{H}^s_r} \|h\|_{\widehat{H}^s_r}
\end{aligned} \tag{47}$$

for smooth $f, g, h \in \widehat{H}^s_r$. By the series expansion of the exponential and (45) we infer that the left-hand side of (47) is bounded by

$$\|h\|_{\widehat{H}^s_r} \|\mathcal{I}f - \mathcal{I}g\|_{\widehat{H}^{s_1}_r} \sum_{n=1}^\infty \frac{c^n}{n!} \sum_{k=0}^{n-1} \|\mathcal{I}f\|^k_{\widehat{H}^{s_1}_r} \|\mathcal{I}g\|^{n-1-k}_{\widehat{H}^{s_1}_r}$$

for $s_1 = \max\{\frac{1}{r}+, s\}$. By the definition of $\mathcal{I}$ it follows that

$$\|\mathcal{I}f - \mathcal{I}g\|_{\widehat{H}_r^{s_1}} \lesssim \||f|^2 - |g|^2\|_{\widehat{H}_r^{s_1-1}}.$$

In the case $s \le \frac{1}{r}$ we have

$$\||f|^2 - |g|^2\|_{\widehat{H}_r^{s_1-1}} \lesssim (\|\widehat{f}\|_{\ell^{2-}} + \|\widehat{g}\|_{\ell^{2-}})\|\widehat{f} - \widehat{g}\|_{\ell^{2-}}.$$

Because $s > \frac{1}{r} - \frac{1}{2}$ we find

$$\|\mathcal{I}f - \mathcal{I}g\|_{\widehat{H}_r^{s_1}} \lesssim \left(\|f\|_{\widehat{H}_r^s} + \|g\|_{\widehat{H}_r^s}\right)\|f - g\|_{\widehat{H}_r^s}.$$

In the case $s > \frac{1}{r}$ we use (46) to deduce

$$\||f|^2 - |g|^2\|_{\widehat{H}_r^{s_1-1}} \lesssim \|f - g\|_{\widehat{H}_r^s}(\|f\|_{\widehat{H}_r^s} + \|g\|_{\widehat{H}_r^s}),$$

and the estimate (47) follows. $\square$

The following lemma is contained in [19] in the $r = 2$ case.

LEMMA 6.3. *Let $1 < r \le 2$ and $s > \frac{1}{r} - \frac{1}{2}$ or $r = 2$ and $s \ge 0$. The translation operators*

$$\tau^{\mp} : C([-T,T], \widehat{H}_r^s(\mathbb{T})) \to C([-T,T], \widehat{H}_r^s(\mathbb{T})),$$
$$\tau^{\mp}u(t,x) := u\left(t, x \mp 2t \fint_0^{2\pi} |u(t,y)|^2 dy\right)$$

are continuous. However, their restrictions to arbitrarily small balls are not uniformly continuous.

Proof. We sketch only the main ideas and refer the reader to [19, Propositions 3.2.1 and 3.2.2] for details in the H^s case which easily carry over to the $\widehat{H}_r^s$ setting: Because the embedding $\widehat{H}_r^s \subset L^2$ is continuous, the continuity statement follows from the continuity of the map

$$\tau : \mathbb{R} \times C([-T,T], \widehat{H}_r^s(\mathbb{T})) \to C([-T,T], \widehat{H}_r^s(\mathbb{T})),$$
$$\tau(h,u)(t,x) := u(t, x + ht).$$

If we fix the time variable, the map

$$\mathbb{R} \times \widehat{H}_r^s(\mathbb{T}) \to \widehat{H}_r^s(\mathbb{T}), \quad (h,f) \mapsto f(\cdot + h)$$

is continuous. This follows from the fact that a translation by a fixed amount is an isometry in $\widehat{H}_r^s(\mathbb{T})$ combined with $e^{i\xi h} \to e^{i\xi h_0}$ for $h \to h_0$ and the dominated convergence theorem. Now, because $[-T,T]$ is compact, we may approximate $u \in C([-T,T], \widehat{H}_r^s(\mathbb{T}))$ uniformly by a piecewise constant (in time) function and apply the result for t fixed.

For $r > 0$, the sequences of functions

$$u_{n,j}(t,x) = u_{n,j}(x) = rn^{-s}e^{inx} + c_{n,j}, \quad n \in \mathbb{N},\ j = 1,2,$$

with $c_{n,1} = \frac{1}{\sqrt{n}}$ and $c_{n,2} = 0$ provide a counterexample to the uniform continuity on balls. $\square$

As in [18, 19] we define $\mathcal{G} = \mathcal{G}_0 \circ \tau^-$, i.e.,

$$\mathcal{G}u(t,x) = (\mathcal{G}_0 u)\left(t, x - 2t \fint_0^{2\pi} |u(t,y)|^2 dy\right). \tag{48}$$

LEMMA 6.4. *Let $u, v \in C([-T,T], H^2) \cap C^1((-T,T), L^2)$ such that $v = \mathcal{G}u$. Then u solves* (DNLS) *if and only if v solves*

$$\begin{aligned} i\partial_t v(t) + \partial_x^2 v(t) &= -i\mathcal{T}(v)(t) - \tfrac{1}{2}\mathcal{Q}(v)(t), \quad t \in (-T,T), \\ v(0) &= \mathcal{G}u(0), \end{aligned} \tag{49}$$

where

$$\mathcal{T}(v) = v^2 \partial_x \bar{v} - 2i \fint_0^{2\pi} \mathrm{Im}\,(v\partial_x \bar{v})\, dx\, v \tag{50}$$

and

$$\mathcal{Q}(v) = \left(|v|^4 - \fint_0^{2\pi} |v|^4 dx\right) v - 2\fint_0^{2\pi} |v|^2 dx \left(|v|^2 - \fint_0^{2\pi} |v|^2 dx\right) v, \tag{51}$$

i.e., $\mathcal{T}(v) = T(v,v,v)$ and $\mathcal{Q}(v) = Q(v,v,v,v,v)$ for T and Q defined in (8) *and* (12)*, respectively. Moreover, the map*

$$\mathcal{G} : C([-T,T], \widehat{H}^s_r(\mathbb{T})) \to C([-T,T], \widehat{H}^s_r(\mathbb{T}))$$

is a homeomorphism with inverse $\mathcal{G}^{-1} = \tau^+ \circ \mathcal{G}_0^{-1}$. The restrictions of $\mathcal{G}$ and $\mathcal{G}^{-1}$ to arbitrarily small balls fail to be uniformly continuous. However, $\mathcal{G}$ is locally bi-Lipschitz on subsets of functions with prescribed L^2-norm.

Proof. To see the equivalence of (DNLS) and (49) the calculations for the periodic case may be found in [18, section 2]. The fact that we may represent $\mathcal{T}$ and $\mathcal{Q}$ via convolution operators on the Fourier side where certain frequency interactions are cancelled out was remarked in [19, Remark 3.2.7]. The verification of the precise formulas are straightforward, using (suppressing the t dependence)

$$\begin{aligned} (2\pi)^{-1} \widehat{v} * \widehat{\partial_x \overline{v}}(0) &= i \fint_0^{2\pi} \mathrm{Im}\,(v\partial_x \overline{v})\, dx, \\ (2\pi)^{-1} \widehat{v} * \widehat{\overline{v}}(0) &= \fint_0^{2\pi} |v|^2\, dx, \\ (2\pi)^{-2} \widehat{v} * \widehat{\overline{v}} * \widehat{v} * \widehat{\overline{v}}(0) &= \fint_0^{2\pi} |v|^4\, dx. \end{aligned}$$

The mapping properties follow from Lemma 6.2. ∎

Remark 4. The cancellation of certain frequency interactions due to the term

$$2\fint_0^{2\pi} \mathrm{Im}\,(v\partial_x \overline{v})\, dx v,$$

which is crucial for our arguments (cf. (8)), is an important feature of the gauge transformation. We observe as well that this expression itself is not well defined in $\widehat{H}^{\frac{1}{2}}_r(\mathbb{T})$ for $1 < r < 2$.

7. Proof of well-posedness. Now, we show that the Cauchy problem (49) is well-posed. Let $\chi \in C_0^\infty(\mathbb{R})$ be nonnegative and symmetric such that $\chi(t) = 0$ for $|t| \geq 2$ and $\chi(t) = 1$ for $|t| \leq 1$. Recall that $Z_r^s := X_{r,2}^{s,\frac{1}{2}} \cap X_{r,\infty}^{s,0}$. We have, similar to the L^2 case, the following linear estimates.

LEMMA 7.1. *Let $s \in \mathbb{R}$, $1 < r < \infty$.*

$$\|\chi S u_0\|_{Z_r^s} \lesssim \|u_0\|_{\widehat{H}_r^s}, \tag{52}$$

$$\left\| \chi \int_0^t S(t-t')u(t')dt' \right\|_{Z_r^s} \lesssim \|u\|_{X_{r,2}^{s,-\frac{1}{2}}} + \|u\|_{X_{r,\infty}^{s,-1}}. \tag{53}$$

Proof. We use the approach from [9, Lemma 3.1]. Let $u_0 \in C^\infty(\mathbb{R})$ be periodic. We calculate $\mathcal{F}(\chi S u_0)(\tau,\xi) = \mathcal{F}_t\chi(\tau+\xi^2)\mathcal{F}_x u_0(\xi)$, and (52) follows because $\mathcal{F}_t\chi$ is rapidly decreasing. It suffices to consider smooth u with $\operatorname{supp}(u) \subset \{(t,x) \mid |t| \leq 2\}$. We rewrite

$$\chi(t) \int_0^t S(t-t')f(t')\,dt' = F_1(t) + F_2(t),$$

where

$$F_1(t) = \frac{1}{2}\chi(t)S(t)\int_{\mathbb{R}} \varphi(t')S(-t')u(t')\,dt',$$
$$F_2(t) = \frac{1}{2}\chi(t)\int_{\mathbb{R}} \varphi(t-t')S(t-t')u(t')\,dt',$$

and $\varphi(t') = \chi(t'/10)\operatorname{sign}(t')$. Concerning φ we have

$$|\mathcal{F}_t\varphi(\tau)| \lesssim \langle\tau\rangle^{-1}. \tag{54}$$

Estimate (52) yields

$$\|F_1\|_{Z_r^s} \lesssim \left\| \int_{\mathbb{R}} \varphi(t')S(-t')u(t')\,dt' \right\|_{\widehat{H}_r^s}.$$

Parseval's equality implies that

$$\mathcal{F}_x\left(\int_{\mathbb{R}} \varphi(t')S(-t')u(t')\,dt'\right)(\xi) = \int_{\mathbb{R}} \overline{\mathcal{F}_t\varphi}(\tau+\xi^2)\mathcal{F}u(\tau,\xi)\,d\tau,$$

which gives

$$\left\| \int_{\mathbb{R}} \varphi(t')S(-t')u(t')\,dt' \right\|_{\widehat{H}_r^s} \lesssim \|u\|_{X_{r,\infty}^{s,-1}}$$

by (54). Now, let us consider F_2. Due to Young's inequality, we may remove the cutoff function χ in front of the integral. The Fourier transform of the remainder is given by

$$\mathcal{F}\left(\int_{\mathbb{R}} \varphi(t-t')S(t-t')u(t')\,dt'\right)(\tau,\xi) = \mathcal{F}_t\varphi(\tau+\xi^2)\mathcal{F}u(\tau,\xi).$$

Estimate (54) proves (53). ∎

A standard application of the fixed point argument gives the following.

THEOREM 7.2. *Let $\frac{4}{3} < q \leq r \leq 2$. Then for every*

$$v_0 \in B_R := \left\{ v_0 \in \widehat{H}^{\frac{1}{2}}_r(\mathbb{T}) \mid \|v_0\|_{\widehat{H}^{\frac{1}{2}}_q} < R \right\}$$

and $T \lesssim R^{-2q'-}$ there exists a solution

$$v \in Z^{\frac{1}{2}}_r(T) \subset C([-T,T], \widehat{H}^{\frac{1}{2}}_r(\mathbb{T}))$$

of the Cauchy problem (49)*. This solution is unique in the space $Z^{\frac{1}{2}}_q(T)$, and the map*

$$\left(B_R, \|\cdot\|_{\widehat{H}^{\frac{1}{2}}_r} \right) \longrightarrow C([-T,T], \widehat{H}^{\frac{1}{2}}_r(\mathbb{T})) : \quad v_0 \mapsto v$$

is locally Lipschitz continuous. Moreover, it is real analytic.

Sketch of proof. As a consequence of the estimates (53), (9), (10), and (15),

$$\Phi(v)(t) := \int_0^t e^{i(t-t')\partial_x^2} \left(-\tfrac{1}{2}\mathcal{Q} - i\mathcal{T}\right) v(t')\, dt'$$

extends to a continuous map $\Phi : Z^{\frac{1}{2}}_r(T) \to Z^{\frac{1}{2}}_r(T)$ for $\frac{4}{3} < r \leq 2$ along with the estimate

$$\begin{aligned} \|\Phi(v_1) - \Phi(v_2)\|_{Z^{\frac{1}{2}}_r(T)} \lesssim{}& T^{\frac{1}{r'}-} \left(\|v_1\|_{Z^{\frac{1}{2}}_r(T)} + \|v_2\|_{Z^{\frac{1}{2}}_r(T)} \right)^2 \|v_1 - v_2\|_{Z^{\frac{1}{2}}_r(T)} \\ &+ T^{\frac{2}{r'}-} \left(\|v_1\|_{Z^{\frac{1}{2}}_r(T)} + \|v_2\|_{Z^{\frac{1}{2}}_r(T)} \right)^4 \|v_1 - v_2\|_{Z^{\frac{1}{2}}_r(T)}, \end{aligned} \tag{55}$$

and with (52) we also have

$$\|e^{it\partial_x^2} v_0 + \Phi(v)\|_{Z^{\frac{1}{2}}_r(T)} \lesssim \|v_0\|_{\widehat{H}^{\frac{1}{2}}_{r_1}} + T^{\frac{1}{r'}-} \|v\|^3_{Z^{\frac{1}{2}}_r(T)} + T^{\frac{2}{r'}-} \|v\|^5_{Z^{\frac{1}{2}}_r(T)}.$$

Hence, for fixed v_0 the operator $e^{it\partial_x^2} v_0 + \Phi : D \to D$ is a strict contraction in some closed ball $D \subset Z^{\frac{1}{2}}_r(T)$ for small enough T. By the contraction mapping principle we find a fixed point $v \in Z^{\frac{1}{2}}_r(T)$ which is a solution of (49) for small times. Similarly, the implicit function theorem shows that the map $v_0 \mapsto v$ is real analytic, hence locally Lipschitz. Uniqueness in $Z^{\frac{1}{2}}_q(T)$ follows by contradiction: A translation in time reduces matters to uniqueness for an arbitrarily short time interval which follows from the estimate (55) with $r = q$. The lower bound on the maximal time of existence is a consequence of the mixed estimate

$$\|v\|_{Z^{\frac{1}{2}}_r(T)} \lesssim \|v(0)\|_{\widehat{H}^{\frac{1}{2}}_r} + T^{\frac{1}{q'}-} \|v\|^2_{Z^{\frac{1}{2}}_q(T)} \|v\|_{Z^{\frac{1}{2}}_r(T)} + T^{\frac{2}{q'}-} \|v\|^4_{Z^{\frac{1}{2}}_q(T)} \|v\|_{Z^{\frac{1}{2}}_r(T)}$$

for solutions v and an iteration argument. □

By combining Theorem 7.2 with Lemma 6.4 and some approximation arguments, Theorem 1.2 follows (this is carried out in detail in [13] for the nonperiodic case). In particular, the claim concerning the nonuniform continuity is a consequence of the properties of $\mathcal{G}$ and of the flow map of (49).

Remark 5. In fact, our estimates also imply uniqueness of solutions of (49) in a restriction space based on the $X_{q,2}^{\frac{1}{2},\frac{1}{2}}$ component only. Hence, the optimal uniqueness statement concerning solutions of (DNLS) provided by our methods is the following: Let $4/3 < q \leq 2$ and $u_1, u_2 \in C([-T,T], \widehat{H}_q^{\frac{1}{2}}(\mathbb{T}))$ be solutions of (DNLS) with $u_1(0) = u_2(0)$. If additionally $\mathcal{G}u_1, \mathcal{G}u_2 \in X_{q,2}^{\frac{1}{2},\frac{1}{2}}$, which also satisfy (49), then $u_1 = u_2$.

Acknowledgment. The authors would like to thank the anonymous referees for their valuable comments.

REFERENCES

[1] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Grundlehren Math. Wiss. 223, Springer-Verlag, Berlin, New York, 1976.

[2] H. A. Biagioni and F. Linares, *Ill-posedness for the derivative Schrödinger and generalized Benjamin-Ono equations*, Trans. Amer. Math. Soc., 353 (2001), pp. 3649–3659.

[3] J. Bourgain, *Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations.* I. *Schrödinger equations*, Geom. Funct. Anal., 3 (1993), pp. 107–156.

[4] T. Cazenave, L. Vega, and M. C. Vilela, *A note on the nonlinear Schrödinger equation in weak L^p spaces*, Commun. Contemp. Math., 3 (2001), pp. 153–162.

[5] M. Christ, *Nonuniqueness of Weak Solutions of the Nonlinear Schrödinger Equation*, http://arxiv.org/abs/math/0503366.

[6] M. Christ, *Power series solution of a nonlinear Schrödinger equation*, in Mathematical Aspects of Nonlinear Dispersive Equations, Ann. of Math. Stud. 163, Princeton University Press, Princeton, NJ, 2007, pp. 131–155.

[7] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Global well-posedness for Schrödinger equations with derivative*, SIAM J. Math. Anal., 33 (2001), pp. 649–669.

[8] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *A refined global well-posedness result for Schrödinger equations with derivative*, SIAM J. Math. Anal., 34 (2002), pp. 64–86.

[9] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Multilinear estimates for periodic KdV equations, and applications*, J. Funct. Anal., 211 (2004), pp. 173–218.

[10] D. De Silva, N. Pavlovic, G. Staffilani, and N. Tzirakis, *Global well-posedness for a periodic nonlinear Schrödinger equation in* 1*D and* 2*D*, Discrete Contin. Dyn. Syst., 19 (2007), pp. 37–65.

[11] J. Ginibre, Y. Tsutsumi, and G. Velo, *On the Cauchy problem for the Zakharov system*, J. Funct. Anal., 151 (1997), pp. 384–436.

[12] A. Grünrock, *An improved local well-posedness result for the modified KdV equation*, Int. Math. Res. Not., 2004 (2004), pp. 3287–3308.

[13] A. Grünrock, *Bi- and trilinear Schrödinger estimates in one space dimension with applications to cubic NLS and DNLS*, Int. Math. Res. Not., 2005 (2005), pp. 2525–2558.

[14] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 5th ed., Clarendon Press, Oxford University Press, New York, 1979.

[15] N. Hayashi, *The initial value problem for the derivative nonlinear Schrödinger equation in the energy space*, Nonlinear Anal., 20 (1993), pp. 823–833.

[16] N. Hayashi and T. Ozawa, *On the derivative nonlinear Schrödinger equation*, Phys. D, 55 (1992), pp. 14–36.

[17] N. Hayashi and T. Ozawa, *Finite energy solutions of nonlinear Schrödinger equations of derivative type*, SIAM J. Math. Anal., 25 (1994), pp. 1488–1503.

[18] S. Herr, *On the Cauchy problem for the derivative nonlinear Schrödinger equation with periodic boundary condition*, Int. Math. Res. Not., 2006 (2006), article 96763.

[19] S. Herr, *Well-Posedness Results for Dispersive Equations with Derivative Nonlinearities*, Dissertation, Universität Dortmund, Dortmund, Germany, 2006.

[20] L. Hörmander, *The Analysis of Linear Partial Differential Operators.* II, Grundlehren Math. Wiss. 257, Springer-Verlag, Berlin, 1983.

[21] D. J. Kaup and A. C. Newell, *An exact solution for a derivative nonlinear Schrödinger equation*, J. Math. Phys., 19 (1978), pp. 798–801.

[22] C. E. Kenig, G. Ponce, and L. Vega, *Quadratic forms for the* 1*-D semilinear Schrödinger equation*, Trans. Amer. Math. Soc., 348 (1996), pp. 3323–3353.

[23] A. Moyua and L. Vega, *Bounds for the maximal function associated to periodic solutions of 1D dispersive equations*, Bull. London Math. Soc., to appear.
[24] S. Selberg, *On an estimate for the wave equation and applications to nonlinear problems*, Differential Integral Equations, 15 (2002), pp. 213–236.
[25] H. Takaoka, *Well-posedness for the one-dimensional nonlinear Schrödinger equation with the derivative nonlinearity*, Adv. Differential Equations, 4 (1999), pp. 561–580.
[26] A. Vargas and L. Vega, *Global wellposedness for 1D nonlinear Schrödinger equation for data with an infinite L^2 norm*, J. Math. Pures Appl. (9), 80 (2001), pp. 1029–1044.
[27] A. Zygmund, *On Fourier coefficients and transforms of functions of two variables*, Studia Math., 50 (1974), pp. 189–201.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 1921–1950

LOWER SEMICONTINUITY OF QUASI-CONVEX BULK ENERGIES IN SBV AND INTEGRAL REPRESENTATION IN DIMENSION REDUCTION*

JEAN-FRANÇOIS BABADJIAN†

Abstract. A result of Larsen concerning the structure of the approximate gradient of certain sequences of functions with bounded variation is used to present a short proof of Ambrosio's lower semicontinuity theorem for quasi-convex bulk energies in SBV. It enables us to generalize to the SBV setting the decomposition lemma for scaled gradients in dimension reduction and also to show that, from the point of view of bulk energies, SBV dimensional reduction problems can be reduced to analog ones in the Sobolev spaces framework.

Key words. dimension reduction, Γ-convergence, functions of bounded variation, free discontinuity problems, quasi convexity, equi-integrability

AMS subject classifications. 74K35, 49J45, 49Q20

DOI. 10.1137/060676416

1. Introduction. Since the pioneering work [22], the modeling of thin films through dimensional reduction techniques and Γ-convergence analysis has become one of the main issues in the field of the calculus of variations. In the membrane theory framework in nonlinear elasticity, the problem rests on the study of the (scaled) elastic energy

$$\frac{1}{\varepsilon}\int_{\Omega_\varepsilon} W(\varepsilon)(y,\nabla v)\,dy$$

of such bodies. Here $\Omega_\varepsilon := \omega \times (-\varepsilon/2, \varepsilon/2)$, where ω is a bounded open subset of $\mathbb{R}^2$ and $\varepsilon > 0$, stands for the reference configuration of a nonlinear elastic thin film, $v : \Omega_\varepsilon \to \mathbb{R}^3$ is the deformation field which maps the reference configuration into a deformed configuration, and $W(\varepsilon) : \Omega_\varepsilon \times \mathbb{R}^{3\times 3} \to [0,+\infty)$ is the stored energy density of the body which is a Carathéodory function satisfying uniform p-growth and p-coercivity conditions (with $1 < p < \infty$). From a mathematical point of view, the previous energy is well defined, provided v is a Sobolev function in $W^{1,p}(\Omega_\varepsilon;\mathbb{R}^3)$.

To study the limit problem as the thickness $\varepsilon \to 0$, it will be useful to recast the energy functional over the varying set Ω_ε into a functional with a fixed domain of integration $\Omega := \omega \times (-1/2, 1/2)$. To this end, denoting by $x_\alpha := (x_1, x_2)$ the in-plane variable, we set $u(x_\alpha, x_3) := v(x_\alpha, \varepsilon\, x_3)$ so that, after the (now standard) change of variables

$$x_\alpha = y_\alpha, \quad x_3 = \frac{y_3}{\varepsilon},$$

we are equivalently led to study the following rescaled functional:

$$\int_\Omega W_\varepsilon\left(x, \nabla_\alpha u \Big| \frac{1}{\varepsilon}\nabla_3 u\right) dx, \quad u \in W^{1,p}(\Omega;\mathbb{R}^3), \tag{1.1}$$

*Received by the editors December 1, 2006; accepted for publication (in revised form) November 6, 2007; published electronically February 29, 2008. This research has been supported by the MULTIMAT Marie Curie Research Training Network MRTN-CT-2004-505226 "Multi-scale modelling and characterisation for phase transformations in advanced materials."
http://www.siam.org/journals/sima/39-6/67641.html
†SISSA, Via Beirut 2-4, 34014 Trieste, Italy (babadjia@sissa.it).

where $W_\varepsilon : \Omega \times \mathbb{R}^{3\times 3} \to [0, +\infty)$ is the rescaled stored energy density expressed in the new variables and defined by $W_\varepsilon(x_\alpha, x_3, \xi) := W(\varepsilon)(x_\alpha, \varepsilon\, x_3, \xi)$. From now on, ∇_α (resp., ∇_3) will stand for the (approximate) gradient with respect to x_α (resp., x_3), $\xi = (\xi_\alpha|\xi_3)$ for some matrix $\xi \in \mathbb{R}^{3\times 3}$, and $z = (z_\alpha|z_3)$ for some vector $z \in \mathbb{R}^3$. Thus in view of the p-growth of the energy, it is important to understand the structure of what we call the scaled gradient of u, i.e.,

$$\left(\nabla_\alpha u \Big| \frac{1}{\varepsilon}\nabla_3 u\right). \tag{1.2}$$

In particular, if $\{u_\varepsilon\} \subset W^{1,p}(\Omega; \mathbb{R}^3)$ is a minimizing sequence uniformly bounded in energy, up to a subsequence, there always exist $u \in W^{1,p}(\Omega; \mathbb{R}^3)$ such that $D_3 u = 0$ in the sense of distributions and $b \in L^p(\Omega; \mathbb{R}^3)$ such that $u_\varepsilon \rightharpoonup u$ in $W^{1,p}(\Omega; \mathbb{R}^3)$ and $(1/\varepsilon)\nabla_3 u_\varepsilon \rightharpoonup b$ in $L^p(\Omega; \mathbb{R}^3)$. The limit function u is nothing but the deformation of the midplane, while b is called the *Cosserat vector*. It thus seems natural to expect a limit model depending on the pair (u, b). Unfortunately, this is still out of reach, and we refer the reader to [19] for a more detailed discussion on the subject. However, in [9] (see also [7]) a simplified model has been considered taking into account the *bending moment* $\bar{b} \in L^p(\omega; \mathbb{R}^3)$, i.e., the average in the transverse direction x_3 of b, instead of the full Cosserat vector field.

In the framework of fracture mechanics, one usually adds a surface energy term, penalizing the presence of the crack. The simplest case consists of just penalizing its area leading to the so-called Griffith surface energy. Thus, for a given crack, one should study the energy given by the competing sum of the bulk and the surface energies. Such fracture mechanics problems belong (among others) to the class of free discontinuity problems, that is, variational problems where the unknown is not only a function but a pair set/function. Based on the idea that the deformation may be discontinuous across the crack, it is convenient to study the weak formulation, replacing the crack by the jump set of the deformation and leading to a variational problem stated in the space of (special) functions with bounded variation. Now the energy in which we are interested is

$$\frac{1}{\varepsilon}\int_{\Omega_\varepsilon} W(\varepsilon)(y, \nabla v)\, dy + \frac{1}{\varepsilon}\mathcal{H}^2(S_v), \quad v \in SBV^p(\Omega_\varepsilon; \mathbb{R}^3),$$

where ∇v is intended as the approximate gradient of v, S_v is the jump set of v, and $\mathcal{H}^2$ stands for the 2-dimensional Hausdorff measure. Writing as before, this energy in the rescaled variables yields

$$\int_\Omega W_\varepsilon\left(x, \nabla_\alpha u \Big| \frac{1}{\varepsilon}\nabla_3 u\right) dx + \int_{S_u} \left|\left((\nu_u)_\alpha \Big| \frac{1}{\varepsilon}(\nu_u)_3\right)\right| d\mathcal{H}^2, \quad u \in SBV^p(\Omega; \mathbb{R}^3), \tag{1.3}$$

where ν_u is the generalized normal to S_u and (1.2) is now referred to as the approximate scaled gradient of u.

The aim of this paper is to study the connections between variational problems (1.1) and (1.3), possibly taking into account the presence of the bending moment vector field. To this end, we will use Theorem 4.1 as the main ingredient, which extends the decomposition lemma for scaled gradients (see [8, Theorem 1.1] or [13, Theorem 3.1]) to the SBV setting. It states that any SBV sequence with bounded rescaled bulk energy and whose derivative's singular part behaves asymptotically well can be energetically replaced, up to a set of vanishing Lebesgue measure, by a sequence of Lipschitz maps whose scaled gradient is p-equi-integrable. Thus it reduces the free discontinuity problem to a usual dimensional reduction one in the framework of Sobolev spaces. This result is nothing but a rescaled version of [21, Lemma 2.1] (see also Theorem 3.1 below). Using this structure theorem, we are able to show two integral representation theorems in SBV (Theorems 6.1 and 7.3) which say that, up to a subsequence, the functional (1.3) Γ-converges (in an appropriate topology) to a functional of the same kind, i.e., the sum of a bulk and a surface energy. Moreover, the surface energy is still of Griffith's type, while the bulk energy is exactly the same as that obtained in the analog Sobolev spaces analysis. The main importance of these representation theorems relies on the fact that results on dimension reduction in Sobolev spaces can now be extended to SBV (see [5, 6, 7, 9, 22]).

Note that an integral representation result for dimensional reduction problems in SBV already exists (see [11, Theorem 2.1] and also [10]). Even if this reference may seem more general from the point of view of the hypothesis, it does not contain as a special case our results because the authors made strong use of the fact that their surface energy had to grow linearly with respect to the deformation jump. This assumption was essential in order to get compactness in $BV(\Omega;\mathbb{R}^3)$ of minimizing sequences. However, they suggested a way to remove that constraint by singular perturbation [11, Remark 2.2]. In our study we use a direct argument based on a trick introduced in [18] and which was already used in [4] in the framework of dimensional reduction. It consists of defining an artificial functional exactly as we usually do for the Γ-$\liminf$, except that we impose the minimizing sequences to be uniformly bounded in $L^\infty(\Omega;\mathbb{R}^3)$. Thanks to a truncation argument (see Lemma 6.2) we show that it actually coincides with the Γ-$\liminf$ for deformations $u \in L^\infty(\Omega;\mathbb{R}^3)$, and the advantage is that now minimizing sequences turn out to be relatively compact in $SBV(\Omega;\mathbb{R}^3)$ thanks to Ambrosio's compactness theorem. We refer the reader to [4] for a deeper insight on that subject.

To close this introduction, we wish to stress that in this paper, we are mostly interested in representation of effective bulk energies arising in 3D-2D dimensional reduction problems stated in SBV. For this reason we will consider a large class for such bulk energies, while surface energies will be restricted to the simplified case of a Griffith-type one. However, we are convinced that the results presented here could be generalized to a larger class of surface energies.

The overall plan of the paper is as follows: after recalling some useful notations in section 2, and in order to show the technique in a more transparent way, we present in section 3 a short proof of Ambrosio's lower semicontinuity result for quasi-convex integrands using [21, Lemma 2.1]. Then in section 4 we prove our main tool, Theorem 4.1, thanks to a slicing argument together with [21, Lemma 2.1]. To reach our goal, we need to prove a general integral representation for the Γ-limit of (1.1) in $W^{1,p}(\Omega;\mathbb{R}^3) \times L^p(\omega;\mathbb{R}^3)$ as a function of the deformation and the bending moment. This is the purpose of Theorem 5.1 in section 5 which contains as particular cases [9, Theorem 3.1] (with $W_\varepsilon(x,\xi) = W(\xi)$) and [7, Theorem 3.4] (with $W_\varepsilon(x,\xi) = W(x,\xi)$). In section 6, we refine the analysis of section 3, adding the difficulties of dimension reduction. From the integral representation in Sobolev spaces, Theorem 5.1, we deduce an analog result in SBV, Theorem 6.1, which says that the Γ-limit of (1.3) in $BV(\Omega;\mathbb{R}^3) \times L^p(\omega;\mathbb{R}^3)$ also has an integral representation and that the bulk energy density is exactly the same one as that obtained in the $W^{1,p}$ analysis. This will be achieved thanks to Theorem 4.1 and a blow-up method which enables us to reduce the problem to affine deformations and constant bending moments. Finally, we deduce a similar result in section 7 without the presence of the bending moment.

2. Notations and preliminaries. If $\Omega \subset \mathbb{R}^N$ is an open set, we consider the Lebesgue spaces $L^p(\Omega;\mathbb{R}^d)$ and the Sobolev spaces $W^{1,p}(\Omega;\mathbb{R}^d)$ in the usual way. When needed, we will make precise what topology the space $L^p(\Omega;\mathbb{R}^d)$ will be endowed. In particular we will denote by $L^p_s(\Omega;\mathbb{R}^d)$ (resp., $L^p_w(\Omega;\mathbb{R}^d)$) the space $L^p(\Omega;\mathbb{R}^d)$ endowed with the strong (resp., weak) topology. Strong convergence will always be denoted by $\to$, while weak (resp., weak*) convergence will be denoted by $\rightharpoonup$ (resp., $\overset{*}{\rightharpoonup}$).

We denote by $\mathcal{M}(\Omega;\mathbb{R}^d)$ the space of vector valued finite Radon measures. If $\mu \in \mathcal{M}(\Omega;\mathbb{R}^d)$ and E is a Borel subset of Ω, we will write $\mu \llcorner E$ for the restriction of μ to E; that is, for every Borel subset F of Ω, $\mu \llcorner E\,(F) = \mu(E \cap F)$. The Lebesgue measure in $\mathbb{R}^N$ will be denoted by $\mathcal{L}^N$, while $\mathcal{H}^{N-1}$ is the $(N-1)$-dimensional Hausdorff measure. We will denote by B the unit ball of $\mathbb{R}^N$ and by $\omega_N := \mathcal{L}^N(B)$ its Lebesgue measure. If $x_0 \in \mathbb{R}^N$ and $\rho > 0$, $B(x_0,\rho) := x_0 + \rho\, B$ is the ball centered at x_0 with radius ρ. The notation $-\!\!\!\!\!\!\int_A$ stands for the average $\mathcal{L}^N(A)^{-1}\int_A$.

The space of functions of bounded variation is denoted by $BV(\Omega;\mathbb{R}^d)$, and we refer the reader to [3] for standard theory of BV functions. We recall here a few facts: if $u \in BV(\Omega;\mathbb{R}^d)$, then its distributional derivative $Du \in \mathcal{M}(\Omega;\mathbb{R}^{d\times N})$, and thanks to Lebesgue's decomposition theorem, we can write $Du = D^a u + D^s u$, where $D^a u$ and $D^s u$ stand for, respectively, the absolutely continuous and singular part of Du with respect to the Lebesgue measure $\mathcal{L}^N$. Let S_u be the complementary of Lebesgue points of u. We say that u is a special function of bounded variation, and we write $u \in SBV(\Omega;\mathbb{R}^d)$, if

$$Du = \nabla u\, \mathcal{L}^N + (u^+ - u^-) \otimes \nu_u\, \mathcal{H}^{N-1} \llcorner S_u,$$

where ∇u is the approximate gradient of u, ν_u is the generalized normal to S_u, and $u^\pm$ are the traces of u on both sides of S_u. If $E \subset \Omega$, we say that E has finite perimeter in Ω, provided $\chi_E \in SBV(\Omega)$. We denote by $\partial^* E$ (resp., $\partial_* E$) the reduced (resp., essential) boundary of E. When $p > 1$, we define

$$SBV^p(\Omega;\mathbb{R}^d) := \Big\{ u \in SBV(\Omega;\mathbb{R}^d) : \nabla u \in L^p(\Omega;\mathbb{R}^{d\times N}) \text{ and } \mathcal{H}^{N-1}(S_u \cap \Omega) < +\infty \Big\}.$$

We say that a sequence $\{u_n\} \subset SBV^p(\Omega;\mathbb{R}^d)$ converges weakly to $u \in SBV^p(\Omega;\mathbb{R}^d)$, and we write $u_n \rightharpoonup u$ in $SBV^p(\Omega;\mathbb{R}^d)$, if

$$\begin{cases} u_n \to u \quad \text{in } L^1(\Omega;\mathbb{R}^d), \\ \nabla u_n \rightharpoonup \nabla u \quad \text{in } L^p(\Omega;\mathbb{R}^{d\times N}), \\ (u_n^+ - u_n^-) \otimes \nu_{u_n} \mathcal{H}^{N-1} \llcorner S_{u_n} \overset{*}{\rightharpoonup} (u^+ - u^-) \otimes \nu_u \mathcal{H}^{N-1} \llcorner S_u \quad \text{in } \mathcal{M}(\Omega;\mathbb{R}^{d\times N}). \end{cases}$$

If $\Omega := \omega \times I$, where ω is a bounded open subset of $\mathbb{R}^2$ and $I := (-1/2, 1/2)$, we will identify the spaces $L^p(\omega;\mathbb{R}^3)$, $W^{1,p}(\omega;\mathbb{R}^3)$, or $SBV^p(\omega;\mathbb{R}^3)$ with the space of functions $v \in L^p(\Omega;\mathbb{R}^3)$, $W^{1,p}(\Omega;\mathbb{R}^3)$, or $SBV^p(\Omega;\mathbb{R}^3)$ such that $D_3 v = 0$ in the sense of distributions.

By $\mathcal{A}(\omega)$ we mean the family all open subsets of ω, while $\mathcal{R}(\omega)$ stands for the countable subfamily of $\mathcal{A}(\omega)$ obtained by taking all finite unions of open cubes contained in ω, centered at rational points and with rational edge length.

In what follows, we will denote by $Q' := (-1/2, 1/2)^2$ the unit cube of $\mathbb{R}^2$ and by $Q'(x_0,\rho) := x_0 + \rho\, Q'$ the cube centered at $x_0 \in \mathbb{R}^2$ and side length $\rho > 0$. Similarly $B' := \{x_\alpha \in \mathbb{R}^2 : |x_\alpha| < 1\}$ stands for the unit ball in $\mathbb{R}^2$, and $B'(x_0,\rho) := x_0 + \rho\, B'$ denotes the ball of $\mathbb{R}^2$ centered at $x_0 \in \mathbb{R}^2$ and of radius $\rho > 0$.

3. Lower semicontinuity of quasi-convex bulk energies in $\boldsymbol{SBV}$. This section is devoted to give a short proof of Ambrosio's lower semicontinuity result for quasi-convex bulk energies in SBV using the following theorem proved in [21, Lemma 2.1].

THEOREM 3.1. *Let $\Omega \subset \mathbb{R}^N$ be a bounded open set with Lipschitz boundary and let $\{u_n\} \subset BV(\Omega;\mathbb{R}^d)$ be such that*

$$\begin{cases} \sup_{n\in\mathbb{N}} \|u_n\|_{BV(\Omega;\mathbb{R}^d)} < +\infty, \\ \sup_{n\in\mathbb{N}} \|\nabla u_n\|_{L^p(\Omega;\mathbb{R}^{d\times N})} < +\infty \quad \text{for some } p>1, \\ |D^s u_n|(\Omega) \to 0. \end{cases}$$

Then there exists a subsequence $\{n_k\} \nearrow +\infty$ and a sequence $\{w_k\} \subset W^{1,\infty}(\Omega;\mathbb{R}^d)$ such that

$$\begin{cases} \sup_{k\in\mathbb{N}} \|w_k\|_{W^{1,p}(\Omega;\mathbb{R}^d)} < +\infty, \\ \{|\nabla w_k|^p\} \text{ is equi-integrable}, \\ \mathcal{L}^N(\{w_k \neq u_{n_k}\} \cup \{\nabla w_k \neq \nabla u_{n_k}\}) \to 0. \end{cases}$$

This theorem is nothing but the BV counterpart of the decomposition lemma, [20, Lemma 1.2], in Sobolev spaces. We now use the previous result to give a short proof of Ambrosio's lower semicontinuity result for quasi-convex bulk energies in SBV (see [2, Theorem 4.3] or [3, Proposition 5.29]). This will enable us to emphasize the techniques used in this paper, occulting the difficulties of dimension reduction. The same kind of arguments will be used in section 6 to prove the lower bound of Theorem 6.1.

THEOREM 3.2. *Let Ω be bounded open subset of $\mathbb{R}^N$ and $f : \Omega \times \mathbb{R}^d \times \mathbb{R}^{d\times N} \to [0,+\infty)$ be a Carathéodory function satisfying, for all $(s,\xi) \in \mathbb{R}^d \times \mathbb{R}^{d\times N}$ and a.e. $x \in \Omega$,*

$$c|\xi|^p \le f(x,s,\xi) \le a(x) + \psi(|s|)(1+|\xi|^p) \tag{3.1}$$

for some $p > 1$, $c > 0$, $a \in L^1(\Omega)$ and some increasing function $\psi : [0,+\infty) \to [0,+\infty)$. If $\xi \mapsto f(x,s,\xi)$ is quasi-convex for every $s \in \mathbb{R}^d$ and a.e. $x \in \Omega$, then

$$\liminf_{n\to+\infty} \int_\Omega f(x,u_n,\nabla u_n)\,dx \ge \int_\Omega f(x,u,\nabla u)\,dx$$

for any sequence $\{u_n\} \subset SBV(\Omega;\mathbb{R}^d)$ converging in $L^1(\Omega;\mathbb{R}^d)$ to $u \in SBV(\Omega;\mathbb{R}^d)$ and satisfying $\sup_n \mathcal{H}^{N-1}(S_{u_n}) < +\infty$.

Proof. The proof is divided into three steps. We first apply the blow-up method to reduce the study to an affine limit function. Then we prove that the resulting sequence can be modified, without increasing the energy too much, into another one uniformly bounded in L^∞. Finally, we apply Theorem 3.1 to replace this last sequence of SBV functions by a sequence of Sobolev functions.

Step 1. Up to a subsequence, there is no loss of generality to assume the existence of nonnegative and finite Radon measures λ and $\mu \in \mathcal{M}(\Omega)$ such that $f(\cdot,u_n,\nabla u_n)\,\mathcal{L}^N \xrightarrow{*} \lambda$ and $\mathcal{H}^{N-1} \llcorner S_{u_n} \xrightarrow{*} \mu$ in $\mathcal{M}(\Omega)$. To prove Theorem 3.2 it is enough to check that

$$\lambda(\Omega) \ge \int_\Omega f(x,u,\nabla u)\,dx,$$

and thanks to Lebesgue's differentiation theorem, it suffices to show that

$$\frac{d\lambda}{d\mathcal{L}^N}(x_0) \geq f(x_0, u(x_0), \nabla u(x_0))$$

for $\mathcal{L}^N$-a.e. $x_0 \in \Omega$. Select $x_0 \in \Omega$ such that

(a) x_0 is a Lebesgue point of u and a and a point of approximate differentiability of u;

(b) the Radon–Nikodým derivative of λ with respect to $\mathcal{L}^N$ exists and is finite;

(c) the following limit exists and

$$\lim_{\rho \to 0^+} \frac{\mu(B(x_0, \rho))}{\omega_{N-1}\rho^{N-1}} = 0; \tag{3.2}$$

(d) for any sequence $\{\rho_i\} \searrow 0^+$ there exist a subsequence $\{\rho_{i(k)}\}$ and a $\mathcal{L}^N$-negligible set $E \subset B$ such that

$$\lim_{k \to +\infty} f(x_0 + \rho_{i(k)} y, u(x_0) + \rho_{i(k)} s, \xi) = f(x_0, u(x_0), \xi) \tag{3.3}$$

locally uniformly in $\mathbb{R}^d \times \mathbb{R}^{d \times N}$ for any $y \in B \setminus E$.

Note that $\mathcal{L}^N$-a.e. points x_0 in Ω satisfy these properties. Items (a) and (b) are immediate, while item (d) is a consequence of [3, Lemma 5.38]. Concerning item (c), we remark that, setting

$$\Theta(x) := \limsup_{\rho \to 0^+} \frac{\mu(B(x, \rho))}{\omega_{N-1}\rho^{N-1}},$$

then $\{\Theta > 0\} = \bigcup_{h=1}^{+\infty} \{\Theta \geq 1/h\}$, and using [3, Theorem 2.56], we get that $\mathcal{H}^{N-1}(\{\Theta \geq 1/h\}) \leq h\mu(\{\Theta \geq 1/h\}) < +\infty$. Thus $\mathcal{L}^N(\{\Theta \geq 1/h\}) = 0$ and consequently $\mathcal{L}^N(\{\Theta > 0\}) = 0$.

Consider a sequence $\{\rho_k\} \searrow 0^+$ such that $0 < \rho_k < 1$, $\mu(\partial B(x_0, \rho_k)) = 0$, and $\lambda(\partial B(x_0, \rho_k)) = 0$ for every $k \in \mathbb{N}$, and (3.3) holds with ρ_k in place of $\rho_{i(k)}$. Then

$$\begin{aligned}
\frac{d\lambda}{d\mathcal{L}^N}(x_0) &= \lim_{k \to +\infty} \frac{\lambda(B(x_0, \rho_k))}{\omega_N \rho_k^N} \\
&= \lim_{k \to +\infty} \lim_{n \to +\infty} \frac{1}{\omega_N \rho_k^N} \int_{B(x_0, \rho_k)} f(x, u_n, \nabla u_n)\, dx \\
&= \lim_{k \to +\infty} \lim_{n \to +\infty} \frac{1}{\omega_N} \int_B f(x_0 + \rho_k y, u(x_0) + \rho_k u_{n,k}, \nabla u_{n,k})\, dy, \qquad (3.4)
\end{aligned}$$

where we set $u_{n,k}(y) = [u_n(x_0 + \rho_k\, y) - u(x_0)]/\rho_k$. Since x_0 is a point of approximate differentiability of u, it follows that

$$\lim_{k \to +\infty} \lim_{n \to +\infty} \|u_{n,k} - w_0\|_{L^1(B;\mathbb{R}^d)} = 0, \tag{3.5}$$

where $w_0(y) := \nabla u(x_0)\, y$. Moreover, by (3.2) we get that

$$\begin{aligned}
\lim_{k \to +\infty} \lim_{n \to +\infty} \mathcal{H}^{N-1}(S_{u_{n,k}} \cap B) &= \lim_{k \to +\infty} \lim_{n \to +\infty} \frac{\mathcal{H}^{N-1}(S_{u_n} \cap B(x_0, \rho_k))}{\rho_k^{N-1}} \\
&= \lim_{k \to +\infty} \frac{\mu(B(x_0, \rho_k))}{\rho_k^{N-1}} = 0. \qquad (3.6)
\end{aligned}$$

From (3.4), (3.5), and (3.6), one can find a sequence $n(k) \nearrow +\infty$ such that, setting $v_k := u_{n(k),k}$, then $v_k \to w_0$ in $L^1(B;\mathbb{R}^d)$, $\mathcal{H}^{N-1}(S_{v_k} \cap B) \to 0$, and

$$\frac{d\lambda}{d\mathcal{L}^N}(x_0) = \lim_{k\to+\infty} \frac{1}{\omega_N} \int_B f(x_0 + \rho_k y, u(x_0) + \rho_k v_k, \nabla v_k)\, dy. \tag{3.7}$$

From now on, all the integrals will be restricted to the unit ball B.

Step 2. We now use the same truncation argument as in the proof of [3, Proposition 5.37]. Define $\hat{v}_k := (\sqrt{1+|v_k - w_0|^2} - 2)^+$ so that by Theorem 3.96 and Proposition 3.64(c) in [3], $\hat{v}_k \in SBV(B)$, $|\nabla \hat{v}_k| \le |\nabla v_k - \nabla w_0|$ $\mathcal{L}^N$-a.e. in B, and $S_{\hat{v}_k} \subset S_{v_k}$. According to the coarea formula in BV [3, Theorem 3.40], we have that

$$\begin{aligned}\int_0^1 \mathcal{H}^{N-1}\big(\partial^*\{\hat{v}_k > t\} \cap (B \setminus S_{\hat{v}_k})\big)\, dt \le |D\hat{v}_k|(B \setminus S_{\hat{v}_k}) = \int_B |\nabla \hat{v}_k|\, dx \\ \le \int_{B\cap\{|v_k - w_0| > \sqrt{3}\}} |\nabla v_k - \nabla w_0|\, dx,\end{aligned}$$

where we have used the fact that $\nabla \hat{v}_k = 0$ $\mathcal{L}^N$-a.e. in $B \cap \{|v_k - w_0| \le \sqrt{3}\}$. From (3.7) and the p-coercivity condition (3.1), the sequence $\{\nabla v_k\}$ is uniformly bounded in $L^p(B;\mathbb{R}^{d\times N})$, and since $p > 1$, it is equi-integrable. Using the fact that $\mathcal{L}^N(B \cap \{|v_k - w_0| > \sqrt{3}\}) \to 0$ we obtain that the right-hand side of the previous relation tends to zero as $k \to +\infty$. Consequently, one can find $t_k \in (0,1)$ such that $A_k := \{\hat{v}_k > t_k\}$ has finite perimeter in B and

$$\lim_{k\to+\infty} \mathcal{H}^{N-1}(B \cap \partial^* A_k \setminus S_{\hat{v}_k}) = 0. \tag{3.8}$$

Define $\tilde{v}_k := v_k \chi_{B\setminus A_k} + w_0 \chi_{B\cap A_k}$ so that $\tilde{v}_k \to w_0$ in $L^1(B;\mathbb{R}^d)$. As $|\hat{v}_k| \le t_k < 1$ in $B \setminus A_k$ it follows that $|v_k - w_0| \le 2\sqrt{2}$ in $B \setminus A_k$ and thus

$$\|\tilde{v}_k\|_{L^\infty(B;\mathbb{R}^d)} \le \|v_k\|_{L^\infty(B\setminus A_k;\mathbb{R}^d)} + \|w_0\|_{L^\infty(B;\mathbb{R}^d)} \le M, \tag{3.9}$$

where $M > 0$ is independent of k. Denoting by v_k^- the exterior trace of v_k on $\partial^* A_k \cap B$ oriented by the inner normal of A_k, [3, Remark 3.85] implies that $|v_k^-(x)| \le M$ for $\mathcal{H}^{N-1}$-a.e. $x \in \partial^* A_k \cap B$ and thus

$$\int_{\partial^* A_k \cap B} |v_k^-|\, d\mathcal{H}^{N-1} \le M \mathcal{H}^{N-1}(\partial^* A_k \cap B) < +\infty$$

so that [3, Theorem 3.84] ensures that $\tilde{v}_k \in SBV(B;\mathbb{R}^d)$. Since $S_{\tilde{v}_k} \subset S_{v_k} \cup \partial_* A_k$, by (3.8) we get that

$$\begin{aligned}\mathcal{H}^{N-1}(B \cap S_{\tilde{v}_k}) &\le \mathcal{H}^{N-1}(B \cap S_{v_k}) + \mathcal{H}^{N-1}(B \cap \partial_* A_k \setminus S_{v_k}) \\ &\le \mathcal{H}^{N-1}(B \cap S_{v_k}) + \mathcal{H}^{N-1}(B \cap \partial^* A_k \setminus S_{\hat{v}_k}) \to 0,\end{aligned}$$

where we used the fact that $S_{\hat{v}_k} \subset S_{v_k}$ and $\mathcal{H}^{N-1}(B \cap \partial_* A_k \setminus \partial^* A_k) = 0$. Using the locality of approximate gradients and the p-growth condition (3.1), we get that

$$\begin{aligned}\int_B & f(x_0+\rho_k y, u(x_0)+\rho_k \tilde v_k, \nabla \tilde v_k)\, dy\\ &= \int_{B\setminus A_k} f(x_0+\rho_k y, u(x_0)+\rho_k v_k, \nabla v_k)\, dy\\ &\qquad + \int_{B\cap A_k} f(x_0+\rho_k y, u(x_0)+\rho_k w_0, \nabla u(x_0))\, dy\\ &\le \int_B f(x_0+\rho_k y, u(x_0)+\rho_k v_k, \nabla v_k)\, dy\\ &\qquad + \int_{B\cap A_k} [a(x_0+\rho_k y)+\psi(|u(x_0)+\rho_k w_0|)(1+|\nabla u(x_0)|^p)]\, dy.\end{aligned}$$

By the choice of x_0, the sequence $\{a(x_0+\rho_k\cdot)\}$ is strongly converging in $L^1(B)$ to $a(x_0)$, and thus it is equi-integrable. Hence as $\mathcal{L}^N(A_k)\le \mathcal{L}^N(\{|v_k-w_0|\ge\sqrt{3}\})\to 0$ we deduce that the second term on the right-hand side of the previous relation tends to zero as $k\to+\infty$, and thanks to (3.7) it follows that

$$\frac{d\lambda}{d\mathcal{L}^N}(x_0)\ge \limsup_{k\to+\infty}\frac{1}{\omega_N}\int_B f(x_0+\rho_k y, u(x_0)+\rho_k\tilde v_k, \nabla\tilde v_k)\, dy. \tag{3.10}$$

Step 3. By (3.9) we have that $|D^s\tilde v_k|(B)\le 2M\mathcal{H}^{N-1}(S_{\tilde v_k}\cap B)\to 0$, while the p-coercivity condition (3.1) and item (b) imply that

$$\sup_{k\in\mathbb{N}}\|\nabla\tilde v_k\|_{L^p(B;\mathbb{R}^{d\times N})}<+\infty.$$

Consequently the sequence $\{\tilde v_k\}$ fulfills the assumptions of Theorem 3.1 so that considering a suitable (not relabeled) subsequence, there exist a Lebesgue measurable set $E_k\subset B$ and a sequence $\{w_k\}\subset W^{1,\infty}(B;\mathbb{R}^d)$ such that $\{|\nabla w_k|^p\}$ is equi-integrable, $w_k=\tilde v_k$ on $B\setminus E_k$, and $\mathcal{L}^N(E_k)\to 0$. From the proof of [21, Lemma 2.1], it can also be checked that $\sup_k\|w_k\|_{L^\infty(B;\mathbb{R}^d)}\le M$. As

$$\int_B |w_k-w_0|\, dy\le \int_{B\setminus E_k}|\tilde v_k-w_0|\, dy+2M\mathcal{L}^N(E_k)\to 0$$

it follows that $w_k\to w_0$ in $L^1(B;\mathbb{R}^d)$, and defining the set $B_k^t:=\{x\in B: |\nabla w_k(x)|\le t\}$, relation (3.10) leads to

$$\frac{d\lambda}{d\mathcal{L}^N}(x_0)\ge \limsup_{t\to+\infty}\limsup_{k\to+\infty}\frac{1}{\omega_N}\int_{B_k^t\setminus E_k} f(x_0+\rho_k y, u(x_0)+\rho_k w_k, \nabla w_k)\, dy.$$

Now using (3.3) with $\rho_{i(k)}=\rho_k$, we obtain that

$$\lim_{k\to+\infty}\int_{B_k^t\setminus E_k}|f(x_0+\rho_k y, u(x_0)+\rho_k w_k, \nabla w_k)-f(x_0,u(x_0),\nabla w_k)|\, dy=0$$

for each $t>0$, implying that

$$\frac{d\lambda}{d\mathcal{L}^N}(x_0)\ge \limsup_{t\to+\infty}\limsup_{k\to+\infty}\frac{1}{\omega_N}\int_{B_k^t\setminus E_k} f(x_0,u(x_0),\nabla w_k)\, dy. \tag{3.11}$$

Since $\mathcal{L}^N(E_k) \to 0$, according to the p-growth condition (3.1) we get that for every $t > 0$,

$$\int_{E_k \cap B_k^t} f(x_0, u(x_0), \nabla w_k)\, dy \leq \big(a(x_0) + \psi(|u(x_0)|)(1+t^p)\big)\mathcal{L}^N(E_k) \xrightarrow[k\to+\infty]{} 0. \tag{3.12}$$

On the other hand, Chebyshev's inequality ensures the existence of a constant $c > 0$ (independent of k and t) such that $\mathcal{L}^N(B \setminus B_k^t) \leq c/t^p \to 0$ as $t \to +\infty$, so that the equi-integrability of $\{|\nabla w_k|^p\}$ yields

$$\begin{aligned} &\sup_{k\in\mathbb{N}} \int_{B\setminus B_k^t} f(x_0, u(x_0), \nabla w_k)\, dy \\ &\leq \sup_{k\in\mathbb{N}} \int_{B\setminus B_k^t} \big(a(x_0) + \psi(|u(x_0)|)(1 + |\nabla w_k|^p)\big)\, dy \xrightarrow[t\to+\infty]{} 0. \end{aligned} \tag{3.13}$$

Gathering (3.11), (3.12), and (3.13), we deduce that

$$\frac{d\lambda}{d\mathcal{L}^N}(x_0) \geq \limsup_{k\to+\infty} \frac{1}{\omega_N} \int_B f(x_0, u(x_0), \nabla w_k)\, dy,$$

and since $w_k \rightharpoonup w_0$ in $W^{1,p}(B;\mathbb{R}^d)$, we can apply [1, Theorem II-4] to conclude that

$$\frac{d\lambda}{d\mathcal{L}^N}(x_0) \geq f(x_0, u(x_0), \nabla u(x_0)),$$

which yields the desired result. ∎

4. Structure of approximate scaled gradients. In this section we prove the following theorem, Theorem 4.1, which is a result similar to Theorem 3.1 in the context of dimension reduction. Note that it generalizes [8, Theorem 1.1] and [13, Theorem 3.1] (with obvious changes for nD-$(n-k)$D dimensional reduction). Its proof relies on a slicing argument similar to that used in [13, Theorem 3.1]. It will be instrumental in section 6 to prove Theorem 6.1 because it will enable us to replace SBV minimizing sequences by Lipschitz ones without increasing the energy.

From now on, $\Omega := \omega \times I$, where ω is a bounded open subset of $\mathbb{R}^2$ and $I := (-1/2, 1/2)$.

THEOREM 4.1. *Assume that ω has a Lipschitz boundary and $p > 1$. Let $\{\varepsilon_n\} \searrow 0^+$ and $\{u_n\} \subset SBV^p(\Omega;\mathbb{R}^3)$ be such that*

$$\sup_{n\in\mathbb{N}} \left\{ \|u_n\|_{L^\infty(\Omega;\mathbb{R}^3)} + \int_\Omega \left| \left(\nabla_\alpha u_n \Big| \frac{1}{\varepsilon_n} \nabla_3 u_n \right) \right|^p dx \right\} < +\infty, \tag{4.1}$$

$$\int_{S_{u_n}} \left| \left(\big(\nu_{u_n}\big)_\alpha \Big| \frac{1}{\varepsilon_n} \big(\nu_{u_n}\big)_3 \right) \right| d\mathcal{H}^2 \to 0 \tag{4.2}$$

and that $u_n \rightharpoonup u$ in $SBV^p(\Omega;\mathbb{R}^3)$, $(1/\varepsilon_n)\nabla_3 u_n \rightharpoonup b$ in $L^p(\Omega;\mathbb{R}^3)$ for some $u \in W^{1,p}(\omega;\mathbb{R}^3)$, and $b \in L^p(\Omega;\mathbb{R}^3)$. Then there exist a subsequence $\{\varepsilon_{n_k}\} \subset \{\varepsilon_n\}$ and a sequence $\{z_k\} \subset W^{1,\infty}(\Omega;\mathbb{R}^3)$ such that $z_k \rightharpoonup u$ in $W^{1,p}(\Omega;\mathbb{R}^3)$, $(1/\varepsilon_{n_k})\nabla_3 z_k \rightharpoonup b$ in $L^p(\Omega;\mathbb{R}^3)$, the sequence $\big\{\big|\big(\nabla_\alpha z_k | \frac{1}{\varepsilon_{n_k}} \nabla_3 z_k\big)\big|^p\big\}$ is equi-integrable, and

$$\mathcal{L}^3(\{z_k \neq u_{n_k}\} \cup \{\nabla z_k \neq \nabla u_{n_k}\}) \to 0.$$

Proof. The proof is based on a slicing argument. We first come back to the nonrescaled cylinder $\Omega_{\varepsilon_n} = \omega \times (-\varepsilon_n/2, \varepsilon_n/2)$ of thickness ε_n setting $v_n(x_\alpha, x_3) := u_n(x_\alpha, x_3/\varepsilon_n)$. It follows that for each $n \in \mathbb{N}$, $v_n \in SBV^p(\Omega_{\varepsilon_n}; \mathbb{R}^3)$, and changing variables in (4.1) we get that

$$\sup_{n\in\mathbb{N}} \left\{ \|v_n\|_{L^\infty(\Omega_{\varepsilon_n};\mathbb{R}^3)} + \frac{1}{\varepsilon_n} \int_{\Omega_{\varepsilon_n}} |\nabla v_n|^p \, dx \right\} < +\infty \tag{4.3}$$

and

$$\mathcal{H}^2(S_{v_n}) = \varepsilon_n \int_{S_{u_n}} \left| \left((\nu_{u_n})_\alpha \Big| \frac{1}{\varepsilon_n} (\nu_{u_n})_3 \right) \right| d\mathcal{H}^2. \tag{4.4}$$

We now periodize the functions v_n in the transverse direction defining

$$\hat{v}_n(x_\alpha, x_3) := \begin{cases} v_n(x_\alpha, -\varepsilon_n - x_3) & \text{if } \ -\varepsilon_n < x_3 \le -\frac{\varepsilon_n}{2}, \\ v_n(x_\alpha, x_3) & \text{if } \ -\frac{\varepsilon_n}{2} < x_3 < \frac{\varepsilon_n}{2}, \\ v_n(x_\alpha, \varepsilon_n - x_3) & \text{if } \ \frac{\varepsilon_n}{2} \le x_3 < \varepsilon_n. \end{cases}$$

Then $\hat{v}_n \in SBV^p\big(\omega \times (-\varepsilon_n, \varepsilon_n); \mathbb{R}^3\big)$ for each $n \in \mathbb{N}$, and from (4.3) and (4.4) it follows that

$$\sup_{n\in\mathbb{N}} \left\{ \|\hat{v}_n\|_{L^\infty(\omega\times(-\varepsilon_n,\varepsilon_n);\mathbb{R}^3)} + \frac{1}{\varepsilon_n} \int_{\omega\times(-\varepsilon_n,\varepsilon_n)} |\nabla \hat{v}_n|^p \, dx \right\} < +\infty \tag{4.5}$$

and

$$\mathcal{H}^2(S_{\hat{v}_n}) = 2\varepsilon_n \int_{S_{u_n}} \left| \left((\nu_{u_n})_\alpha \Big| \frac{1}{\varepsilon_n} (\nu_{u_n})_3 \right) \right| d\mathcal{H}^2. \tag{4.6}$$

We are now in a position to extend $\hat{v}_n$ by periodicity in the x_3 direction. Note that we do not create any additional jump set because periodicity ensures continuity at the interface of each slice. Let

$$N_n := \begin{cases} \dfrac{1}{4\varepsilon_n} - \dfrac{1}{2} & \text{if } \ \dfrac{1}{4\varepsilon_n} - \dfrac{1}{2} \in \mathbb{N}, \\ \left[\dfrac{1}{4\varepsilon_n} + \dfrac{1}{2} \right] & \text{otherwise}, \end{cases}$$

where $[t]$ denotes the integer part of t. For every $i \in \{-N_n, \dots, N_n\}$, we set $I_{i,n} := \big((2i-1)\varepsilon_n, (2i+1)\varepsilon_n\big)$ and $\Omega_{i,n} := \omega \times I_{i,n}$. Note that N_n is the largest integer such that $\Omega \cap \Omega_{i,n} \neq \emptyset$ for every $i \in \{-N_n, \dots, N_n\}$. We define the function $\tilde{v}_n$ on $\Omega(n) := \omega \times (-(2N_n+1)\varepsilon_n, (2N_n+1)\varepsilon_n)$ by extending $\hat{v}_n$ by periodicity in the x_3 direction on $\Omega(n)$:

$$\tilde{v}_n(x_\alpha, x_3) = \hat{v}_n(x_\alpha, x_3 - 2i\varepsilon_n) \quad \text{if } x_3 \in I_{i,n}.$$

Since $\Omega \subset \Omega(n)$, $\tilde{v}_n \in SBV^p(\Omega; \mathbb{R}^3)$ and thanks to (4.5) and the definition of N_n, we have that

$$\sup_{n\in\mathbb{N}} \left\{ \|\tilde{v}_n\|_{L^\infty(\Omega;\mathbb{R}^3)} + \int_\Omega |\nabla \tilde{v}_n|^p \, dx \right\} < +\infty, \tag{4.7}$$

while (4.6), together with (4.2), implies that

$$\mathcal{H}^2(S_{\tilde{v}_n}) \le c \int_{S_{u_n}} \left| \left((\nu_{u_n})_\alpha \Big| \frac{1}{\varepsilon_n} (\nu_{u_n})_3 \right) \right| d\mathcal{H}^2 \to 0. \tag{4.8}$$

As a consequence of (4.7) and (4.8), the sequence $\{\tilde{v}_n\}$ fulfills the assumptions of Theorem 3.1. Hence there exist a subsequence $\{\varepsilon_{n_k}\} \subset \{\varepsilon_n\}$ and a sequence $\{w_k\} \subset W^{1,\infty}(\Omega;\mathbb{R}^3)$ such that

$$\begin{cases} \sup_{k\in\mathbb{N}} \|w_k\|_{W^{1,p}(\Omega;\mathbb{R}^3)} < +\infty, \\ \{|\nabla w_k|^p\} \text{ is equi-integrable}, \\ \mathcal{L}^3(\{\tilde{v}_{n_k} \neq w_k\} \cup \{\nabla \tilde{v}_{n_k} \neq \nabla w_k\}) \to 0. \end{cases}$$

From de La Vallée Poussin's criterion, one can find an increasing and continuous function $\vartheta : [0,+\infty) \to [0,+\infty]$ such that $\vartheta(t)/t \to +\infty$ as $t \to +\infty$ and

$$\sup_{k\in\mathbb{N}} \int_\Omega \vartheta(|\nabla w_k|^p)\, dx < +\infty.$$

We claim that for at least half of the indexes $i \in \{-N_{n_k}+1, \dots, N_{n_k}-1\}$, there holds that

$$\begin{aligned} \frac{2N_{n_k}-1}{2} \int_{\Omega_{i,n_k}} &\left[\vartheta(|\nabla w_k|^p) + |w_k|^p + |\nabla w_k|^p\right] dx \\ &\le \int_\Omega \left[\vartheta(|\nabla w_k|^p) + |w_k|^p + |\nabla w_k|^p\right] dx. \end{aligned} \tag{4.9}$$

If not, define J_k to be the set of indexes $i \in \{-N_{n_k}+1, \dots, N_{n_k}-1\}$ such that (4.9) does not hold. Then it would imply that $\#(J_k) > (2N_{n_k}-1)/2$ and

$$\begin{aligned} \int_\Omega &\left[\vartheta(|\nabla w_k|^p) + |w_k|^p + |\nabla w_k|^p\right] dx \\ &\ge \sum_{i\in J_k} \int_{\Omega_{i,n_k}} \left[\vartheta(|\nabla w_k|^p) + |w_k|^p + |\nabla w_k|^p\right] dx \\ &> \frac{2}{2N_{n_k}-1} \#(J_k) \int_\Omega \left[\vartheta(|\nabla w_k|^p) + |w_k|^p + |\nabla w_k|^p\right] dx, \end{aligned}$$

which is absurd. Similarly, one can show that for at least half of the indexes satisfying (4.9), we have that

$$\begin{aligned} \frac{2N_{n_k}-1}{4} &\mathcal{L}^3 \llcorner \Omega_{i,n_k}(\{\tilde{v}_{n_k} \neq w_k\} \cup \{\nabla \tilde{v}_{n_k} \neq \nabla w_k\}) \\ &\le \mathcal{L}^3(\{\tilde{v}_{n_k} \neq w_k\} \cup \{\nabla \tilde{v}_{n_k} \neq \nabla w_k\}). \end{aligned} \tag{4.10}$$

Let $i_k \in \{-N_{n_k}+1, \dots, N_{n_k}-1\}$ be such that (4.9) and (4.10) hold at the same time. Now define $z_k(x_\alpha, x_3) := w_k(x_\alpha, \varepsilon_{n_k} x_3 + 2\varepsilon_{n_k} i_k)$. Changing variables in (4.9) and (4.10) and using the construction of $\tilde{v}_{n_k}$ from u_{n_k} we get that

$$\begin{aligned} \varepsilon_{n_k} \frac{2N_{n_k}-1}{2} \int_\Omega &\left[\vartheta\left(\left|\left(\nabla_\alpha z_k \Big| \frac{1}{\varepsilon_{n_k}} \nabla_3 z_k\right)\right|^p\right) + |z_k|^p + \left|\left(\nabla_\alpha z_k \Big| \frac{1}{\varepsilon_{n_k}} \nabla_3 z_k\right)\right|^p\right] dx \\ &\le \int_\Omega \left[\vartheta(|\nabla w_k|^p) + |w_k|^p + |\nabla w_k|^p\right] dx \end{aligned}$$

and

$$\varepsilon_{n_k} \frac{2N_{n_k}-1}{4} \mathcal{L}^3 \llcorner \Omega \left(\{u_{n_k} \neq z_k\} \cup \{\nabla u_{n_k} \neq \nabla z_k\}\right)$$
$$\leq \mathcal{L}^3(\{\tilde{v}_{n_k} \neq w_k\} \cup \{\nabla \tilde{v}_{n_k} \neq \nabla w_k\}).$$

Since $\varepsilon_{n_k}(2N_{n_k}-1) \geq 1/4$ for k large enough, it follows that

$$\begin{cases} \displaystyle\sup_{k\in\mathbb{N}} \int_\Omega \left[\vartheta\left(\left|\left(\nabla_\alpha z_k \Big| \frac{1}{\varepsilon_{n_k}}\nabla_3 z_k\right)\right|^p\right) + |z_k|^p + \left|\left(\nabla_\alpha z_k \Big| \frac{1}{\varepsilon_{n_k}}\nabla_3 z_k\right)\right|^p\right] dx < +\infty, \\ \mathcal{L}^3(\{u_{n_k} \neq z_k\} \cup \{\nabla u_{n_k} \neq \nabla z_k\}) \to 0, \end{cases}$$

and the equi-integrability of $\left\{\left|\left(\nabla_\alpha z_k \big| \frac{1}{\varepsilon_{n_k}}\nabla_3 z_k\right)\right|^p\right\}$ follows from de La Vallée Poussin's criterion.

It remains to prove the weak convergence of z_k and $(1/\varepsilon_{n_k})\nabla_3 z_k$. Let $v \in L^{p'}(\Omega;\mathbb{R}^3)$ with $1/p + 1/p' = 1$; then

$$\int_\Omega (z_k - u)\cdot v\,dx = \int_{\{z_k = u_{n_k}\}} (u_{n_k} - u)\cdot v\,dx + \int_{\{z_k \neq u_{n_k}\}} (z_k - u)\cdot v\,dx.$$

As $\mathcal{L}^3(\{z_k \neq u_{n_k}\}) \to 0$, it follows that $v\chi_{\{z_k = u_{n_k}\}} \to v$ in $L^{p'}(\Omega;\mathbb{R}^3)$. Then, using Hölder's inequality, the fact that $\{z_k\}$ is uniformly bounded in $L^p(\Omega;\mathbb{R}^3)$, and the fact that $u_{n_k} \rightharpoonup u$ in $L^p(\Omega;\mathbb{R}^3)$, we obtain that

$$\lim_{k\to+\infty} \left|\int_\Omega (z_k - u)\cdot v\,dx\right| \leq \lim_{k\to+\infty} \left|\int_\Omega (u_{n_k} - u)\cdot v\chi_{\{z_k = u_{n_k}\}}\,dx\right|$$
$$+ \lim_{k\to+\infty} \|z_k - u\|_{L^p(\Omega;\mathbb{R}^3)} \|v\chi_{\{z_k \neq u_{n_k}\}}\|_{L^{p'}(\Omega;\mathbb{R}^3)} = 0.$$

Similarly we may show that $\nabla z_k \rightharpoonup \nabla u$ in $L^p(\Omega;\mathbb{R}^{3\times 3})$ and that $(1/\varepsilon_{n_k})\nabla_3 z_k \rightharpoonup b$ in $L^p(\Omega;\mathbb{R}^3)$. ◻

5. Integral representation for dimension reduction problems in Sobolev spaces involving the bending moment.

Consider a Carathéodory function $W_\varepsilon : \Omega \times \mathbb{R}^{3\times 3} \to [0,+\infty)$ satisfying uniform p-growth and p-coercivity conditions: there exist $0 < \beta' \leq \beta < +\infty$ and $1 < p < +\infty$ such that

$$\beta'|\xi|^p \leq W_\varepsilon(x,\xi) \leq \beta(1 + |\xi|^p) \tag{5.1}$$

for a.e. $x \in \Omega$ and all $\xi \in \mathbb{R}^{3\times 3}$. Define $\mathcal{J}_\varepsilon : L^p(\Omega;\mathbb{R}^3) \times L^p(\omega;\mathbb{R}^3) \times \mathcal{A}(\omega) \to [0,+\infty]$ by

$$\mathcal{J}_\varepsilon(u,\overline{b},A) := \begin{cases} \displaystyle\int_{A\times I} W_\varepsilon\left(x, \nabla_\alpha u \Big| \frac{1}{\varepsilon}\nabla_3 u\right) dx & \text{if } \begin{cases} u \in W^{1,p}(A\times I;\mathbb{R}^3), \\ \overline{b} = \frac{1}{\varepsilon}\int_I \nabla_3 u(\cdot,x_3)\,dx_3, \end{cases} \\ +\infty & \text{otherwise.} \end{cases}$$

We prove the following integral representation for the Γ-limit.

THEOREM 5.1. *For every sequence $\{\varepsilon_n\} \searrow 0^+$, there exist a subsequence (not relabeled) and a Carathéodory function $W^* : \omega \times \mathbb{R}^{3\times 2} \times \mathbb{R}^3 \to [0,+\infty)$ (depending on* *the subsequence) such that for every $A \in \mathcal{A}(\omega)$, the sequence $\mathcal{J}_{\varepsilon_n}(\cdot,\cdot,A)$ Γ-converges in $L^p_s(A\times I;\mathbb{R}^3)\times L^p_w(A;\mathbb{R}^3)$ to $\mathcal{J}(\cdot,\cdot,A)$, where*

$$\mathcal{J}(u,\overline{b},A) = \begin{cases} \displaystyle\int_A W^*(x_\alpha, \nabla_\alpha u(x_\alpha)|\overline{b}(x_\alpha))\,dx_\alpha & \text{if } u \in W^{1,p}(A;\mathbb{R}^3),\\ +\infty & \text{otherwise.}\end{cases}$$

Proof. For every $\{\varepsilon_n\} \searrow 0^+$, $u \in L^p(\Omega;\mathbb{R}^3)$, $\overline{b} \in L^p(\omega;\mathbb{R}^3)$, and $A \in \mathcal{A}(\omega)$, let

$$\mathcal{J}(u,\overline{b},A) := \inf_{\{u_n,\overline{b}_n\}} \Big\{ \liminf_{n\to+\infty} \mathcal{J}_{\varepsilon_n}(u_n,\overline{b}_n,A) : u_n \to u \text{ in } L^p(A\times I;\mathbb{R}^3) \text{ and } \overline{b}_n \rightharpoonup \overline{b} \text{ in } L^p(A;\mathbb{R}^3)\Big\}.$$

Repeating word for word the (standard) proof of [9, Lemma 2.1] one can show that there exists a subsequence, still labeled $\{\varepsilon_n\}$, such that for any $A \in \mathcal{A}(\omega)$, $\mathcal{J}(\cdot,\cdot,A)$ is the Γ-limit in $L^p_s(A\times I;\mathbb{R}^3)\times L^p_w(A;\mathbb{R}^3)$ of $\mathcal{J}_{\varepsilon_n}(\cdot,\cdot,A)$, that $\mathcal{J}(u,\overline{b},A) = +\infty$ if $u \in L^p(\Omega;\mathbb{R}^3)\setminus W^{1,p}(A;\mathbb{R}^3)$, and that for every $(u,\overline{b}) \in W^{1,p}(\omega;\mathbb{R}^3)\times L^p(\omega;\mathbb{R}^3)$, the set function $\mathcal{J}(u,\overline{b},\cdot)$ is the restriction to $\mathcal{A}(\omega)$ of a Radon measure absolutely continuous with respect to the Lebesgue measure $\mathcal{L}^2$. The remainder of the proof is very close to that of [14, Theorem 1.1]; thus we will point out only the main changes. Let $\overline{\xi} \in \mathbb{R}^{3\times 2}$, $z \in \mathbb{R}^3$, and $x_0 \in \omega$ and define

$$W^*(x_0,\overline{\xi}|z) := \limsup_{\rho\to 0^+} \frac{\mathcal{J}(u_{\overline{\xi}},\overline{b}_z,Q'(x_0,\rho))}{\rho^2},$$

where we have denoted $u_{\overline{\xi}}(x_\alpha) := \overline{\xi}\,x_\alpha$ and $\overline{b}_z(x_\alpha) := z$. Since $\mathcal{J}(u_{\overline{\xi}},\overline{b}_z,\cdot)$ is (the restriction of) a Radon measure absolutely continuous with respect to $\mathcal{L}^2$, we have for every $A \in \mathcal{A}(\omega)$,

$$\mathcal{J}(u_{\overline{\xi}},\overline{b}_z,A) = \int_A W^*(x_\alpha,\overline{\xi}|z)\,dx_\alpha = \int_A W^*(x_\alpha,\nabla_\alpha u_{\overline{\xi}}|\overline{b}_z)\,dx_\alpha. \tag{5.2}$$

By additivity, it is clear that

$$\mathcal{J}(u,\overline{b},A) = \int_A W^*(x_\alpha,\nabla_\alpha u|\overline{b})\,dx_\alpha \tag{5.3}$$

holds whenever u is piecewise affine and $\overline{b}$ is piecewise constant in A and we wish to extend (5.3) to arbitrary functions $u \in W^{1,p}(A;\mathbb{R}^3)$ and $\overline{b} \in L^p(A;\mathbb{R}^3)$.

Using the lower semicontinuity of $\mathcal{J}$ and a suitable choice of sequence, one can show as in [14, Theorem 1.1] that $\overline{\xi} \mapsto W^*(x_0,\overline{\xi}|z)$ is rank one convex. We claim that $z \mapsto W^*(x_0,\overline{\xi}|z)$ is convex. To see this let $\theta \in [0,1]$, z_1, $z_2 \in \mathbb{R}^3$, and $\overline{\xi} \in \mathbb{R}^{3\times 2}$. Fix $x_0 \in \omega$, $\rho > 0$ and take an open set $A \subset Q'(x_0,\rho)$ such that $\mathcal{L}^2(\partial A) = 0$ and $\mathcal{L}^2(A) = \theta\rho^2$ (take, e.g., $A = Q'(x_0,\sqrt{\theta}\rho)$). Define

$$\overline{b}_n(x_\alpha) := z_1\chi(nx_\alpha) + z_2(1-\chi(nx_\alpha)),$$

where χ is the characteristic function of A in $Q'(x_0,\rho)$ which has been extended to $\mathbb{R}^2$ by ρ-periodicity. The Riemann–Lebesgue lemma asserts that $\overline{b}_n \rightharpoonup \overline{b}_{\theta z_1+(1-\theta)z_2}$ in

$L^p(Q'(x_0,\rho);\mathbb{R}^3)$, and since $\mathcal{J}(u_{\overline{\xi}},\cdot,Q'(x_0,\rho))$ is sequentially weakly lower semicontinuous in $L^p(Q'(x_0,\rho);\mathbb{R}^3)$, it follows that

$$\begin{aligned}
&\mathcal{J}(u_{\overline{\xi}},\overline{b}_{\theta z_1+(1-\theta)z_2},Q'(x_0,\rho))\\
&\qquad\le \liminf_{n\to+\infty}\mathcal{J}(u_{\overline{\xi}},\overline{b}_n,Q'(x_0,\rho))\\
&\qquad= \liminf_{n\to+\infty}\left\{\mathcal{J}(u_{\overline{\xi}},\overline{b}_{z_1},A_n)+\mathcal{J}(u_{\overline{\xi}},\overline{b}_{z_2},Q'(x_0,\rho)\setminus\overline{A}_n)\right\},
\end{aligned}\tag{5.4}$$

where $A_n:=\{x_\alpha\in Q'(x_0,\rho):\chi(nx_\alpha)=1\}$ is an open set. Note that in the last equality, we have used the fact that since $\mathcal{L}^2(\partial A_n)=0$, then $\mathcal{J}(u_{\overline{\xi}},\overline{b}_n,\partial A_n)=0$ as well, and that $\mathcal{J}$ is local on open sets. Using once more the Riemann–Lebesgue lemma together with (5.2), we get that

$$\begin{aligned}
\lim_{n\to+\infty}\mathcal{J}(u_{\overline{\xi}},\overline{b}_{z_1},A_n)&=\lim_{n\to+\infty}\int_{Q'(x_0,\rho)}\chi(nx_\alpha)W^*(x_\alpha,\overline{\xi}|z_1)\,dx_\alpha\\
&=\theta\int_{Q'(x_0,\rho)}W^*(x_\alpha,\overline{\xi}|z_1)\,dx_\alpha\\
&=\theta\mathcal{J}(u_{\overline{\xi}},\overline{b}_{z_1},Q'(x_0,\rho))
\end{aligned}$$

and similarly for the second term of (5.4). Hence we deduce that

$$\mathcal{J}(u_{\overline{\xi}},\overline{b}_{\theta z_1+(1-\theta)z_2},Q'(x_0,\rho))\le\theta\mathcal{J}(u_{\overline{\xi}},\overline{b}_{z_1},Q'(x_0,\rho))+(1-\theta)\mathcal{J}(u_{\overline{\xi}},\overline{b}_{z_2},Q'(x_0,\rho)),$$

and the convexity of $W^*(x_0,\overline{\xi}|\cdot)$ arises after dividing the previous inequality by ρ^2 and taking the lim sup as ρ tends to zero. It follows that $(\overline{\xi}|z)\mapsto W^*(x_0,\overline{\xi}|z)$ is separately convex for a.e. $x_0\in\omega$, and since the following p-growth and p-coercivity conditions hold,

$$\beta'(|\overline{\xi}|^p+|z|^p)\le W^*(x_0,\overline{\xi}|z)\le\beta(1+|\overline{\xi}|^p+|z|^p),\tag{5.5}$$

for a.e. $x_0\in\omega$ and all $(\overline{\xi},z)\in\mathbb{R}^{3\times2}\times\mathbb{R}^3$, we conclude that $(\overline{\xi}|z)\mapsto W^*(x_0,\overline{\xi}|z)$ is continuous for a.e. $x_0\in\omega$, which proves that W^* is a Carathéodory function.

We now prove that (5.3) holds for any $(u,\overline{b})\in W^{1,p}(A;\mathbb{R}^3)\times L^p(A;\mathbb{R}^3)$. By approximation and thanks to the lower semicontinuity of $\mathcal{J}(\cdot,\cdot,A)$ for the strong $W^{1,p}(A;\mathbb{R}^3)\times L^p(A;\mathbb{R}^3)$ topology, there holds that

$$\mathcal{J}(u,\overline{b},A)\le\int_A W^*(x_\alpha,\nabla_\alpha u|\overline{b})\,dx_\alpha$$

for any $(u,\overline{b})\in W^{1,p}(A;\mathbb{R}^3)\times L^p(A;\mathbb{R}^3)$, and it remains to prove the converse inequality. This is achieved exactly as in the final step of the proof of [14, Theorem 1.1], by considering the translated functional

$$\widetilde{\mathcal{J}}(v,\overline{c},A):=\mathcal{J}(u+v,\overline{b}+\overline{c},A),$$

where $(u,\overline{b})$ are arbitrary functions in $W^{1,p}(A;\mathbb{R}^3)\times L^p(A;\mathbb{R}^3)$. □

We refer the reader to [8, 9] for more explicit formulas for the integrand W^* in particular cases.

The following technical proposition states some kind of blow-up result for functionals through Γ-convergence. It will be of use in the proof of the lower bound in Theorem 6.1 because at some point, we will need to get rid of small residual terms occurring inside the integrand W_ε. In [5, 6, 7], this difficulty was treated thanks to a decoupling variable method which consisted of replacing the function W_ε by a much more regular one thanks to the Scorza–Dragoni and the Tietze extension theorems, and the set where these two integrands did not match was controlled thanks to the equi-integrability result [8, Theorem 1.1]. This method was quite powerful in that context since the manner on which W_ε was depending on ε was completely known. However, in the generalized framework considered here, it no longer applies since we have no information on the way W_ε depends on ε. The following blow-up result, together with a diagonalization argument (see Remark 5.3 below), will enable us to overcome that problem.

PROPOSITION 5.2. *There exists a set $N \subset \omega$ with $\mathcal{L}^2(N) = 0$ such that for every $\{\rho_k\} \searrow 0^+$ and every $x_0 \in \omega \setminus N$, the functional $J_k : L^p(B' \times I;\mathbb{R}^3) \times L^p(B';\mathbb{R}^3) \to [0,+\infty]$ defined by*

$$J_k(u,\bar{b}) = \begin{cases} \displaystyle\int_{B'} W^*(x_0 + \rho_k x_\alpha, \nabla_\alpha u(x_\alpha)|\bar{b}(x_\alpha))\, dx_\alpha & \text{if } u \in W^{1,p}(B';\mathbb{R}^3), \\ +\infty & \text{otherwise} \end{cases}$$

Γ-converges in $L^p_s(B' \times I;\mathbb{R}^3) \times L^p_w(B';\mathbb{R}^3)$ to $J : L^p(B' \times I;\mathbb{R}^3) \times L^p(B';\mathbb{R}^3) \to [0,+\infty]$, where

$$J(u,\bar{b}) = \begin{cases} \displaystyle\int_{B'} W^*(x_0, \nabla_\alpha u(x_\alpha)|\bar{b}(x_\alpha))\, dx_\alpha & \text{if } u \in W^{1,p}(B';\mathbb{R}^3), \\ +\infty & \text{otherwise.} \end{cases}$$

Proof. The proof relies on the Scorza–Dragoni theorem (see, e.g., [17, Chapter VIII]). For any $q \in \mathbb{N}$, there exists a compact set $K_q \subset \omega$ with $\mathcal{L}^2(\omega \setminus K_q) < 1/q$ and such that W^* is continuous on $K_q \times \mathbb{R}^{3\times 2} \times \mathbb{R}^3$. Let $N := \omega \setminus \bigcup_q K_q^*$, where

$$K_q^* := \left\{ x \in K_q : \lim_{\rho \to 0} \frac{\mathcal{L}^2(B'(x_0,\rho) \setminus K_q)}{\mathcal{L}^2(B'(x_0,\rho))} = 0 \right\}. \tag{5.6}$$

Since $\mathcal{L}^2(K_q \setminus K_q^*) = 0$, then $\mathcal{L}^2(N) \le \mathcal{L}^2(\omega \setminus K_q^*) = \mathcal{L}^2(\omega \setminus K_q) < 1/q \to 0$. Select a point $x_0 \in \omega \setminus N$, so that $x_0 \in K_q^*$ for some $q \in \mathbb{N}$.

The upper bound. Assume first that $u \in W^{1,\infty}(B';\mathbb{R}^3)$ and $\bar{b} \in L^\infty(B';\mathbb{R}^3)$ and set $M := \|(\nabla_\alpha u|\bar{b})\|_{L^\infty(B';\mathbb{R}^{3\times 3})}$. Then according to the p-growth condition (5.5)

$$\begin{aligned} J_k(u,\bar{b}) &= \int_{B'} W^*(x_0 + \rho_k x_\alpha, \nabla_\alpha u|\bar{b})\, dx_\alpha \\ &\le \int_{B' \cap \left(\frac{K_q - x_0}{\rho_k}\right)} W^*(x_0 + \rho_k x_\alpha, \nabla_\alpha u|\bar{b})\, dx_\alpha \\ &\quad + \beta(1 + M^p)\mathcal{L}^2\left(B' \setminus \left(\frac{K_q - x_0}{\rho_k}\right)\right). \end{aligned} \tag{5.7}$$

As W^* is uniformly continuous on $K_q \times B(0,M)$, there exists a continuous and increasing function $\eta : [0,+\infty) \to [0,+\infty)$ such that $\eta(0) = 0$ and

$$\int_{B' \cap \left(\frac{K_q - x_0}{\rho_k}\right)} |W^*(x_0 + \rho_k x_\alpha, \nabla_\alpha u|\bar{b}) - W^*(x_0, \nabla_\alpha u|\bar{b})|\, dx_\alpha \leq \eta(\rho_k). \tag{5.8}$$

Gathering (5.6), (5.7), and (5.8) and passing to the limit as $k \to +\infty$ yields

$$\Gamma\text{-}\limsup_{k\to+\infty} J_k(u,\bar{b}) \leq \limsup_{k\to+\infty} J_k(u,\bar{b}) \leq J(u,\bar{b}).$$

The general case follows from the density of the space $W^{1,\infty}(B';\mathbb{R}^3) \times L^\infty(B';\mathbb{R}^3)$ in $W^{1,p}(B';\mathbb{R}^3) \times L^p(B';\mathbb{R}^3)$, the lower continuity of the Γ-limsup, and the continuity of J for the strong $W^{1,p}(B';\mathbb{R}^3) \times L^p(B';\mathbb{R}^3)$-topology.

The lower bound. Let $(u,\bar{b}) \in L^p(B' \times I;\mathbb{R}^3) \times L^p(B';\mathbb{R}^3)$ and $\{(u_k,\bar{b}_k)\} \subset L^p(B' \times I;\mathbb{R}^3) \times L^p(B';\mathbb{R}^3)$ such that $u_k \to u$ in $L^p(B' \times I;\mathbb{R}^3)$, $\bar{b}_k \rightharpoonup \bar{b}$ in $L^p(B';\mathbb{R}^3)$, and

$$\liminf_{k\to+\infty} J_k(u_k,\bar{b}_k) < +\infty.$$

Up to a subsequence (not relabeled) we can suppose that u and $u_k \in W^{1,p}(B';\mathbb{R}^3)$ for each $k \in \mathbb{N}$ and that $u_k \rightharpoonup u$ in $W^{1,p}(B';\mathbb{R}^3)$. According to the decomposition lemma [20, Lemma 1.2] and Chacon's biting lemma [3, Lemma 5.32], there is no loss of generality to assume that $\{|\nabla_\alpha u_k|^p\}$ and $\{|\bar{b}_k|^p\}$ are equi-integrable. Define the set $A_k^t := \left\{x_\alpha \in B' : |(\nabla_\alpha u_k(x_\alpha)|\bar{b}_k(x_\alpha))| \leq t\right\}$. From Chebyshev's inequality we have that $\mathcal{L}^2(B' \setminus A_k^t) \leq c/t^p$ for some constant $c > 0$ independent of t and k, and arguing exactly as in the proof of the upper bound, one can show that for each $t > 0$,

$$\liminf_{k\to+\infty} J_k(u_k,\bar{b}_k) \geq \liminf_{k\to+\infty} \int_{A_k^t \cap \left(\frac{K_q - x_0}{\rho_k}\right)} W^*(x_0, \nabla_\alpha u_k|\bar{b}_k)\, dx_\alpha. \tag{5.9}$$

According to the p-growth condition (5.5) and (5.6),

$$\begin{aligned} &\int_{A_k^t \setminus \left(\frac{K_q - x_0}{\rho_k}\right)} W^*(x_0, \nabla_\alpha u_k|\bar{b}_k)\, dx_\alpha \\ &\qquad \leq \beta(1+t^p)\mathcal{L}^2\left(B' \setminus \left(\frac{K_q - x_0}{\rho_k}\right)\right) \xrightarrow[k\to+\infty]{} 0, \end{aligned} \tag{5.10}$$

while the equi-integrability of $\{|\nabla_\alpha u_k|^p\}$ and $\{|\bar{b}_k|^p\}$ and the fact that $\mathcal{L}^2(B' \setminus A_k^t) \to 0$ as $t \to +\infty$ imply that

$$\begin{aligned} &\sup_{k\in\mathbb{N}} \int_{B' \setminus A_k^t} W^*(x_0, \nabla_\alpha u_k|\bar{b}_k)\, dx_\alpha \\ &\qquad \leq \beta \sup_{k\in\mathbb{N}} \int_{B' \setminus A_k^t} (1 + |\nabla_\alpha u_k|^p + |\bar{b}_k)|^p)\, dx_\alpha \xrightarrow[t\to+\infty]{} 0. \end{aligned} \tag{5.11}$$

Hence gathering (5.9), (5.10), and (5.11) yields

$$\liminf_{k\to+\infty} J_k(u_k,\bar{b}_k) \geq \liminf_{k\to+\infty} J(u_k,\bar{b}_k) \geq J(u,\bar{b}),$$

where the last inequality holds because J is sequentially weakly lower semicontinuous in $W^{1,p}(B';\mathbb{R}^3) \times L^p(B';\mathbb{R}^3)$. ◻

Remark 5.3. One can show that in Theorem 5.1, the value of $\mathcal{J}$ does not change when replacing W_{ε_n} by its quasi-convexification QW_{ε_n} defined by

$$QW_{\varepsilon_n}(x,\xi) := \inf_{\varphi \in W_0^{1,\infty}((0,1)^3;\mathbb{R}^3)} \int_{(0,1)^3} W_{\varepsilon_n}(x, \xi + \nabla\varphi(y))\, dy \tag{5.12}$$

for all $\xi \in \mathbb{R}^{3\times 3}$ and a.e. $x \in \Omega$. Hence there is no loss of generality to assume in Theorem 5.1 that W_ε is quasi-convex. Since the weak topology on every normed bounded subset of $L^p(B';\mathbb{R}^3)$ is metrizable, it follows from a diagonalization argument, Theorem 5.1, Proposition 5.2, and the fact that Γ-convergence of coercive and lower semicontinuous functionals on a metric space is metrizable (see [16, Theorem 10.22(a)]) that for every $M > 0$ and every sequence $\{\rho_k\} \searrow 0^+$, there exists a subsequence $n(k) \nearrow +\infty$ such that $\varepsilon_{n(k)}/\rho_k \to 0$ and for every $(u,\bar{b}) \in L^p(B' \times I;\mathbb{R}^3) \times L^p(B';\mathbb{R}^3)$ with $\|\bar{b}\|_{L^p(B';\mathbb{R}^3)} \le M$, the Γ-limit in $L^p_s(B' \times I;\mathbb{R}^3) \times L^p_w(B';\mathbb{R}^3)$ of

$$\begin{cases} \displaystyle\int_{B'\times I} W_{\varepsilon_{n(k)}}\left(x_0 + \rho_k x_\alpha, x_3, \nabla_\alpha u \Big| \frac{\rho_k}{\varepsilon_{n(k)}} \nabla_3 u\right) dx & \text{if } \begin{cases} u \in W^{1,p}(B' \times I;\mathbb{R}^3), \\ \bar{b} = \frac{\rho_k}{\varepsilon_{n(k)}} \int_I \nabla_3 u(\cdot, x_3)\, dx_3, \end{cases} \\ +\infty & \text{otherwise} \end{cases}$$

coincides with

$$\begin{cases} \displaystyle\int_{B'} W^*(x_0, \nabla_\alpha u|\bar{b})\, dx_\alpha & \text{if } u \in W^{1,p}(B';\mathbb{R}^3), \\ +\infty & \text{otherwise} \end{cases}$$

for every $x_0 \in \omega \setminus N$, where $N \subset \omega$ is the same exceptional set as in Proposition 5.2.

6. Integral representation for dimension reduction problems in *SBV* involving the bending moment. We now come to the heart of this study, that is, dealing with a similar problem to that in Theorem 5.1 but in the framework of special functions with bounded variation, adding a surface energy term. Let us define $\mathcal{G}_\varepsilon : BV(\Omega;\mathbb{R}^3) \times L^p(\omega;\mathbb{R}^3) \to [0,+\infty]$ by

$$\mathcal{G}_\varepsilon(u,\bar{b}) := \begin{cases} \displaystyle\int_\Omega W_\varepsilon\left(x, \nabla_\alpha u \Big| \frac{1}{\varepsilon}\nabla_3 u\right) dx & \\ \displaystyle\qquad + \int_{S_u} \left| \left((\nu_u)_\alpha \,\Big|\, \frac{1}{\varepsilon}\,(\nu_u)_3 \right) \right| d\mathcal{H}^2 & \text{if } \begin{cases} u \in SBV^p(\Omega;\mathbb{R}^3), \\ \bar{b} = \frac{1}{\varepsilon} \int_I \nabla_3 u(\cdot, x_3)\, dx_3, \end{cases} \\ +\infty & \text{otherwise.} \end{cases}$$

Then the following Γ-convergence result holds.

THEOREM 6.1. *For every sequence $\{\varepsilon_n\} \searrow 0^+$, there exists a subsequence, still labeled $\{\varepsilon_n\}$, such that $\mathcal{G}_{\varepsilon_n}$ Γ-converges in $L^1_s(\Omega;\mathbb{R}^3) \times L^p_w(\omega;\mathbb{R}^3)$ to $\mathcal{G} : BV(\Omega;\mathbb{R}^3) \times L^p(\omega;\mathbb{R}^3) \to [0,+\infty]$ defined by*

$$\mathcal{G}(u,\bar{b}) := \begin{cases} \displaystyle\int_\omega W^*(x_\alpha, \nabla_\alpha u|\bar{b})\, dx_\alpha + \mathcal{H}^1(S_u) & \textit{if } u \in SBV^p(\omega;\mathbb{R}^3), \\ +\infty & \textit{otherwise,} \end{cases}$$

where W^* is given by Theorem 5.1.

The remainder of this section is devoted to proving Theorem 6.1. We will first localize the functional $\mathcal{G}_\varepsilon$ on $\mathcal{A}(\omega)$, and noticing that minimizing sequences are not necessarily weakly relatively compact in BV, we will use the same truncation argument as in [4] (see also [18]), introducing an artificial functional. Then we will show that it actually coincides with the Γ-limit whenever $u \in BV(\Omega;\mathbb{R}^3) \cap L^\infty(\Omega;\mathbb{R}^3)$ (see Lemma 6.2 and Remark 6.3), and it will enable us to show that for such u's the Γ-limit is a measure absolutely continuous with respect to $\mathcal{L}^2 + \mathcal{H}^1 \llcorner S_u$ (see Lemma 6.6). Together with a blow-up argument, this property will be useful to prove the upper bound in section 6.3, while the lower bound, in section 6.4, will be obtained thanks to Theorem 4.1 and a suitable diagonalization argument (see Remark 5.3).

6.1. Localization. We start by localizing our functional on $\mathcal{A}(\omega)$, defining $\mathcal{G}_\varepsilon : BV(\Omega;\mathbb{R}^3) \times L^p(\omega;\mathbb{R}^3) \times \mathcal{A}(\omega) \to [0,+\infty]$ by

$$\mathcal{G}_\varepsilon(u,\bar{b},A) := \begin{cases} \displaystyle\int_{A\times I} W_\varepsilon\left(x, \nabla_\alpha u \Big| \frac{1}{\varepsilon}\nabla_3 u\right) dx & \\ \displaystyle\quad + \int_{S_u \cap (A\times I)} \left| \left((\nu_u)_\alpha \,\Big|\, \frac{1}{\varepsilon}\,(\nu_u)_3 \right) \right| d\mathcal{H}^2 & \text{if } \begin{cases} u \in SBV^p(A\times I;\mathbb{R}^3), \\ \bar{b} = \frac{1}{\varepsilon}\int_I \nabla_3 u(\cdot,x_3)\,dx_3, \end{cases} \\ +\infty & \text{otherwise.} \end{cases}$$

For every sequence $\{\varepsilon_n\} \searrow 0^+$ and all $(u,\bar{b},A) \in BV(\Omega;\mathbb{R}^3) \times L^p(\omega;\mathbb{R}^3) \times \mathcal{A}(\omega)$, we define

$$\mathcal{E}(u,\bar{b},A) := \inf_{\{u_n,\bar{b}_n\}} \Big\{ \liminf_{n\to+\infty} \mathcal{G}_{\varepsilon_n}(u_n,\bar{b}_n,A) : u_n \to u \text{ in } L^1(A\times I;\mathbb{R}^3)$$

$$\text{and } \bar{b}_n \rightharpoonup \bar{b} \text{ in } L^p(A;\mathbb{R}^3) \Big\}. \tag{6.1}$$

Theorem 8.5 and Corollary 8.12 in [16] together with a diagonalization argument imply the existence of a subsequence, still denoted $\{\varepsilon_n\}$, such that, for any $A \in \mathcal{R}(\omega)$ (or $A = \omega$), $\mathcal{E}(\cdot,\cdot,A)$ is the Γ-limit of $\mathcal{G}_{\varepsilon_n}(\cdot,\cdot,A)$ in $L^1_s(A\times I;\mathbb{R}^3) \times L^p_w(A;\mathbb{R}^3)$. Extracting if necessary a further subsequence, one may assume that $\{\varepsilon_n\}$ is chosen so that Theorem 5.1 holds. To prove Theorem 6.1, it is enough to show that $\mathcal{E}(u,\bar{b},\omega) = \mathcal{G}(u,\bar{b})$.

6.2. A truncation argument. As pointed out in [4], the main problem with the definition of $\mathcal{E}$ in (6.1) is that minimizing sequences are not necessarily bounded in $BV(\Omega;\mathbb{R}^3)$ and thus not necessarily weakly convergent in this space. Following [4], we define for all $(u,\bar{b},A) \in BV(\Omega;\mathbb{R}^3) \times L^p(\omega;\mathbb{R}^3) \times \mathcal{A}(\omega)$

$$\mathcal{E}_\infty(u,\bar{b},A) := \inf_{\{u_n,\bar{b}_n\}} \Big\{ \liminf_{n\to+\infty} \mathcal{G}_{\varepsilon_n}(u_n,\bar{b}_n,A) : u_n \to u \text{ in } L^1(A\times I;\mathbb{R}^3),$$

$$\bar{b}_n \rightharpoonup \bar{b} \text{ in } L^p(A;\mathbb{R}^3) \text{ and } \sup_{n\in\mathbb{N}} \|u_n\|_{L^\infty(A\times I;\mathbb{R}^3)} < +\infty \Big\}.$$

It is immediate that $\mathcal{E}(u,\bar{b},A) \le \mathcal{E}_\infty(u,\bar{b},A)$, while we will show that equality holds when u belongs to $BV(\Omega;\mathbb{R}^3) \cap L^\infty(\Omega;\mathbb{R}^3)$. This will be obtained as a consequence of Lemma 6.2. It means that for such deformation fields $u \in BV(\Omega;\mathbb{R}^3) \cap L^\infty(\Omega;\mathbb{R}^3)$, strong $L^1(\Omega;\mathbb{R}^3)$-convergence and weak $BV(\Omega;\mathbb{R}^3)$-convergence are, in a sense, equivalent for the computation of the Γ-limit.

LEMMA 6.2. *Let $A \in \mathcal{A}(\omega)$, $u \in BV(\Omega;\mathbb{R}^3) \cap L^\infty(\Omega;\mathbb{R}^3)$, and $\overline{b} \in L^p(\omega;\mathbb{R}^3)$. If $\{u_n\} \subset SBV^p(A \times I;\mathbb{R}^3)$ is such that $u_n \to u$ in $L^1(A \times I;\mathbb{R}^3)$, $\frac{1}{\varepsilon_n}\int_I \nabla_3 u_n(\cdot,x_3)\, dx_3 \rightharpoonup \overline{b}$ in $L^p(A;\mathbb{R}^3)$ and the limit*

$$L := \lim_{n\to+\infty} \mathcal{G}_{\varepsilon_n}\left(u_n, \frac{1}{\varepsilon_n}\int_I \nabla_3 u_n(\cdot,x_3)\, dx_3, A\right)$$

exists and is finite. Then for any $\eta > 0$ one can find $C > 0$ and $\{w_n\} \subset SBV^p(A \times I;\mathbb{R}^3)$ such that $w_n \to u$ in $L^1(A \times I;\mathbb{R}^3)$, $\frac{1}{\varepsilon_n}\int_I \nabla_3 w_n(\cdot,x_3)\, dx_3 \rightharpoonup \overline{b}$ in $L^p(A;\mathbb{R}^3)$, $\sup_n \|w_n\|_{L^\infty(A\times I;\mathbb{R}^3)} \le C$, and

$$L \ge \limsup_{n\to+\infty} \mathcal{G}_{\varepsilon_n}\left(w_n, \frac{1}{\varepsilon_n}\int_I \nabla_3 w_n(\cdot,x_3)\, dx_3, A\right) - \eta.$$

Proof. Let us define a smooth truncation function $\varphi_i \in \mathcal{C}^1_c(\mathbb{R}^3;\mathbb{R}^3)$ satisfying

$$\varphi_i(s) = \begin{cases} s & \text{if} \quad |s| < e^i, \\ 0 & \text{if} \quad |s| \ge e^{i+1} \end{cases} \qquad \text{and} \qquad |\nabla\varphi_i(s)| \le 2. \tag{6.2}$$

Let $w_{n,i} := \varphi_i(u_n)$, thanks to the chain rule formula [3, Theorem 3.96], $w_{n,i} \in SBV^p(A \times I;\mathbb{R}^3)$, and

$$\begin{cases} \|w_{n,i}\|_{L^\infty(A\times I;\mathbb{R}^3)} \le e^{i+1}, \\ S_{w_{n,i}} \subset S_{u_n}, \\ \nabla w_{n,i} = \nabla\varphi_i(u_n)\nabla u_n \quad \mathcal{L}^3\text{-a.e. in } A \times I. \end{cases} \tag{6.3}$$

Since $u \in L^\infty(\Omega;\mathbb{R}^3)$, we can choose i large enough ($i \ge m := [\ln(\|u\|_{L^\infty(\Omega;\mathbb{R}^3)})] + 1$) so that $u = \varphi_i(u)$ and thus according to (6.2)

$$\|w_{n,i} - u\|_{L^1(A\times I;\mathbb{R}^3)} = \|\varphi_i(u_n) - \varphi_i(u)\|_{L^1(A\times I;\mathbb{R}^3)} \le 2\|u_n - u\|_{L^1(A\times I;\mathbb{R}^3)}. \tag{6.4}$$

Since (a subsequence of) $u_n \to u$ a.e. in $A \times I$ and $\nabla\varphi_i$ is continuous, it follows that $\nabla\varphi_i(u_n) \to \nabla\varphi_i(u) = \mathrm{Id}$ a.e. in $A \times I$ as $n \to +\infty$. Take $v \in L^{p'}(A;\mathbb{R}^3)$, where $1/p + 1/p' = 1$; as $|\nabla\varphi_i(u_n)^T v| \le 2|v| \in L^{p'}(A)$, the dominated convergence theorem implies that $\nabla\varphi_i(u_n)^T v \to v$ in $L^{p'}(A \times I;\mathbb{R}^3)$ and thus

$$\lim_{n\to+\infty} \int_A \left(\frac{1}{\varepsilon_n}\int_I \nabla_3 w_{n,i}(x_\alpha,x_3)\, dx_3\right) \cdot v(x_\alpha)\, dx_\alpha$$

$$= \lim_{n\to+\infty} \int_{A\times I} \frac{1}{\varepsilon_n}\nabla_3 u_n \cdot \left(\nabla\varphi_i(u_n)^T v\right)\, dx = \int_A \overline{b} \cdot v\, dx_\alpha,$$

where we used the fact that $(1/\varepsilon_n)\nabla_3 u_n \rightharpoonup b$ in $L^p(A \times I;\mathbb{R}^3)$ and $\overline{b} = \int_I b(\cdot,x_3)\, dx_3$. Hence

$$\frac{1}{\varepsilon_n}\int_I \nabla_3 w_{n,i}(\cdot,x_3)\, dx_3 \xrightharpoonup[n\to+\infty]{} \overline{b} \text{ in } L^p(A;\mathbb{R}^3) \quad \text{for all } i \ge m. \tag{6.5}$$

The growth condition (5.1), (6.2), and (6.3) imply that

$$\begin{aligned}
&\int_{A\times I} W_{\varepsilon_n}\left(x, \nabla_\alpha w_{n,i}\Big|\frac{1}{\varepsilon_n}\nabla_3 w_{n,i}\right)dx \\
&\qquad\le \int_{\{|u_n|<e^i\}} W_{\varepsilon_n}\left(x, \nabla_\alpha u_n\Big|\frac{1}{\varepsilon_n}\nabla_3 u_n\right)dx + \beta\mathcal{L}^3(\{|u_n|\ge e^{i+1}\}) \\
&\qquad\quad + \int_{\{e^i\le|u_n|<e^{i+1}\}} W_{\varepsilon_n}\left(x, \nabla\varphi_i(u_n)\nabla_\alpha u_n\Big|\frac{1}{\varepsilon_n}\nabla\varphi_i(u_n)\nabla_3 u_n\right)dx \\
&\qquad\le \int_{A\times I} W_{\varepsilon_n}\left(x, \nabla_\alpha u_n\Big|\frac{1}{\varepsilon_n}\nabla_3 u_n\right)dx + \beta\, e^{-i}\,\|u_n\|_{L^1(A\times I;\mathbb{R}^3)} \\
&\qquad\quad + 2^p\beta\int_{\{e^i\le|u_n|<e^{i+1}\}}\left|\left(\nabla_\alpha u_n\Big|\frac{1}{\varepsilon_n}\nabla_3 u_n\right)\right|^p dx,
\end{aligned}\tag{6.6}$$

where we have used Chebyshev's inequality. Since $\nu_{w_{n,i}}(x) = \pm\nu_{u_n}(x)$ for $\mathcal{H}^2$-a.e. $x\in S_{w_{n,i}}$, (6.3) yields

$$\begin{aligned}
&\int_{S_{w_{n,i}}\cap(A\times I)}\left|\left(\left(\nu_{w_{n,i}}\right)_\alpha\Big|\frac{1}{\varepsilon_n}\left(\nu_{w_{n,i}}\right)_3\right)\right|d\mathcal{H}^2 \\
&\qquad\qquad\le \int_{S_{u_n}\cap(A\times I)}\left|\left(\left(\nu_{u_n}\right)_\alpha\Big|\frac{1}{\varepsilon_n}\left(\nu_{u_n}\right)_3\right)\right|d\mathcal{H}^2.
\end{aligned}\tag{6.7}$$

Let $M\in\mathbb{N}$; from (6.6) and (6.7), a summation for $i=m$ to M implies that

$$\begin{aligned}
&\frac{1}{M-m+1}\sum_{i=m}^{M}\left[\int_{A\times I} W_{\varepsilon_n}\left(x, \nabla_\alpha w_{n,i}\Big|\frac{1}{\varepsilon_n}\nabla_3 w_{n,i}\right)dx\right. \\
&\qquad\qquad\left. + \int_{S_{w_{n,i}}\cap(A\times I)}\left|\left(\left(\nu_{w_{n,i}}\right)_\alpha\Big|\frac{1}{\varepsilon_n}\left(\nu_{w_{n,i}}\right)_3\right)\right|d\mathcal{H}^2\right] \\
&\quad\le \int_{A\times I} W_{\varepsilon_n}\left(x, \nabla_\alpha u_n\Big|\frac{1}{\varepsilon_n}\nabla_3 u_n\right)dx + \int_{S_{u_n}\cap(A\times I)}\left|\left(\left(\nu_{u_n}\right)_\alpha\Big|\frac{1}{\varepsilon_n}\left(\nu_{u_n}\right)_3\right)\right|d\mathcal{H}^2 \\
&\qquad\qquad + \frac{c}{M-m+1},
\end{aligned}$$

where $0<c<+\infty$ is given by

$$c = \beta\sup_{n\in\mathbb{N}}\|u_n\|_{L^1(A\times I;\mathbb{R}^3)}\sum_{i\ge1}e^{-i} + 2^p\beta\sup_{n\in\mathbb{N}}\left\|\left(\nabla_\alpha u_n\Big|\frac{1}{\varepsilon_n}\nabla_3 u_n\right)\right\|^p_{L^p(A\times I;\mathbb{R}^{3\times3})}.$$

We may find some $i_n\in\{m,\dots,M\}$ such that, setting $w_n := w_{n,i_n}$, then

$$\begin{aligned}
&\int_{A\times I} W_{\varepsilon_n}\left(x, \nabla_\alpha w_n\Big|\frac{1}{\varepsilon_n}\nabla_3 w_n\right)dx + \int_{S_{w_n}\cap(A\times I)}\left|\left(\left(\nu_{w_n}\right)_\alpha\Big|\frac{1}{\varepsilon_n}\left(\nu_{w_n}\right)_3\right)\right|d\mathcal{H}^2 \\
&\quad\le \int_{A\times I} W_{\varepsilon_n}\left(x, \nabla_\alpha u_n\Big|\frac{1}{\varepsilon_n}\nabla_3 u_n\right)dx + \int_{S_{u_n}\cap(A\times I)}\left|\left(\left(\nu_{u_n}\right)_\alpha\Big|\frac{1}{\varepsilon_n}\left(\nu_{u_n}\right)_3\right)\right|d\mathcal{H}^2 \\
&\qquad\qquad + \frac{c}{M-m+1}.
\end{aligned}\tag{6.8}$$

Moreover, in view of (6.4) and (6.5), $w_n \to u$ in $L^1(A \times I; \mathbb{R}^3)$, $\frac{1}{\varepsilon_n} \int_I \nabla_3 w_n(\cdot, x_3)\, dx_3 \rightharpoonup \overline{b}$ in $L^p(A; \mathbb{R}^3)$, and (6.3) implies that $\|w_n\|_{L^\infty(A \times I; \mathbb{R}^3)} \le e^{i_n+1} \le e^{M+1} =: C$. The proof is achieved passing to the limit as n tends to $+\infty$ in (6.8) and choosing M large enough so that $c/(M - m + 1) \le \eta$. ∎

Remark 6.3. As a consequence of Lemma 6.2, we get that for any $A \in \mathcal{R}(\omega)$ (or $A = \omega$), every $u \in BV(\Omega; \mathbb{R}^3) \cap L^\infty(\Omega; \mathbb{R}^3)$, and every $\overline{b} \in L^p(\omega; \mathbb{R}^3)$, then $\mathcal{E}(u, \overline{b}, A) = \mathcal{E}_\infty(u, \overline{b}, A)$.

Remark 6.4. A similar statement of Lemma 6.2 can be proved in the framework of Sobolev spaces, replacing $\mathcal{G}_{\varepsilon_n}$ by $\mathcal{J}_{\varepsilon_n}$.

Remark 6.5. Using a relaxation argument as in the proof of [4, Lemma 3.4] and Lemma 6.2, one can show that if $u \in SBV^p(\omega; \mathbb{R}^3) \cap L^\infty(\omega; \mathbb{R}^3)$ and if $\overline{b} \in L^p(\omega; \mathbb{R}^3)$, the value of $\mathcal{E}_\infty$ does not change when replacing W_{ε_n} by its quasi-convexification QW_{ε_n} defined in (5.12). The main point is that the diagonalization argument can still be used despite the weak $L^p(\omega; \mathbb{R}^3)$-convergence of the bending moment since the dual of $L^p(\omega; \mathbb{R}^3)$ is separable. Hence we may assume without loss of generality that W_ε is quasi-convex. In particular (see [15, Lemma 2.2, Chapter 4]), the following p-Lipschitz condition holds:

$$|W_\varepsilon(x, \xi_1) - W_\varepsilon(x, \xi_2)| \le c(1 + |\xi_1|^{p-1} + |\xi_2|^{p-1})|\xi_1 - \xi_2| \tag{6.9}$$

for all $\xi_1,\ \xi_2 \in \mathbb{R}^{3\times 3}$ and a.e. $x \in \Omega$.

Lemma 6.2 and Remark 6.3 are essential for the proof of the following result because they allow us to replace strong $L^1(\Omega; \mathbb{R}^3)$-convergence of any minimizing sequence by strong $L^p(\Omega; \mathbb{R}^3)$-convergence.

LEMMA 6.6. *For all $u \in SBV^p(\omega; \mathbb{R}^3) \cap L^\infty(\omega; \mathbb{R}^3)$ and all $\overline{b} \in L^p(\omega; \mathbb{R}^3)$, $\mathcal{E}_\infty(u, \overline{b}, \cdot)$ is the restriction to $\mathcal{A}(\omega)$ of a Radon measure absolutely continuous with respect to $\mathcal{L}^2 + \mathcal{H}^1 \llcorner S_u$.*

Proof. Let $u \in SBV^p(\omega; \mathbb{R}^3) \cap L^\infty(\omega; \mathbb{R}^3)$, $A \in \mathcal{A}(\omega)$ and assume first that $\overline{b}$ is smooth. Then taking $u_n(x_\alpha, x_3) := u(x_\alpha) + \varepsilon_n x_3 \overline{b}(x_\alpha)$ and $\overline{b}_n(x_\alpha) := \overline{b}(x_\alpha)$ as test functions for $\mathcal{E}_\infty(u, \overline{b}, A)$ and using the p-growth condition (5.1), we get that

$$\mathcal{E}_\infty(u, \overline{b}, A) \le \beta \int_A (1 + |\nabla_\alpha u|^p + |\overline{b}|^p)\, dx_\alpha + \mathcal{H}^1(S_u \cap A). \tag{6.10}$$

The same inequality holds for arbitrary functions $\overline{b} \in L^p(\omega; \mathbb{R}^3)$ thanks to the density of smooth maps into $L^p(\omega; \mathbb{R}^3)$ and the sequential weak lower semicontinuity of $\mathcal{E}_\infty(u, \cdot, A)$ in $L^p(A; \mathbb{R}^3)$. The remainder of the proof is very classical and is essentially the same as that of [4, Lemma 3.6]. As usual, the most delicate point is to prove the subadditivity of $\mathcal{E}_\infty(u, \overline{b}, \cdot)$, and this is done by gluing together suitable minimizing sequences by means of a cut-off function. The argument still works with the presence of the bending moment since the cut-off function is chosen independently of x_3. One should once more be careful when applying a diagonalization argument because of the weak convergence in L^p. As already mentioned in Remark 6.5, it is still allowed in the case where we include the bending moment since the dual of L^p is separable. ∎

As a consequence of Lemma 6.6 and Lebesgue's decomposition theorem, there exists a $\mathcal{L}^2$-measurable function f and a $\mathcal{H}^1 \llcorner S_u$-measurable function g such that for every $A \in \mathcal{A}(\omega)$,

$$\mathcal{E}_\infty(u, \overline{b}, A) = \int_A f\, d\mathcal{L}^2 + \int_{A \cap S_u} g\, d\mathcal{H}^1. \tag{6.11}$$

Since the measures $\mathcal{L}^2$ and $\mathcal{H}^1 \llcorner S_u$ are mutually singular, f is the Radon–Nikodým derivative of $\mathcal{E}_\infty(u,\bar{b},\cdot)$ with respect to $\mathcal{L}^2$,

$$f(x_0) = \lim_{\rho\to 0} \frac{\mathcal{E}_\infty(u,\bar{b},B'(x_0,\rho))}{\mathcal{L}^2(B'(x_0,\rho))} \quad \text{for } \mathcal{L}^2\text{-a.e. } x_0 \in \omega,$$

and g is the Radon–Nikodým derivative of $\mathcal{E}_\infty(u,\bar{b},\cdot)$ with respect to $\mathcal{H}^1 \llcorner S_u$,

$$g(x_0) = \lim_{\rho\to 0} \frac{\mathcal{E}_\infty(u,\bar{b},B'(x_0,\rho))}{\mathcal{H}^1(S_u \cap B'(x_0,\rho))} \quad \text{for } \mathcal{H}^1\text{-a.e. } x_0 \in S_u.$$

6.3. The upper bound. We first show the upper bound. To this end, we will use the locality property of the Γ-limit proved in the previous subsection when $u \in BV(\Omega;\mathbb{R}^3) \cap L^\infty(\Omega;\mathbb{R}^3)$ and the analog Γ-convergence result in Sobolev spaces (Theorem 5.1).

LEMMA 6.7. *For all $u \in BV(\Omega;\mathbb{R}^3)$ and all $\bar{b} \in L^p(\omega;\mathbb{R}^3)$, $\mathcal{E}(u,\bar{b},\omega) \le \mathcal{G}(u,\bar{b})$.*

Proof. It is enough to consider the case where $\mathcal{G}(u,\bar{b}) < +\infty$ and thus $u \in SBV^p(\omega;\mathbb{R}^3)$. In fact, we will first restrict ourselves to the case where $u \in L^\infty(\omega;\mathbb{R}^3) \cap SBV^p(\omega;\mathbb{R}^3)$ because, thanks to Remark 6.3, it allows us to replace $\mathcal{E}$ by $\mathcal{E}_\infty$. According to (6.11) and the definition of $\mathcal{G}$, we must show that $g(x_0) \le 1$ for $\mathcal{H}^1$-a.e. $x_0 \in S_u$ and $f(x_0) \le W^*(x_0, \nabla_\alpha u(x_0)|\bar{b}(x_0))$ for $\mathcal{L}^2$-a.e. $x_0 \in \omega$.

Let us first treat the surface term. By virtue of (6.10) with $A = B'(x_0,\rho)$, we have that for $\mathcal{H}^1$-a.e. $x_0 \in S_u$,

$$\begin{aligned} g(x_0) &= \lim_{\rho\to 0} \frac{\mathcal{E}_\infty(u,\bar{b},B'(x_0,\rho))}{\mathcal{H}^1(S_u \cap B'(x_0,\rho))} \\ &\le \limsup_{\rho\to 0} \frac{1}{\mathcal{H}^1(S_u \cap B'(x_0,\rho))} \Bigg\{ \beta \int_{B'(x_0,\rho)} (1 + |\nabla_\alpha u|^p + |\bar{b}|^p)\, dx_\alpha \\ &\qquad\qquad + \mathcal{H}^1(S_u \cap B'(x_0,\rho)) \Bigg\} \\ &= \limsup_{\rho\to 0} \frac{\mu(B'(x_0,\rho))}{\mathcal{H}^1(S_u \cap B'(x_0,\rho))} + 1, \end{aligned}$$

where we set $\mu := \beta(1 + |\nabla_\alpha u|^p + |\bar{b}|^p)\mathcal{L}^2$. But since μ and $\mathcal{H}^1 \llcorner S_u$ are mutually singular, we have for $\mathcal{H}^1$-a.e. $x_0 \in S_u$

$$\lim_{\rho\to 0} \frac{\mu(B'(x_0,\rho))}{\mathcal{H}^1(S_u \cap B'(x_0,\rho))} = 0,$$

which shows that $g(x_0) \le 1$ for $\mathcal{H}^1$-a.e. $x_0 \in S_u$.

Concerning the bulk term, choose $x_0 \in \omega$ to be a Lebesgue point of u, $\nabla_\alpha u$, $\bar{b}$, and $W^*(\cdot,\nabla_\alpha u(\cdot)|\bar{b}(\cdot))$ and such that

$$\lim_{\rho\to 0} \frac{\mathcal{H}^1(S_u \cap B'(x_0,\rho))}{\mathcal{L}^2(B'(x_0,\rho))} = 0. \tag{6.12}$$

Remark that $\mathcal{L}^2$-a.e. points x_0 in ω satisfy these properties and set $u_0(x_\alpha) := \nabla_\alpha u(x_0)\, x_\alpha$ and $\bar{b}_0(x_\alpha) := \bar{b}(x_0)$. For every $\rho > 0$, Theorem 5.1 implies the existence of a sequence $\{v_n^\rho\} \subset W^{1,p}(B'(x_0,\rho)\times I;\mathbb{R}^3)$ such that $v_n^\rho \to u_0$ in $L^p(B'(x_0,\rho)\times I;\mathbb{R}^3)$

(thus in $L^1(B'(x_0,\rho)\times I;\mathbb{R}^3)$ as well), $\frac{1}{\varepsilon_n}\int_I \nabla_3 v_n^\rho(\cdot,x_3)\,dx_3 \rightharpoonup \bar b_0$ in $L^p(B'(x_0,\rho);\mathbb{R}^3)$, and

$$\lim_{n\to+\infty}\int_{B'(x_0,\rho)\times I} W_{\varepsilon_n}\left(x,\nabla_\alpha v_n^\rho\Big|\frac{1}{\varepsilon_n}\nabla_3 v_n^\rho\right)dx = \int_{B'(x_0,\rho)} W^*(x_\alpha,\nabla_\alpha u(x_0)|\bar b(x_0))\,dx_\alpha.$$

Since $u_0\in L^\infty(\omega;\mathbb{R}^3)$, by Lemma 6.2 and Remark 6.4, for any $\eta>0$ we can find a sequence $\{w_n^\rho\}\subset W^{1,p}(B'(x_0,\rho)\times I;\mathbb{R}^3)$ and $C_\rho>0$ such that

$$\sup_{n\in\mathbb{N}}\|w_n^\rho\|_{L^\infty(B'(x_0,\rho)\times I;\mathbb{R}^3)}\le C_\rho,$$

$w_n^\rho\to u_0$ in $L^p(B'(x_0,\rho)\times I;\mathbb{R}^3)$, $\frac{1}{\varepsilon_n}\int_I\nabla_3 w_n^\rho(\cdot,x_3)\,dx_3\rightharpoonup \bar b_0$ in $L^p(B'(x_0,\rho);\mathbb{R}^3)$, and

$$\limsup_{n\to+\infty}\int_{B'(x_0,\rho)\times I} W_{\varepsilon_n}\left(x,\nabla_\alpha w_n^\rho\Big|\frac{1}{\varepsilon_n}\nabla_3 w_n^\rho\right)dx \le \int_{B'(x_0,\rho)} W^*(x_\alpha,\nabla_\alpha u(x_0)|\bar b(x_0))\,dx_\alpha + \mathcal{L}^2(B'(x_0,\rho))\eta.$$

Thanks to (5.5) and the separately convex character of $W^*(x_0,\cdot|\cdot)$ (see the proof of Theorem 5.1), it follows that $W^*(x_0,\cdot|\cdot)$ is p-Lipschitz. Thus our choice of x_0 implies that

$$\limsup_{\rho\to 0}\limsup_{n\to+\infty}⨍_{B'(x_0,\rho)\times I} W_{\varepsilon_n}\left(x,\nabla_\alpha w_n^\rho\Big|\frac{1}{\varepsilon_n}\nabla_3 w_n^\rho\right)dx \le W^*(x_0,\nabla_\alpha u(x_0)|\bar b(x_0))+\eta, \tag{6.13}$$

and from the coercivity condition (5.1), we get

$$\sup_{\rho>0,\,n\in\mathbb{N}}⨍_{B'(x_0,\rho)\times I}\left|\left(\nabla_\alpha w_n^\rho\Big|\frac{1}{\varepsilon_n}\nabla_3 w_n^\rho\right)\right|^p dx<+\infty. \tag{6.14}$$

Let $\bar b_k\in\mathcal{C}_c^\infty(\omega;\mathbb{R}^3)$ be such that $\bar b_k\to\bar b$ in $L^p(\omega;\mathbb{R}^3)$ and define

$$u_{n,k}^\rho(x) := u(x_\alpha)+\varepsilon_n x_3(\bar b_k(x_\alpha)-\bar b(x_0))+w_n^\rho(x_\alpha,x_3)-\nabla_\alpha u(x_0)\,x_\alpha.$$

Then we have that $u_{n,k}^\rho\to u$ in $L^1(B'(x_0,\rho)\times I;\mathbb{R}^3)$, $\frac{1}{\varepsilon_n}\int_I\nabla_3 u_{n,k}^\rho(\cdot,x_3)\,dx_3\rightharpoonup\bar b_k$ in $L^p(B'(x_0,\rho);\mathbb{R}^3)$ as $n\to+\infty$, and $\sup_n\|u_{n,k}^\rho\|_{L^\infty(B'(x_0,\rho)\times I;\mathbb{R}^3)}<+\infty$. Thus, since $S_{u_{n,k}^\rho}\cap(B'(x_0,\rho)\times I)=(S_u\cap B'(x_0,\rho))\times I$, we get that

$$\begin{aligned}\mathcal{E}_\infty(u,\bar b_k,B'(x_0,\rho)) &\le \liminf_{n\to+\infty}\left\{\int_{B'(x_0,\rho)\times I} W_{\varepsilon_n}\left(x,\nabla_\alpha u_{n,k}^\rho\Big|\frac{1}{\varepsilon_n}\nabla_3 u_{n,k}^\rho\right)dx\right.\\ &\qquad\left.+\int_{S_{u_{n,k}^\rho}\cap(B'(x_0,\rho)\times I)}\left|\left(\left(\nu_{u_{n,k}^\rho}\right)_\alpha\Big|\frac{1}{\varepsilon_n}\left(\nu_{u_{n,k}^\rho}\right)_3\right)\right|d\mathcal{H}^2\right\}\\ &\le \liminf_{n\to+\infty}\int_{B'(x_0,\rho)\times I} W_{\varepsilon_n}\Big(x,\nabla_\alpha u(x_\alpha)-\nabla_\alpha u(x_0)+\nabla_\alpha w_n^\rho(x)\\ &\qquad+\varepsilon_n x_3\nabla_\alpha\bar b_k(x_\alpha)\Big|\frac{1}{\varepsilon_n}\nabla_3 w_n^\rho(x)+\bar b_k(x_\alpha)-\bar b(x_0)\Big)\,dx\\ &\qquad+\mathcal{H}^1(S_u\cap B'(x_0,\rho)).\end{aligned}$$

Thus from (6.12), we obtain

$$f(x_0) \leq \liminf_{\rho\to 0} \liminf_{k\to+\infty} \liminf_{n\to+\infty} \fint_{B'(x_0,\rho)\times I} W_{\varepsilon_n}\Big(x, \nabla_\alpha u(x_\alpha) - \nabla_\alpha u(x_0) + \nabla_\alpha w_n^\rho(x) + \varepsilon_n x_3 \nabla_\alpha \bar{b}_k(x_\alpha)\Big| \frac{1}{\varepsilon_n}\nabla_3 w_n^\rho(x) + \bar{b}_k(x_\alpha) - \bar{b}(x_0)\Big)\, dx.$$

Relations (6.9), (6.13), (6.14), and Hölder's inequality yield

$$\begin{aligned} f(x_0) \leq \liminf_{\rho\to 0} \liminf_{k\to+\infty} \liminf_{n\to+\infty} \Bigg\{ & \fint_{B'(x_0,\rho)\times I} W_{\varepsilon_n}\left(x, \nabla_\alpha w_n^\rho \Big| \frac{1}{\varepsilon_n}\nabla_3 w_n^\rho\right) dx \\ & + c \fint_{B'(x_0,\rho)\times I} \Big(1 + |\nabla_\alpha u(x_\alpha) - \nabla_\alpha u(x_0)|^{p-1} + |\bar{b}_k(x_\alpha) - \bar{b}(x_0)|^{p-1} \\ & + \left|\left(\nabla_\alpha w_n^\rho(x)\Big|\frac{1}{\varepsilon_n}\nabla_3 w_n^\rho(x)\right)\right|^{p-1} + \varepsilon_n^{p-1}|\nabla_\alpha \bar{b}_k(x_\alpha)|^{p-1}\Big) \\ & \times\Big(|\nabla_\alpha u(x_\alpha) - \nabla_\alpha u(x_0)| + \varepsilon_n|\nabla_\alpha \bar{b}_k(x_\alpha)| + |\bar{b}_k(x_\alpha) - \bar{b}(x_0)|\Big)\, dx \Bigg\} \\ \leq{} & W^*(x_0, \nabla_\alpha u(x_0)|\bar{b}(x_0)) + \eta \\ & + c \limsup_{\rho\to 0} \left\{ \fint_{B'(x_0,\rho)} \left(1 + |\nabla_\alpha u - \nabla_\alpha u(x_0)|^p + |\bar{b} - \bar{b}(x_0)|^p\right) dx_\alpha \right\}^{(p-1)/p} \\ & \times \left\{ \fint_{B'(x_0,\rho)} \left(|\nabla_\alpha u - \nabla_\alpha u(x_0)|^p + |\bar{b} - \bar{b}(x_0)|^p\right) dx_\alpha \right\}^{1/p}. \end{aligned}$$

Thanks to our choice of x_0 and letting $\eta \to 0$, we conclude that

$$f(x_0) \leq W^*(x_0, \nabla_\alpha u(x_0)|\bar{b}(x_0))$$

for $\mathcal{L}^2$-a.e. $x_0 \in \omega$, which completes the proof in the case where $u \in L^\infty(\omega;\mathbb{R}^3) \cap SBV^p(\omega;\mathbb{R}^3)$. The general case can in turn be treated by approximation exactly as in the proof of [4, Lemma 3.8]. □

6.4. The lower bound. Let us now prove the lower bound. The proof is essentially based on Theorem 4.1 and a blow-up argument.

LEMMA 6.8. *For all $u \in BV(\Omega;\mathbb{R}^3)$ and all $\bar{b} \in L^p(\omega;\mathbb{R}^3)$, $\mathcal{E}(u,\bar{b},\omega) \geq \mathcal{G}(u,\bar{b})$.*

Proof. It is not restrictive to assume that $\mathcal{E}(u,\bar{b},\omega) < +\infty$. By Γ-convergence, there exists a recovery sequence $\{u_n\} \subset SBV^p(\Omega;\mathbb{R}^3)$ such that $u_n \to u$ in $L^1(\Omega;\mathbb{R}^3)$, $\frac{1}{\varepsilon_n}\int_I \nabla_3 u_n(\cdot,x_3)\,dx_3 \rightharpoonup \bar{b}$ in $L^p(\omega;\mathbb{R}^3)$, and

$$\lim_{n\to+\infty}\left[\int_\Omega W_{\varepsilon_n}\left(x, \nabla_\alpha u_n \Big| \frac{1}{\varepsilon_n}\nabla_3 u_n\right) dx + \int_{S_{u_n}} \left|\left(\left(\nu_{u_n}\right)_\alpha \Big| \frac{1}{\varepsilon_n}\left(\nu_{u_n}\right)_3\right)\right| d\mathcal{H}^2\right] = \mathcal{E}(u,\bar{b},\omega). \tag{6.15}$$

Arguing exactly as in the proof of [4, Lemma 3.9], we can actually show that $u \in SBV^p(\omega;\mathbb{R}^3)$ and that $u_n \rightharpoonup u$ in $SBV^p(\Omega;\mathbb{R}^3)$. Now for every Borel set $E \subset \omega$, define the following sequences of Radon measures:

$$\lambda_n(E) := W_{\varepsilon_n}\left(\cdot, \nabla_\alpha u_n \Big| \frac{1}{\varepsilon_n}\nabla_3 u_n\right)\mathcal{L}^3 \llcorner (E \times I) \\ + \left|\left(\left(\nu_{u_n}\right)_\alpha \Big| \frac{1}{\varepsilon_n}\left(\nu_{u_n}\right)_3\right)\right| \mathcal{H}^2 \llcorner (S_{u_n} \cap (E \times I))$$

and

$$\mu_n(E) := \left|\left(\left(\nu_{u_n}\right)_\alpha \Big| \frac{1}{\varepsilon_n}\left(\nu_{u_n}\right)_3\right)\right| \mathcal{H}^2 \llcorner (S_{u_n} \cap (E \times I)).$$

Then for a subsequence (not relabeled), there exist nonnegative and finite Radon measures λ and $\mu \in \mathcal{M}(\omega)$ such that $\lambda_n \overset{*}{\rightharpoonup} \lambda$ and $\mu_n \overset{*}{\rightharpoonup} \mu$ in $\mathcal{M}(\omega)$. By the Besicovitch differentiation theorem [3, Theorem 2.22], one can find three mutually singular nonnegative Radon measures λ^a, λ^j, and λ^c such that $\lambda = \lambda^a + \lambda^j + \lambda^c$, where $\lambda^a \ll \mathcal{L}^2$ and $\lambda^j \ll \mathcal{H}^1 \llcorner S_u$. It is enough to check that

$$\frac{d\lambda^j}{d\mathcal{H}^1 \llcorner S_u}(x_0) \geq 1 \quad \text{for } \mathcal{H}^1\text{-a.e. } x_0 \in S_u \tag{6.16}$$

and

$$\frac{d\lambda^a}{d\mathcal{L}^2}(x_0) \geq W^*(x_0, \nabla_\alpha u(x_0)|\bar{b}(x_0)) \quad \text{for } \mathcal{L}^2\text{-a.e. } x_0 \in \omega. \tag{6.17}$$

Indeed, if (6.16) and (6.17) hold, we obtain from (6.15) that

$$\begin{aligned} \mathcal{E}(u, \bar{b}, \omega) &\geq \lambda(\omega) = \lambda^a(\omega) + \lambda^j(\omega) + \lambda^c(\omega) \\ &\geq \int_\omega W^*(x_\alpha, \nabla_\alpha u|\bar{b})\, dx_\alpha + \mathcal{H}^1(S_u) = \mathcal{G}(u, \bar{b}). \end{aligned}$$

We first prove (6.16). Fix a point $x_0 \in S_u$ such that

$$\frac{d\lambda^j}{d\mathcal{H}^1 \llcorner S_u}(x_0) = \frac{d\lambda}{d\mathcal{H}^1 \llcorner S_u}(x_0)$$

exists and is finite and remark that $\mathcal{H}^1$-a.e. points in S_u satisfy this property. Let $\{\rho_k\} \searrow 0^+$ be such that $\lambda(\partial B'(x_0, \rho_k)) = 0$ for each $k \in \mathbb{N}$. Then

$$\begin{aligned} \frac{d\lambda}{d\mathcal{H}^1 \llcorner S_u}(x_0) &= \lim_{k\to+\infty} \frac{\lambda(B'(x_0,\rho_k))}{\mathcal{H}^1(S_u \cap B'(x_0,\rho_k))} \\ &= \lim_{k\to+\infty} \lim_{n\to+\infty} \frac{\lambda_n(B'(x_0,\rho_k))}{\mathcal{H}^1(S_u \cap B'(x_0,\rho_k))} \\ &\geq \liminf_{k\to+\infty} \liminf_{n\to+\infty} \frac{\mathcal{H}^2(S_{u_n} \cap (B'(x_0,\rho_k) \times I))}{\mathcal{H}^1(S_u \cap B'(x_0,\rho_k))}. \end{aligned}$$

By [3, Theorem 4.36], we have that

$$\liminf_{n\to+\infty} \mathcal{H}^2(S_{u_n} \cap (B'(x_0,\rho_k) \times I)) \geq \mathcal{H}^1(S_u \cap B'(x_0,\rho_k));$$

hence we obtain (6.16).

Let us prove that (6.17) holds at every point $x_0 \in \omega \setminus N$ (where $N \subset \omega$ is the exceptional set introduced in Proposition 5.2) which is a Lebesgue point of both $\nabla_\alpha u$ and $\bar{b}$, a point of approximate differentiability of u such that

$$\frac{d\lambda^a}{d\mathcal{L}^2}(x_0) = \frac{d\lambda}{d\mathcal{L}^2}(x_0)$$

exists and is finite and which satisfies

$$\lim_{\rho \to 0} \frac{\mu(B'(x_0,\rho))}{2\rho} = 0. \tag{6.18}$$

It turns out that $\mathcal{L}^2$-a.e. points x_0 in ω satisfy these properties. Indeed, the verification of (6.18) is similar to the one of (3.2) used in the proof of Theorem 3.2. As before, let $\{\rho_k\} \searrow 0^+$ be such that $\lambda(\partial B'(x_0,\rho_k)) = 0$ for every $k \in \mathbb{N}$; then

$$\begin{aligned}
\frac{d\lambda}{d\mathcal{L}^2}(x_0) &= \lim_{k\to+\infty} \frac{\lambda(B'(x_0,\rho_k))}{\mathcal{L}^2(B'(x_0,\rho_k))} \\
&= \lim_{k\to+\infty} \lim_{n\to+\infty} \frac{\lambda_n(B'(x_0,\rho_k))}{\mathcal{L}^2(B'(x_0,\rho_k))} \\
&\geq \limsup_{k\to+\infty} \limsup_{n\to+\infty} \fint_{B'(x_0,\rho_k)\times I} W_{\varepsilon_n}\left(x, \nabla_\alpha u_n \middle| \frac{1}{\varepsilon_n}\nabla_3 u_n\right) dx \\
&= \limsup_{k\to+\infty} \limsup_{n\to+\infty} \fint_{B'\times I} W_{\varepsilon_n}\left(x_0+\rho_k x_\alpha, x_3, \nabla_\alpha u_{n,k} \middle| \frac{\rho_k}{\varepsilon_n}\nabla_3 u_{n,k}\right) dx, \qquad (6.19)
\end{aligned}$$

where $u_{n,k}(x_\alpha,x_3) = [u_n(x_0+\rho_k x_\alpha, x_3) - u(x_0)]/\rho_k$. Since x_0 is a point of approximate differentiability of u, we have that

$$\lim_{k\to+\infty} \lim_{n\to+\infty} \int_{B'\times I} |u_{n,k}(x) - \nabla_\alpha u(x_0) x_\alpha|\, dx = 0, \tag{6.20}$$

and using the fact that x_0 is a Lebesgue point of $\bar{b}$, for every $v \in L^{p'}(B';\mathbb{R}^3)$ we get that

$$\lim_{k\to+\infty} \lim_{n\to+\infty} \int_{B'} \left(\frac{\rho_k}{\varepsilon_n}\int_I \nabla_3 u_{n,k}(x_\alpha,x_3)\, dx_3\right)\cdot v(x_\alpha)\, dx_\alpha = \int_{B'} \bar{b}(x_0)\cdot v\, dx_\alpha. \tag{6.21}$$

Changing variables in the surface term and thanks to (6.18), it yields

$$\begin{aligned}
&\limsup_{k\to+\infty} \limsup_{n\to+\infty} \int_{S_{u_{n,k}}\cap(B'\times I)} \left| \left(\left(\nu_{u_{n,k}}\right)_\alpha \middle| \frac{\rho_k}{\varepsilon_n}\left(\nu_{u_{n,k}}\right)_3 \right) \right| d\mathcal{H}^2 \\
&\quad = \limsup_{k\to+\infty} \limsup_{n\to+\infty} \frac{1}{\rho_k} \int_{S_{u_n}\cap(B'(x_0,\rho_k)\times I)} \left| \left(\left(\nu_{u_n}\right)_\alpha \middle| \frac{1}{\varepsilon_n}\left(\nu_{u_n}\right)_3 \right) \right| d\mathcal{H}^2 \\
&\quad \leq \limsup_{k\to+\infty} \limsup_{n\to+\infty} \frac{\mu_n(\overline{B'(x_0,\rho_k)})}{\rho_k} \leq \limsup_{k\to+\infty} \frac{\mu(\overline{B'(x_0,\rho_k)})}{\rho_k} = 0 \qquad (6.22)
\end{aligned}$$

because $\mu(\partial B'(x_0,\rho_k)) \le \lambda(\partial B'(x_0,\rho_k)) = 0$. Set

$$M := \max\left\{\left(\frac{\mathcal{L}^2(B')}{\beta'}\left(\left|\frac{d\lambda}{d\mathcal{L}^2}(x_0)\right|+1\right)\right)^{1/p}, |\bar{b}(x_0)|\mathcal{L}^2(B')^{1/p}\right\} < +\infty. \tag{6.23}$$

From (6.19)–(6.22), using a diagonalization argument, the fact that $L^{p'}(B';\mathbb{R}^3)$ is separable, and Remark 5.3, we can find a sequence $n(k) \nearrow +\infty$ such that, setting $\delta_k := \varepsilon_{n(k)}/\rho_k$, $v_k := u_{n(k),k}$, $u_0(x_\alpha) := \nabla u(x_0)\,x_\alpha$, and $\bar{b}_0(x_\alpha) := \bar{b}(x_0)$, then $\delta_k \to 0$, $v_k \to u_0$ in $L^1(B'\times I;\mathbb{R}^3)$, $\frac{1}{\delta_k}\int_I \nabla_3 v_k(\cdot,x_3)\,dx_3 \rightharpoonup \bar{b}_0$ in $L^p(B';\mathbb{R}^3)$,

$$\lim_{k\to+\infty}\int_{S_{v_k}}\left|\left((\nu_{v_k})_\alpha\Big|\frac{1}{\delta_k}(\nu_{v_k})_3\right)\right|d\mathcal{H}^2 = 0, \tag{6.24}$$

$$\frac{d\lambda}{d\mathcal{L}^2}(x_0) \ge \limsup_{k\to+\infty}\frac{1}{\mathcal{L}^2(B')}\int_{B'\times I} W_{\varepsilon_{n(k)}}\left(x_0+\rho_k x_\alpha, x_3, \nabla_\alpha v_k\Big|\frac{1}{\delta_k}\nabla_3 v_k\right)dx, \tag{6.25}$$

and for every $(u,\bar{b}) \in L^p(B'\times I;\mathbb{R}^3)\times L^p(B';\mathbb{R}^3)$ with $\|\bar{b}\|_{L^p(B';\mathbb{R}^3)} \le M$, the Γ-limit in $L^p_s(B'\times I;\mathbb{R}^3)\times L^p_w(B';\mathbb{R}^3)$ of

$$\begin{cases} \displaystyle\int_{B'\times I} W_{\varepsilon_{n(k)}}\left(x_0+\rho_k x_\alpha, x_3, \nabla_\alpha u\Big|\frac{1}{\delta_k}\nabla_3 u\right)dx & \text{if } \begin{cases} u \in W^{1,p}(B'\times I;\mathbb{R}^3),\\ \bar{b} = \frac{1}{\delta_k}\int_I \nabla_3 u(\cdot,x_3)\,dx_3, \end{cases}\\ +\infty & \text{otherwise} \end{cases}$$

coincides with

$$\begin{cases} \displaystyle\int_{B'} W^*(x_0,\nabla_\alpha u|\bar{b})\,dx_\alpha & \text{if } u \in W^{1,p}(B';\mathbb{R}^3),\\ +\infty & \text{otherwise.} \end{cases}$$

From (6.24), (6.25), and (a slight variant of) Lemma 6.2, for any $0<\eta<1$, there exist a constant $C>0$ and $\{w_k\}\subset SBV^p(B'\times I;\mathbb{R}^3)$ such that $w_k \to u_0$ in $L^1(B'\times I;\mathbb{R}^3)$, $\frac{1}{\delta_k}\int_I \nabla_3 w_k(\cdot,x_3)\,dx_3 \rightharpoonup \bar{b}_0$ in $L^p(B';\mathbb{R}^3)$, $\sup_k \|w_k\|_{L^\infty(B'\times I;\mathbb{R}^3)} \le C$,

$$\lim_{k\to+\infty}\int_{S_{w_k}}\left|\left((\nu_{w_k})_\alpha\Big|\frac{1}{\delta_k}(\nu_{w_k})_3\right)\right|d\mathcal{H}^2 = 0,$$

and

$$\frac{d\lambda}{d\mathcal{L}^2}(x_0) \ge \limsup_{k\to+\infty}\frac{1}{\mathcal{L}^2(B')}\int_{B'\times I} W_{\varepsilon_{n(k)}}\left(x_0+\rho_k x_\alpha, x_3, \nabla_\alpha w_k\Big|\frac{1}{\delta_k}\nabla_3 w_k\right)dx - \eta.$$

From the p-coercivity condition (5.1) and [3, Theorem 4.36], the sequence $\{w_k\}$ converges weakly to u in $SBV^p(\Omega;\mathbb{R}^3)$, and it fulfills the assumptions of Theorem 4.1. Thus, for a not relabeled subsequence, one can find another sequence $\{z_k\}\subset W^{1,\infty}(B'\times I;\mathbb{R}^3)$ such that $z_k \rightharpoonup u_0$ in $W^{1,p}(B'\times I;\mathbb{R}^3)$, $\frac{1}{\delta_k}\int_I \nabla_3 z_k(\cdot,x_3)\,dx_3 \rightharpoonup \bar{b}_0$ in $L^p(B';\mathbb{R}^3)$, $\left\{\left|\left(\nabla_\alpha z_k|\frac{1}{\delta_k}\nabla_3 z_k\right)\right|^p\right\}$ is equi-integrable, and $\mathcal{L}^3(\{z_k \ne w_k\}\cup\{\nabla z_k \ne \nabla w_k\}) \to 0$. Hence

$$\frac{d\lambda}{d\mathcal{L}^2}(x_0) \ge \limsup_{k\to+\infty}\frac{1}{\mathcal{L}^2(B')}\int_{\{w_k=z_k\}} W_{\varepsilon_{n(k)}}\left(x_0+\rho_k x_\alpha, x_3, \nabla_\alpha z_k\Big|\frac{1}{\delta_k}\nabla_3 z_k\right)dx - \eta,$$

and using the p-growth condition (5.1), the fact that $\left\{\left|\left(\nabla_\alpha z_k\middle|\frac{1}{\delta_k}\nabla_3 z_k\right)\right|^p\right\}$ is equi-integrable, and the fact that $\mathcal{L}^3(\{z_k \neq w_k\}) \to 0$ we get that

$$\limsup_{k\to+\infty} \int_{\{w_k\neq z_k\}} W_{\varepsilon_{n(k)}}\left(x_0+\rho_k x_\alpha, x_3, \nabla_\alpha z_k\middle|\frac{1}{\delta_k}\nabla_3 z_k\right)dx = 0.$$

As a consequence

$$\frac{d\lambda}{d\mathcal{L}^2}(x_0) \geq \limsup_{k\to+\infty} \frac{1}{\mathcal{L}^2(B')}\int_{B'\times I} W_{\varepsilon_{n(k)}}\left(x_0+\rho_k x_\alpha, x_3, \nabla_\alpha z_k\middle|\frac{1}{\delta_k}\nabla_3 z_k\right)dx - \eta,$$

and by the p-coercivity condition (5.1) and (6.23),

$$\left\|\frac{1}{\delta_k}\int_I \nabla_3 z_k(\cdot,x_3)\,dx_3\right\|_{L^p(B';\mathbb{R}^3)} \leq M, \quad \|\bar{b}_0\|_{L^p(B';\mathbb{R}^3)} \leq M.$$

Thus by our choice of the subsequence $n(k)$ and Remark 5.3, we get that

$$\frac{d\lambda}{d\mathcal{L}^2}(x_0) \geq W^*(x_0, \nabla_\alpha u(x_0)|\bar{b}(x_0)) - \eta.$$

Letting η tend to zero completes the proof of (6.17). □

Remark 6.9. Note that it seems difficult to think of applying the decoupling variable method introduced in [7] and further developed in [5, 6]. Indeed, this generalized framework has the drawback that we have no information on the way that W_ε depends on ε, and it requires application of such abstract results as metrizability of Γ-convergence. Remark also that the same kind of blow-up argument considered here could have been used in [5, 6, 7] in place of the decoupling variable method, in order to treat the presence of the spatial variable.

7. Case without bending moment. In this last section, we deduce from Theorem 6.1 a similar result without the presence of the bending moment. Define $\mathcal{I}_\varepsilon : L^p(\Omega;\mathbb{R}^3) \to [0,+\infty]$ by

$$\mathcal{I}_\varepsilon(u) := \begin{cases} \displaystyle\int_\Omega W_\varepsilon\left(x, \nabla_\alpha u\middle|\frac{1}{\varepsilon}\nabla_3 u\right)dx & \text{if } u \in W^{1,p}(\Omega;\mathbb{R}^3), \\ +\infty & \text{otherwise.} \end{cases}$$

In [12, Theorem 2.5], the following integral representation result has been proved.

THEOREM 7.1. *For every sequence* $\{\varepsilon_n\} \searrow 0^+$, *there exist a subsequence (not relabeled) and a Carathéodory function* $\widehat{W} : \omega \times \mathbb{R}^{3\times 2} \to [0,+\infty)$ *(depending on the subsequence) such that the sequence* $\mathcal{I}_{\varepsilon_n}$ Γ*-converges in* $L^p_s(\Omega;\mathbb{R}^3)$ *to* $\mathcal{I}$, *where*

$$\mathcal{I}(u) = \begin{cases} \displaystyle\int_\omega \widehat{W}(x_\alpha, \nabla_\alpha u)\,dx_\alpha & \textit{if } u \in W^{1,p}(\omega;\mathbb{R}^3), \\ +\infty & \textit{otherwise.} \end{cases}$$

We refer the reader to [22, 12, 7, 5, 6] for more explicit formulas in particular cases.

Remark 7.2. As it has been pointed out in [7] in the case where W_ε was independent of ε (see also [9]), it can still be seen here that

$$\widehat{W}(x_0,\bar{\xi}) = \min_{z\in\mathbb{R}^3} W^*(x_0,\bar{\xi}|z)$$

for all $\bar{\xi} \in \mathbb{R}^{3\times 2}$ and a.e. $x_0 \in \omega$.

Now define $\mathcal{F}_\varepsilon : BV(\Omega;\mathbb{R}^3) \to [0,+\infty]$ by

$$\mathcal{F}_\varepsilon(u) := \begin{cases} \displaystyle\int_\Omega W_\varepsilon\left(x, \nabla_\alpha u \Big| \frac{1}{\varepsilon}\nabla_3 u\right) dx & \\ \qquad\qquad + \displaystyle\int_{S_u} \left| \left((\nu_u)_\alpha \left| \frac{1}{\varepsilon} (\nu_u)_3 \right.\right) \right| d\mathcal{H}^2 & \text{if } u \in SBV^p(\Omega;\mathbb{R}^3), \\ +\infty & \text{otherwise.} \end{cases}$$

As a consequence of Theorem 6.1, Theorem 7.1, Remark 7.2, and a standard measurability selection criterion (see, e.g., [17, Theorem 1.2, Chapter VIII]) we get the following integral representation result for dimension reduction problems in SBV without bending moment.

THEOREM 7.3. *For every sequence $\{\varepsilon_n\} \searrow 0^+$, there exists a subsequence, still labeled $\{\varepsilon_n\}$, such that $\mathcal{F}_{\varepsilon_n}$ Γ-converges in $L^1_s(\Omega;\mathbb{R}^3)$ to $\mathcal{F} : BV(\Omega;\mathbb{R}^3) \to [0,+\infty]$ defined by*

$$\mathcal{F}(u) := \begin{cases} \displaystyle\int_\omega \widehat{W}(x_\alpha, \nabla_\alpha u)\, dx_\alpha + \mathcal{H}^1(S_u) & \textit{if } u \in SBV^p(\omega;\mathbb{R}^3), \\ +\infty & \textit{otherwise,} \end{cases}$$

where $\widehat{W}$ is given by Theorem 7.1.

Acknowledgments. The author wishes to thank Irene Fonseca and Gilles Francfort for having drawn this problem to his attention.

REFERENCES

[1] E. Acerbi and N. Fusco, *Semicontinuity problems in the calculus of variations*, Arch. Rational Mech. Anal., 86 (1984), pp. 125–145.
[2] L. Ambrosio, *On the lower semicontinuity of quasi-convex integrals in SBV*, Nonlinear Anal., 23 (1994), pp. 405–425.
[3] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press, Oxford, UK, 2000.
[4] J.-F. Babadjian, *Quasistatic evolution of a brittle thin film*, Calc. Var. Partial Differential Equations, 26 (2006), pp. 69–118.
[5] J.-F. Babadjian and M. Baía, *3D-2D analysis of a thin film with periodic microstructure*, Proc. Roy. Soc. Edinburgh Sect. A, 136 (2006), pp. 223–243.
[6] J.-F. Babadjian and M. Baía, *Multiscale nonconvex relaxation and application to thin films*, Asymptot. Anal., 48 (2006), pp. 173–218.
[7] J.-F. Babadjian and G. A. Francfort, *Spatial heterogeneity in 3D-2D dimensional reduction*, ESAIM Control Optim. Calc. Var., 11 (2005), pp. 139–160.
[8] M. Bocea and I. Fonseca, *Equi-integrability results for 3D-2D dimension reduction problems*, ESAIM Control Optim. Calc. Var., 7 (2002), pp. 443–470.
[9] G. Bouchitté, I. Fonseca, and L. Mascarenhas, *Bending moment in membrane theory*, J. Elasticity, 73 (2003), pp. 75–99.
[10] G. Bouchitté, I. Fonseca, G. Leoni, and L. Mascarenhas, *A global method for relaxation in $W^{1,p}$ and in SBV_p*, Arch. Rational Mech. Anal., 165 (2002), pp. 187–242.
[11] A. Braides and I. Fonseca, *Brittle thin films*, Appl. Math. Optim., 44 (2001), pp. 299–323.
[12] A. Braides, I. Fonseca, and G. A. Francfort, *3D-2D asymptotic analysis for inhomogeneous thin films*, Indiana Univ. Math. J., 49 (2000), pp. 1367–1404.
[13] A. Braides and C. I. Zeppieri, *A note on equi-integrability in dimension reduction problems*, Calc. Var. Partial Differential Equations, 29 (2007), pp. 231–238.
[14] G. Buttazzo and G. Dal Maso, *Integral representation and relaxation of local functionals*, Nonlinear Anal., 9 (1985), pp. 515–532.
[15] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, 1989.
[16] G. Dal Maso, *An Introduction to Γ-Convergence*, Birkhäuser, Boston, 1993.

[17] I. Ekeland and R. Temam, *Analyse Convexe et Problèmes Variationnels*, Dunod, Gauthiers–Villars, Paris, 1974.

[18] I. Fonseca and G. A. Francfort, *Relaxation in BV versus quasiconvexification in $W^{1,p}$; a model for the interaction between fracture and damage*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 407–446.

[19] I. Fonseca, G. A. Francfort, and G. Leoni, *Thin elastic films: The impact of higher order perturbations*, Quart. Appl. Math., 65 (2007), pp. 69–98.

[20] I. Fonseca, S. Müller, and P. Pedregal, *Analysis of concentration and oscillation effects generated by gradients*, SIAM J. Math. Anal., 29 (1998), pp. 736–756.

[21] C. J. Larsen, *On the representation of effective energy densities*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 529–538.

[22] H. Le Dret and A. Raoult, *The nonlinear membrane model as variational limit of nonlinear three-dimensional elasticity*, J. Math. Pures Appl. (9), 74 (1995), pp. 549–578.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 1951–1977

LONG TAILS IN THE LONG-TIME ASYMPTOTICS OF QUASI-LINEAR HYPERBOLIC-PARABOLIC SYSTEMS OF CONSERVATION LAWS*

GUILLAUME VAN BAALEN†, NIKOLA POPOVIĆ†, AND C. EUGENE WAYNE†

Abstract. The long-time behavior of solutions of systems of conservation laws has been extensively studied. In particular, Liu and Zeng [*Mem. Amer. Math. Soc.*, 125 (1997), pp. viii–120] have given a detailed exposition of the leading order asymptotics of solutions close to a constant background state. In this paper, we extend the analysis of Liu and Zeng by examining higher order terms in the asymptotics in the framework of the so-called two-dimensional *p-system*, though we believe that our methods and results also apply to more general systems. We give a constructive procedure for obtaining these terms, and we show that their structure is determined by the interplay of the parabolic and hyperbolic parts of the problem. In particular, we prove that the corresponding solutions develop *long tails*.

Key words. long-time asymptotics, long tails, hyberbolic-parabolic conservation laws, p-system

AMS subject classifications. 35L65, 35B40, 35K55, 35M10, 35C20

DOI. 10.1137/070684835

1. Introduction.

In this paper, we consider the long-time behavior of solutions of systems of viscous conservation laws. This topic has been extensively studied. In particular, for the case of solutions close to a constant background state, [6] (building on work of [2]) contains a detailed exposition of the leading order long-time behavior of such solutions. More precisely, it is shown in [6] that the leading order asymptotics are given as a sum of contributions moving with the characteristic speeds of the undamped system of conservation laws and that each contribution evolves either as a Gaussian solution of the heat equation or as a self-similar solution of the viscous Burger's equation. Thus, with the exception of the translation along characteristics, these leading order terms reflect primarily the dissipative aspects of the problem.

In this paper, in an effort to better understand the interplay between the hyperbolic and parabolic aspects of the problem, we examine higher order terms in the asymptotics. We work with a specific two-dimensional system of equations–the *p-system*, but we believe that its behavior is prototypical. In particular, we think that our methods and results would extend to more complicated systems such as the "full gas dynamics" and the equations of magnetohydrodynamics (MHD) as considered in [6].

The specific set of equations we consider is the following:

$$\begin{aligned} &\partial_t a = c_1 \partial_x b, && a(x,0) = a_0(x), \\ &\partial_t b = c_2 \partial_x a + \partial_x g(a,b) + \alpha \left(\partial_x^2 b + \partial_x (f(a,b) \partial_x b) \right), && b(x,0) = b_0(x). \end{aligned} \tag{1.1}$$

We will make precise the assumptions on the nonlinear terms f and g below, but in order to describe our results informally, we basically assume that $|g(a,b)| \sim \mathcal{O}((|a| +$

*Received by the editors March 9, 2007; accepted for publication (in revised form) September 17, 2007; published electronically March 18, 2008.
http://www.siam.org/journals/sima/39-6/68483.html

†Department of Mathematics and Statistics and Center for Biodynamics, Boston University, Boston, MA 02215 (gvb@math.bu.edu, popovic@math.bu.edu, cew@math.bu.edu). The first author was supported by the *Fond National Suisse pour la Recherche Scientifique*. The second author was supported by NSF grant DMS-0109427. The third author was supported in part by NSF grant DMS-0405724.

$|b|)^2)$ and $|f(a,b)| \sim \mathcal{O}((|a|+|b|))$. We also note that, without loss of generality, we can set $c_1 = c_2 = 1$ and $\alpha = 2$ in (1.1), which can be achieved by appropriate scalings of space, time and the dependent variables, and possible redefinition of the functions f and g.

Physically, (1.1) is a model for compressible, constant entropy flow, where a represents the volume fraction (i.e., the reciprocal of the density) and b is the fluid velocity. The first of the two equations in (1.1) is the consistency relation between these two physical quantities. In particular, it would not be physically reasonable to include a dissipative term in this equation, whereas such a term arises naturally in the second equation which is essentially Newton's law, in which internal frictional forces are often present. As a consequence of the form of the dissipation the damping here is not "diagonalizable" in the terminology of [6].

Next, we note that, with the scaling $c_1 = c_2 = 1$ and $\alpha = 2$ in (1.1), the characteristic speeds are ± 1. If the initial conditions a_0 and b_0 in (1.1) decay sufficiently fast as $|x| \to \infty$, Liu and Zeng [6] showed that $a(x,t) \pm b(x,t) = \frac{1}{\sqrt{1+t}} g_0^{\pm}(\frac{x \pm t}{\sqrt{1+t}}) + \mathcal{O}((1+t)^{-\frac{3}{4}})$, where $g_0^{\pm}$ are self-similar solutions either of the heat equation or of Burger's equation, depending on the detailed form of the nonlinear terms. In this paper we derive similar expressions for the higher order terms in the asymptotics through a constructive procedure that can be carried out to arbitrary order.

More precisely, we show that, for any $N \geq 1$, there exist (universal) functions $\{g_n^{\pm}\}_{n=1}^N$ and constants $\{d_n^{\pm}\}_{n=1}^N$ determined by the initial conditions such that

$$\begin{aligned} a(x,t) + b(x,t) &= \frac{1}{\sqrt{1+t}} g_0^+ \left(\frac{x+t}{\sqrt{1+t}} \right) + \sum_{n=1}^N \frac{1}{(1+t)^{\delta_n}} d_n^+ g_n^+ \left(\frac{x+t}{\sqrt{1+t}} \right) \\ &\quad + \mathcal{O}\left(\frac{1}{(1+t)^{\delta_{N+1}}} \right), \\ a(x,t) - b(x,t) &= \frac{1}{\sqrt{1+t}} g_0^- \left(\frac{x-t}{\sqrt{1+t}} \right) + \sum_{n=1}^N \frac{1}{(1+t)^{\delta_n}} d_n^- g_n^- \left(\frac{x-t}{\sqrt{1+t}} \right) \\ &\quad + \mathcal{O}\left(\frac{1}{(1+t)^{\delta_{N+1}}} \right), \end{aligned} \tag{1.2}$$

where

$$\delta_n = 1 - \frac{1}{2^{n+1}}.$$

We give explicit expressions for the functions $g_n^{\pm}$ below, but focusing for the moment on the case $N = 1$ and the variable a, we have

$$\begin{aligned} a(x,t) = &\frac{1}{2\sqrt{1+t}} \left(g_0^+ \left(\frac{x+t}{\sqrt{1+t}} \right) + g_0^- \left(\frac{x-t}{\sqrt{1+t}} \right) \right) \\ &+ \frac{1}{2(1+t)^{\frac{3}{4}}} \left(d_1^+ g_1^+ \left(\frac{x+t}{\sqrt{1+t}} \right) + d_1^- g_1^- \left(\frac{x-t}{\sqrt{1+t}} \right) \right) + \mathcal{O}\left(\frac{1}{(1+t)^{\frac{7}{8}}} \right), \end{aligned}$$

where the functions $g_0^{\pm}(z)$ and $g_1^{\pm}(z)$ are solutions of the following ordinary differential equations:

$$\partial_z^2 g_0^{\pm}(z) + \frac{1}{2} z \partial_z g_0^{\pm}(z) + \frac{1}{2} g_0^{\pm}(z) + c_{\pm} \partial_z (g_0^{\pm}(z)^2) = 0, \tag{1.3}$$

$$\partial_z^2 g_1^{\pm}(z) + \frac{1}{2} z \partial_z g_1^{\pm}(z) + \frac{3}{4} g_1^{\pm}(z) + 2c_{\pm} \partial_z (g_0^{\pm}(z) g_1^{\pm}(z)) = 0. \tag{1.4}$$

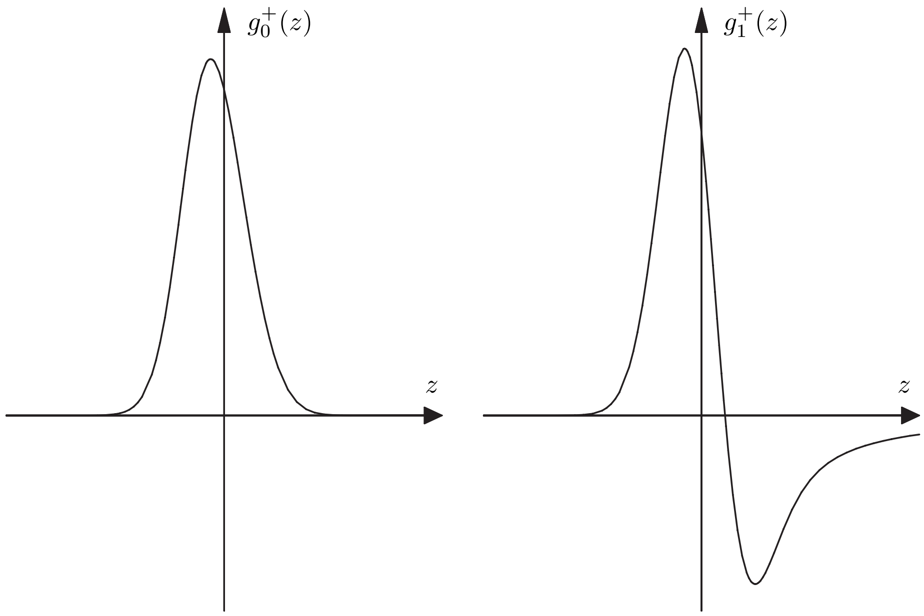


FIG. 1. *Graphs of the functions g_0^+ (left panel) and g_1^+ (right panel). Note the* long tail *of g_1^+ as $z \to \infty$.*

Here $c_\pm$ are constants that depend on the Hessian matrix of $g(a, b)$ at $a = b = 0$ and that will be specified in the course of our analysis. We will prove that, while all solutions of (1.3) have Gaussian decay as $|x| \to \infty$, general solutions of the *linear* equation (1.4) are linear combinations of two functions $g_{1,\pm}(z)$, where $g_{1,\pm}(z)$ decays like a Gaussian as $z \to \mp\infty$ but only like $|z|^{-\frac{3}{2}}$ as $z \to \pm\infty$; see also [5]. The graphs of the functions $g_0^+(z)$ and $g_1^+(z)$ are presented in Figure 1.

Thus, the higher order terms in the asymptotics develop *long tails*. These tails are a manifestation of the hyperbolic part of the problem (or perhaps more precisely of the interplay between the parabolic and hyperbolic parts). Were we to consider just the asymptotic behavior of the viscous Burger's equation which gives the leading order behavior of the solutions, we would find that, if the initial data are well-localized, the higher order terms in the long-time asymptotics decay rapidly in space and have temporal decay rates given by half-integers.

We also note one additional fact about the expansion in (1.2). Prior research [3, 9] has shown that for both parabolic equations and damped wave equations the eigenfunctions of the operator

$$\mathcal{L}u(z) = \partial_z^2 u + \frac{1}{2} z \partial_z u$$

play an important role for the asymptotics. In particular, on appropriate function spaces this operator has a sequence of isolated eigenvalues whose associated eigenfunctions can be used to construct an expansion for the long-time asymptotics. In this connection we prove that the functions $g_n^\pm$ are closely approximated by eigenfunctions of $\mathcal{L}$ with eigenvalues $\lambda_n = -\frac{1}{2} + 2^{-(n+1)}$; more precisely, the functions $g_n^\pm$ are eigenfunctions of a compact perturbation of $\mathcal{L}$; see, e.g., (1.4). However, so far we have not succeeded in finding a function space which both contains these eigenfunctions (the functions $g_n^\pm$ decay slowly as $z \to \pm\infty$) and in which the corresponding eigenvalues are isolated points in the spectrum. We plan to investigate this point further in future research.

Before moving to a precise statement of our results, we note that our approach makes no use of Kawashima's energy estimates for hyperbolic-parabolic conservation laws [4]. Instead we prove existence by directly studying the integral form of (1.1).

We now state our results on the Cauchy problem (1.1). We begin by stating the precise assumptions we make on the nonlinearities f and g in (1.1).

DEFINITION 1. *The maps* $f, g : \mathbf{R}^2 \to \mathbf{R}$ *are admissible nonlinearities for* (1.1) *if there is a quadratic map* $g_0 : \mathbf{R}^2 \to \mathbf{R}$ *and a constant* C *such that, for all* $|\mathbf{z}|$, $|\mathbf{z}_1|$, *and* $|\mathbf{z}_2|$ *small enough,*

$$
\begin{aligned}
|g(\mathbf{z})| &\le C|\mathbf{z}|^2, & |g(\mathbf{z}_1) - g(\mathbf{z}_2)| &\le C|\mathbf{z}_1 - \mathbf{z}_2|(|\mathbf{z}_1| + |\mathbf{z}_2|),\\
|\Delta g(\mathbf{z})| &\le C|\mathbf{z}|^3, & |\Delta g(\mathbf{z}_1) - \Delta g(\mathbf{z}_2)| &\le C|\mathbf{z}_1 - \mathbf{z}_2|(|\mathbf{z}_1| + |\mathbf{z}_2|)^2,\\
|f(\mathbf{z})| &\le C|\mathbf{z}|, \quad \textit{and} & |f(\mathbf{z}_1) - f(\mathbf{z}_2)| &\le C|\mathbf{z}_1 - \mathbf{z}_2|,
\end{aligned}
$$

where $\Delta g(\mathbf{z}) \equiv g(\mathbf{z}) - g_0(\mathbf{z})$.

The main result of this paper can be formulated as follows.

THEOREM 2. *Fix* $N > 0$. *There exists* $\epsilon_0 > 0$ *sufficiently small such that if*

(i) $|a_0|_{\mathrm{H}^1(\mathbb{R})} + |a_0|_{\mathrm{L}^1(\mathbb{R})} < \epsilon_0$ *and* $|b_0|_{\mathrm{H}^2(\mathbb{R})} + |b_0|_{\mathrm{L}^1(\mathbb{R})} < \epsilon_0$,

(ii) $|x^2 a_0|_{\mathrm{L}^2(\mathbb{R})} + |x^2 b_0|_{\mathrm{L}^2(\mathbb{R})} < \infty$,

then (1.1) *has a unique (mild) solution with initial conditions* a_0 *and* b_0. *Moreover, there exist functions* $\{g_n^\pm\}_{n=0}^N$ *(independent of initial conditions for* $n \ge 1$*) and constants* C_N, $\{d_n^\pm\}_{n=1}^N$ *determined by the initial conditions such that*

$$
\begin{aligned}
a(x,t) + b(x,t) &= \frac{1}{\sqrt{1+t}} g_0^+\left(\frac{x+t}{\sqrt{1+t}}\right) + \sum_{n=1}^N \frac{1}{(1+t)^{1-\frac{1}{2^{n+1}}}} d_n^+ g_n^+\left(\frac{x+t}{\sqrt{1+t}}\right) + R_u^N(x,t),\\
a(x,t) - b(x,t) &= \frac{1}{\sqrt{1+t}} g_0^-\left(\frac{x-t}{\sqrt{1+t}}\right) + \sum_{n=1}^N \frac{1}{(1+t)^{1-\frac{1}{2^{n+1}}}} d_n^- g_n^-\left(\frac{x-t}{\sqrt{1+t}}\right) + R_v^N(x,t),
\end{aligned}
\tag{1.5}
$$

where the remainders R_u^N *and* R_v^N *satisfy the estimates*

$$
\begin{aligned}
\sup_{t\ge0}(1+t)^{\frac34 - \frac{1}{2^{N+2}}} \|R^N_{\{u,v\}}(\cdot,t)\|_{\mathrm{L}^2(\mathbb{R})} &\le C_N,\\
\sup_{t\ge0}(1+t)^{\frac54 - \frac{1}{2^{N+2}}} \|\partial_x R^N_{\{u,v\}}(\cdot,t)\|_{\mathrm{L}^2(\mathbb{R})} &\le C_N.
\end{aligned}
\tag{1.6}
$$

Furthermore, for $n \ge 1$, *the functions* $g_n^\pm$ *satisfy* $g_n^\pm(z) \sim |z|^{-2+\frac{1}{2^n}}$ *as* $z \to \pm\infty$.

There is a slight incongruity in this result in that the norm in which we estimate the remainder term is weaker than the one we use on the initial data; namely, we do not give estimates for the remainder in $\mathrm{H}^2(\mathbb{R})$ or in the localization norms $\mathrm{L}^1(\mathbb{R})$ and the weighted $\mathrm{L}^2(\mathbb{R})$-norm (on that aspect of the problem, see Remark 3 below). Theorem 2 actually holds for slightly more general initial conditions than those satisfying (i)–(ii). Furthermore, we will prove that the estimates (1.6) hold for all initial conditions (a_0, b_0) in a subset $\mathcal{D}_2 \subset \mathrm{H}_1 \times \mathrm{H}_2$ that is *positively invariant* under the flow of (1.1). However, since the topology used to define the subset $\mathcal{D}_2$ is somewhat nonstandard, we have chosen to state the result initially in this slightly weaker, but hopefully more comprehensible, form to keep the introduction as simple as possible.

Remark 3. It is interesting to note (see Proposition 7 below) that $\|x^2 a(\cdot,t)\|_{\mathrm{L}^2(\mathbb{R})} + \|x^2 b(\cdot,t)\|_{\mathrm{L}^2(\mathbb{R})}$ is finite for all finite $t > 0$ but that the terms with $n \ge 1$ in the asymptotic expansion do not satisfy this property due to the long tails of the functions $g_n^{\pm}$.

Remark 4. As the asymmetry in the degree of x derivatives in (1.1) suggests, we require more spatial regularity from the second component (the b variable) than from the first (the a variable). It is then natural to expect that R_u^N or R_v^N are not necessarily in H^2 but that only their difference is.

We conclude this section with a few remarks. Define $u_{\pm}(x,t) = a(x,t) \pm b(x,t)$. Then the asymptotics of the solutions of (1.1) in the variables $u_{\pm}$ are the same as those of the two-dimensional (generalized) Burger's equation

$$\begin{aligned} \partial_t u_+ &= \partial_x^2 u_+ + \partial_x u_+ + \partial_x (c_+ u_+^2 - c_- u_-^2), \\ \partial_t u_- &= \partial_x^2 u_- - \partial_x u_- + \partial_x (c_- u_-^2 - c_+ u_+^2), \end{aligned} \tag{1.7}$$

where the constants $c_{\pm}$ are determined by the Hessian of $g(a,b)$ at $a = b = 0$ through

$$c_{\pm} = \pm \frac{1}{8} (1, \pm 1) \cdot \begin{pmatrix} \partial_a^2 g & \partial_a \partial_b g \\ \partial_a \partial_b g & \partial_b^2 g \end{pmatrix}\bigg|_{a=b=0} \cdot \begin{pmatrix} 1 \\ \pm 1 \end{pmatrix}.$$

We will show that the hyperbolic effects manifest themselves through the "source" terms $-c_- u_-^2$ (respectively, $c_+ u_+^2$) in the first (respectively, second) equation in (1.7). In particular, none of the terms $g_n^{\pm}$, with $n \geq 1$, would be present in the asymptotic expansion if those terms were absent.

Finally, note that we have chosen to state Theorem 2 for finite N. As it turns out, the sums appearing in (1.5) converge in the limit as $N \to \infty$, in which case the estimates (1.6) hold with time weights replaced by $(1+t)^{\frac{3}{4}} \ln(2+t)^{-1}$ and $(1+t)^{\frac{5}{4}} \ln(2+t)^{-1}$. The proof can easily be done with the techniques used in this paper and is left to the reader.

The remainder of the paper is organized as follows: In section 2, we discuss the well-posedness of the Cauchy problem (1.1) in an appropriately defined topology. In section 3, we explain our strategy for proving our main result (Theorem 2) on the long-time asymptotics of solutions of (1.1). Namely, we decompose that proof into a series of simpler subproblems which are then tackled in subsequent sections: In sections 4 and 5, we investigate properties of solutions of Burger-type equations (respectively, of inhomogeneous heat equations) as they occur naturally in the asymptotic analysis. In section 6, we collect some estimates that are used in the proof of the well-posedness of (1.1). Finally, in section 7, we specify the sense in which the semigroup of the linearization of (1.1) is close to heat kernels translating along the characteristics, and we give estimates on the remainder terms occurring in Theorem 2.

2. Cauchy problem. To motivate our technical treatment of the problem and in particular our choice of function spaces, we first note that, upon taking the Fourier transform of the linearization of (1.1), it follows that

$$\partial_t \begin{pmatrix} a \\ b \end{pmatrix} = \mathrm{L} \begin{pmatrix} a \\ b \end{pmatrix} \equiv \begin{pmatrix} 0 & ik \\ ik & -2k^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}. \tag{2.1}$$

We then find that the (Fourier transform of) the semigroup associated with (2.1) is

$$\mathrm{e}^{\mathrm{L}t} = \mathrm{e}^{-k^2 t} \begin{pmatrix} \cos(kt\Delta) + \frac{k}{\Delta} \sin(kt\Delta) & \frac{i}{\Delta} \sin(kt\Delta) \\ \frac{i}{\Delta} \sin(kt\Delta) & \cos(kt\Delta) - \frac{k}{\Delta} \sin(kt\Delta) \end{pmatrix}, \tag{2.2}$$

where $\Delta = \sqrt{1-k^2}$. The most important fact about the semigroup $\mathrm{e}^{\mathrm{L}t}$ is that it is close to $\mathrm{e}^{\mathrm{L}_0 t}$, the semigroup associated with the problem

$$\partial_t \begin{pmatrix} u \\ v \end{pmatrix} = \mathrm{L}_0 \begin{pmatrix} u \\ v \end{pmatrix} \equiv \begin{pmatrix} \partial_x^2 + \partial_x & 0 \\ 0 & \partial_x^2 - \partial_x \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}. \tag{2.3}$$

Formally, $\mathrm{e}^{\mathrm{L}_0 t}$ can be obtained by setting $\Delta = 1$ in $\mathrm{e}^{\mathrm{L}t}$ and by conjugating with the matrix

$$\mathcal{S} \equiv \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{2.4}$$

These two operations correspond to a long wavelength expansion and a change of dependent variables to quantities that move along the characteristics. More precisely, we will prove that $\mathrm{e}^{\mathrm{L}t}$ satisfies the intertwining property

$$\mathcal{S}\mathrm{e}^{\mathrm{L}t} \approx \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S},$$

where the symbol $\approx$ means that the action of these two operators is the same in the large-scale—long-time limit; see Lemma 19 at the beginning of section 7 for details.

Furthermore, $\mathrm{e}^{\mathrm{L}t}$ satisfies paraboliclike estimates

$$|\mathrm{e}^{\mathrm{L}t}| \le C\mathrm{e}^{-\min(k^2,1)\frac{t}{4}} \begin{pmatrix} 1 & \frac{1}{\sqrt{1+k^2}} \\ \frac{1}{\sqrt{1+k^2}} & 1 \end{pmatrix}, \tag{2.5}$$

$$\left| \mathrm{e}^{\mathrm{L}t} \begin{pmatrix} 0 \\ ik \end{pmatrix} \right| \le C \frac{\mathrm{e}^{-\min(k^2,1)\frac{t}{4}}}{\sqrt{t}} \begin{pmatrix} 1 \\ \frac{1}{\sqrt{1+k^2}} \end{pmatrix} \tag{2.6}$$

uniformly in $t \ge 0$ and $k \in \mathbf{R}$.

Hence, to summarize, $\mathrm{e}^{\mathrm{L}t}$ behaves like a superposition of heat kernels translating along the characteristics of the underlying hyperbolic problem. In view of the above observations as well as of classical techniques for parabolic PDEs (see, e.g., [7, 1]), we will consider (1.1) in the following (somewhat nonstandard) topology (cf. also [8]).

Definition 5. *We define $\mathcal{B}_0$ (respectively, $\mathcal{B}$) as the closure of $\mathcal{C}_0^\infty(\mathbf{R}, \mathbf{R}^2)$ (respectively, $\mathcal{C}_0^\infty(\mathbf{R} \times [0,\infty), \mathbf{R}^2)$) under the norm $|\cdot|$ (respectively, $\|\cdot\|$), where, for $\mathbf{z}_0 = (a_0, b_0) : \mathbf{R} \to \mathbf{R}^2$ and $\mathbf{z} = (a,b) : \mathbf{R} \times [0,\infty) \to \mathbf{R}^2$, we define*

$$|\mathbf{z}_0| = \|\widehat{\mathbf{z}}_0\|_\infty + \|\mathbf{z}_0\|_2 + \|\mathrm{D}\mathbf{z}_0\|_2 + \|\mathrm{D}^2 b_0\|_2,$$
$$\|\mathbf{z}\| = \|\hat{\mathbf{z}}\|_{\infty,0} + \|\mathbf{z}\|_{2,\frac{1}{4}} + \|\mathrm{D}\mathbf{z}\|_{2,\frac{3}{4}} + \|\mathrm{D}^2 b\|_{2,\frac{5}{4}\star}.$$

Here $(Da)(x,t) \equiv \partial_x a(x,t)$, $\hat{a}(k,t)$ is the Fourier transform of $a(x,t)$,

$$\|f\|_{p,q} = \sup_{t\ge 0}(1+t)^q \|f(\cdot,t)\|_p, \qquad \|f\|_{p,q^\star} = \sup_{t\ge 0} \frac{(1+t)^q}{\ln(2+t)} \|f(\cdot,t)\|_p,$$

and $\|\cdot\|_p$ is the standard $\mathrm{L}^p(\mathbf{R})$-norm.

Before turning to the Cauchy problem with initial data in $\mathcal{B}_0$, we collect a few comments on our choice of function spaces.

Consider first the requirements on the initial conditions in (1.1). While the use of H^1 space is quite natural in this context, we choose to replace the L^1-norm by the (weaker) control of the L^∞-norm in Fourier space. This has the great advantage that all estimates can then be done in Fourier space, where the semigroup $\mathrm{e}^{\mathrm{L}t}$ has the simple, explicit, form (2.2).

In turn, our choice of q-exponents in the norm $\|\cdot\|$ is motivated by the fact that these are the highest possible exponents for which the $\|\cdot\|$-norm of the leading order asymptotic term $\frac{1}{\sqrt{1+t}}g_0^{\pm}(\frac{x\pm t}{\sqrt{1+t}})$ is bounded. Note also that, for the linear evolution (2.1), we have

$$\|\mathrm{e}^{\mathrm{L}t}\mathbf{z}_0\| \leq C|\mathbf{z}_0|, \tag{2.7}$$

since $\hat{j}(k,t) = \mathrm{e}^{-\min(k^2,1)t}u_0(k)$ satisfies

$$\|\mathrm{D}^n j(\cdot,t)\|_2 \leq C\Big(\mathrm{e}^{-t}\|\mathrm{D}^n u_0\|_2 + \min\Big(t^{-\frac{1}{4}-\frac{n}{2}}\|\hat{u}_0\|_\infty, \|D^n u_0\|_2\Big)\Big)$$

for all $n = 0, 1, \ldots$.

Finally, we note that, for admissible nonlinearities in the sense of Definition 1, the map $h(a,b) = f(a,b)\partial_x b + g(a,b) = h(\mathbf{z})$ satisfies

$$\|h(\mathbf{z})\|_{1,\frac{1}{2}} + \|h(\mathbf{z})\|_{2,\frac{3}{4}} + \|Dh(\mathbf{z})\|_{2,\frac{5}{4}} \leq C\|\mathbf{z}\|^2, \tag{2.8}$$

$$\|h(\mathbf{z}_1) - h(\mathbf{z}_2)\|_{1,\frac{1}{2}} + \|h(\mathbf{z}_1) - h(\mathbf{z}_2)\|_{2,\frac{3}{4}} \leq C\|\mathbf{z}_1 - \mathbf{z}_2\|(\|\mathbf{z}_1\| + \|\mathbf{z}_2\|), \tag{2.9}$$

$$\|D(h(\mathbf{z}_1) - h(\mathbf{z}_2))\|_{2,\frac{5}{4}} \leq C\|\mathbf{z}_1 - \mathbf{z}_2\|(\|\mathbf{z}_1\| + \|\mathbf{z}_2\|). \tag{2.10}$$

We are now fully equipped to study the Cauchy problem (1.1) in $\mathcal{B}$.

THEOREM 6. *For all $\mathbf{z}_0 \in \mathcal{B}_0$ with $|\mathbf{z}_0| = |(a_0, b_0)| \leq \epsilon_0$ small enough, the Cauchy problem* (1.1) *is (locally) well posed in $\mathcal{B}$ if the nonlinearities are admissible in the sense of Definition* 1*. In particular, the solution satisfies $\|\mathbf{z}\| \leq c\epsilon_0$ for some $c > 1$ and is unique among functions in $\mathcal{B}$ satisfying this bound.*

Proof. Upon taking the Fourier transform of (1.1), we get

$$\partial_t\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0 & ik \\ ik & -2k^2 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} 0 \\ ikh \end{pmatrix}, \tag{2.11}$$

which gives the following representation for the solution:

$$\mathbf{z}(t) \equiv \begin{pmatrix} a(t) \\ b(t) \end{pmatrix} = \mathrm{e}^{\mathrm{L}t}\begin{pmatrix} a_0 \\ b_0 \end{pmatrix} + \int_0^t \mathrm{d}s\ \mathrm{e}^{\mathrm{L}(t-s)}\begin{pmatrix} 0 \\ \partial_x h(\mathbf{z}(s)) \end{pmatrix} \equiv \mathrm{e}^{\mathrm{L}t}\mathbf{z}_0 + \mathcal{N}[\mathbf{z}](t). \tag{2.12}$$

We will prove below that for all $\mathbf{z}_i \in \mathcal{B}$, $i = 1, 2$, we have

$$\|\mathcal{N}[\mathbf{z}]\| \leq C\|\mathbf{z}\|^2 \quad \text{and} \quad \|\mathcal{N}[\mathbf{z_1}] - \mathcal{N}[\mathbf{z_2}]\| \leq C\|\mathbf{z}_1 - \mathbf{z}_2\|(\|\mathbf{z}_1\| + \|\mathbf{z}_2\|) \tag{2.13}$$

for some constant C. The proof of Theorem 6 then follows from the fact that, for all $\mathbf{z}_0 \in \mathcal{B}_0$ with $|\mathbf{z}_0| \leq \epsilon_0$ small enough and $c > 1$, the right-hand side (r.h.s.) of (2.12) defines a contraction map from some (small) ball of radius $c\epsilon_0$ in $\mathcal{B}$ onto itself.

The general rule for proving the various estimates involved in (2.13) is to split the integration interval into two parts, with $s \in \mathcal{I}_1 \equiv [0, \frac{t}{2}]$ and $s \in \mathcal{I}_2 \equiv [\frac{t}{2}, t]$. In $\mathcal{I}_1$, we place as many derivatives (or, equivalently, factors of k) as possible on the semigroup $\mathrm{e}^{\mathrm{L}(t-s)}$, while on $\mathcal{I}_2$, (most of) these derivatives need to act on h, since the integral would otherwise be divergent at $s = t$.

Additional difficulties arise from the fact that $\mathrm{e}^{\mathrm{L}t}$ has very few smoothing properties (slow or no decay in k as $|k| \to \infty$), so that in some cases we need to consider separately the large-k part and the small-k part of the L^2-norm, say. This is done through the use of $\mathbb{P}$, defined as the Fourier multiplier with the characteristic function on $[-1, 1]$.

We decompose the proof of $\|\mathcal{N}[\mathbf{z}]\| \leq C\|\mathbf{z}\|^2$ into that of

$$\begin{aligned}\|\mathcal{N}[\mathbf{z}]\| \leq{}& \|\widehat{\mathcal{N}[\mathbf{z}]}\|_{\infty,0} + \|\mathcal{N}[\mathbf{z}]\|_{2,\frac{1}{4}} + \|\mathbb{P}\mathrm{D}\mathcal{N}[\mathbf{z}]\|_{2,\frac{3}{4}} + \|(1-\mathbb{P})\mathrm{D}\mathcal{N}[\mathbf{z}]\|_{2,\frac{3}{4}}\\ &+ \|(1-\mathbb{P})\mathrm{D}^2\mathcal{N}[\mathbf{z}]_2\|_{2,\frac{5}{4}^\star} + \|(1-\mathbb{Q})\mathbb{P}\mathrm{D}^2\mathcal{N}[\mathbf{z}]_2\|_{2,\frac{5}{4}^\star} + \|\mathbb{Q}\mathbb{P}\mathrm{D}^2\mathcal{N}[\mathbf{z}]_2\|_{2,\frac{5}{4}^\star}\\ \leq{}& C\|\mathbf{z}\|^2, \end{aligned} \tag{2.14}$$

where $\mathbb{Q}$ is the characteristic function for $t \geq 1$ and $\mathcal{N}[\mathbf{z}]_2$ denotes the second component of $\mathcal{N}[\mathbf{z}]$.

We now consider $\|\mathbb{P}\mathrm{D}\mathcal{N}[\mathbf{z}]\|_{2,\frac{3}{4}}$ as an example of the way we prove the above estimates. We have

$$\begin{aligned}\|\mathbb{P}\mathrm{D}\mathcal{N}[\mathbf{z}](\cdot,t)\|_2 \leq{}& \|h(\mathbf{z})\|_{2,\frac{3}{4}} \left(\sup_{|k|\leq 1,\tau\geq 0} |k|\sqrt{\tau}\mathrm{e}^{-\frac{k^2\tau}{4}}\right)\int_0^{\frac{t}{2}} \mathrm{d}s\ \frac{(1+s)^{-\frac{3}{4}}}{t-s}\\ &+ \|\mathrm{D}h(\mathbf{z})\|_{2,\frac{5}{4}}\left(\sup_{|k|\leq 1,\tau\geq 0} \mathrm{e}^{-\frac{k^2\tau}{4}}\right)\int_{\frac{t}{2}}^t \mathrm{d}s\ \frac{(1+s)^{-\frac{5}{4}}}{\sqrt{t-s}}\\ \leq{}& C\|\mathbf{z}\|^2\left(\frac{2}{t}\int_0^{\frac{t}{2}}\frac{\mathrm{d}s}{(1+s)^{\frac{3}{4}}} + \frac{1}{(1+\frac{t}{2})^{\frac{5}{4}}}\int_{\frac{t}{2}}^t\frac{\mathrm{d}s}{\sqrt{t-s}}\right)\\ \leq{}& C\|\mathbf{z}\|^2(1+t)^{-\frac{3}{4}}\end{aligned} \tag{2.15}$$

for all $t \geq 0$, which shows that $\|\mathbb{P}\mathrm{D}\mathcal{N}[\mathbf{z}]\|_{2,\frac{3}{4}} \leq C\|\mathbf{z}\|^2$. All other estimates in (2.14) can be done similarly; we postpone their proof to section 6 below.

Finally, we note that the Lipschitz-type estimate in (2.13) can be obtained in the same manner, mutatis mutandis, due to the similarity between (2.9) and (2.10) with (2.8); we omit the details. ∎

We can now turn to the question of the asymptotic structure of the solutions of (1.1) provided by Theorem 6. Note that already if we wanted to prove that $\mathrm{e}^{\mathrm{L}t}\mathbf{z}_0$ satisfies "Gaussian asymptotics" we would need more localization properties on $\mathbf{z}_0$ than those provided by the $\mathcal{B}_0$-topology. It will turn out to be sufficient to require $\mathbf{z}_0 \in \mathcal{B}_0 \cap \mathrm{L}^2(\mathbb{R}, x^m\mathrm{d}x)$ for (some) $m \geq 2$. We now prove that this requirement is *forward invariant* under the flow of (1.1).

PROPOSITION 7. *Let* $\rho_m(x) = |x|^m$, *and define*

$$\mathcal{D}_m = \left\{ \mathbf{z}_0 \in \mathcal{B}_0 \textit{ such that } |\mathbf{z}_0| + \|\rho_m\mathbf{z}_0\|_2 < \infty \right\}.$$

If $\mathbf{z}_0 \in \mathcal{D}_m$ *and* $|\mathbf{z}_0| \leq \epsilon_0$ *such that Theorem* 6 *holds, then the corresponding solution* $\mathbf{z}(t)$ *of* (1.1) *satisfies* $\mathbf{z}(t) \in \mathcal{D}_m$ *for all finite* $t > 0$. *Furthermore, there holds* $|\mathbf{z}(t)| \leq (1+\delta)\epsilon_0$ *for some (small) constant* δ.

Proof. Note first that, by Theorem 6, $|\mathbf{z}(t)| \leq \|\mathbf{z}\| \leq (1+\delta)\epsilon_0$ since $\mathbf{z}_0 \in \mathcal{B}_0$ and $|\mathbf{z}_0| \leq \epsilon_0$. Then fix $m \in \mathbf{N}$, $m \geq 1$. The proof of Theorem 6 can easily be adapted to show that (1.1) is *locally* (in time) well posed in $\mathcal{D}_m$. Global existence then follows from the fact that the quantity

$$N(t) = \frac{1}{2}\|\rho_m\mathbf{z}(\cdot,t)\|^2 = \frac{1}{2}\int_{-\infty}^{\infty}\mathrm{d}x\ |x|^m(a(x,t)^2 + b(x,t)^2)$$

grows *at most exponentially* as $t \to \infty$. Namely, we have

$$\begin{aligned}
\partial_t N(t) &= \int_{-\infty}^{\infty} \mathrm{d}x\ |x|^m \Big(\partial_x(ab) + 2b\partial_x^2 b + b\partial_x\Big(f(a,b)\partial_x b + g(a,b)\Big)\Big) \\
&= -\int_{-\infty}^{\infty} \mathrm{d}x\ m|x|^{m-1}\mathrm{sign}(x)\Big(b(a+g(a,b)) + (2+f(a,b))b\partial_x b\Big) \\
&\quad - \int_{-\infty}^{\infty} \mathrm{d}x\ |x|^m(\partial_x b)^2\Big(2+f(a,b)\Big) \\
&\le \int_{-\infty}^{\infty} \mathrm{d}x\ \Big((m-1)^{m-1} + |x|^m\Big)\Big|\ b(a+g(a,b)) + (2+f(a,b))b\partial_x b\ \Big| \\
&\quad - \int_{-\infty}^{\infty} \mathrm{d}x\ |x|^m(\partial_x b)^2\Big(2+f(a,b)\Big) \\
&\le \int_{-\infty}^{\infty} \mathrm{d}x\ \Big((m-1)^{m-1} + |x|^m\Big)\Big(|b(a+g(a,b))| + 2^{-1}|2+f(a,b)|b^2\Big) \\
&\le C_1(m,\epsilon_0) + C_2(\epsilon_0)N(t),
\end{aligned}$$

due to the estimates $\|f(a,b)\|_\infty \le C\epsilon_0 \ll 2$ and $\|\frac{g(a,b)}{\sqrt{a^2+b^2}}\|_\infty \le C\epsilon_0$. $\square$

3. Asymptotic structure—proof of Theorem 2. We can now state our main result on the asymptotic structure of solutions of (1.1) in a definitive manner.

THEOREM 8. *Let $\mathcal{D}_m$ be as in Proposition 7 with $m \ge 2$, let $\mathbf{z}_0 \in \mathcal{D}_m$ with $|\mathbf{z}_0| \le \epsilon_0$ such that Theorem 6 holds, and write $\mathbf{z}(t) = (a(t), b(t))$ for the corresponding solution of* (1.1)*. Then there exist functions $\{g_n^\pm\}_{n=0}^N$ (independent of $\mathbf{z}_0$ for $n \ge 1$) and constants C_N, $\{d_n^\pm\}_{n=1}^N$ determined by $\mathbf{z}_0$ such that*

$$\begin{aligned}
a(x,t) + b(x,t) &= \frac{1}{\sqrt{1+t}} g_0^+\left(\frac{x+t}{\sqrt{1+t}}\right) + \sum_{n=1}^N \frac{1}{(1+t)^{1-\frac{1}{2^{n+1}}}} d_n^+ g_n^+\left(\frac{x+t}{\sqrt{1+t}}\right) + R_u^N(x,t), \\
a(x,t) - b(x,t) &= \frac{1}{\sqrt{1+t}} g_0^-\left(\frac{x-t}{\sqrt{1+t}}\right) + \sum_{n=1}^N \frac{1}{(1+t)^{1-\frac{1}{2^{n+1}}}} d_n^- g_n^-\left(\frac{x-t}{\sqrt{1+t}}\right) + R_v^N(x,t),
\end{aligned} \tag{3.1}$$

where the remainders R_u^N and R_v^N satisfy the estimates

$$\begin{aligned}
\sup_{t\ge0}(1+t)^{\frac{3}{4}-\frac{1}{2^{N+2}}}\|R_{\{u,v\}}^N(\cdot,t)\|_{\mathrm{L}^2(\mathbb{R})} &\le C_N, \\
\sup_{t\ge0}(1+t)^{\frac{5}{4}-\frac{1}{2^{N+2}}}\|\partial_x R_{\{u,v\}}^N(\cdot,t)\|_{\mathrm{L}^2(\mathbb{R})} &\le C_N.
\end{aligned} \tag{3.2}$$

Furthermore, for $n \ge 1$, the functions $g_n^\pm$ satisfy $g_n^\pm(z) \sim |z|^{-2+\frac{1}{2^n}}$ as $z \to \pm\infty$.

Remark 9. As will be apparent from the proof of Theorem 8, any hyperbolic-parabolic system of the form

$$\partial_t \mathbf{z} + f(\mathbf{z})_x = (B(\mathbf{z})\mathbf{z}_x)_x$$

with admissible nonlinearities in the sense of (the natural extension of) Definition 1 gives rise to solutions having the same asymptotic structure as those of the p-system as long as the following two conditions are satisfied:

1. There exist two matrices $\mathcal{S}$ and A, with $\mathcal{S}$ nonsingular and A diagonal having eigenvalues of multiplicity 1 for which $\mathcal{S}\mathrm{e}^{\mathrm{L}t} \approx \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}$ in the sense of Lemma 19 (see section 7), where $\mathrm{L}_0 = \partial_x^2 + \mathrm{A}\partial_x$ and $\mathrm{L} = B(0)\partial_x^2 - f'(0)\partial_x$.

2. The Cauchy problem with the initial condition in the corresponding function space (the natural extension of $\mathcal{B}_0$ to the problem considered) is well-posed and satisfies the analogues of Theorem 6 and Proposition 7.

We now briefly comment on the above assumptions for specific systems such as the "full gas dynamics" and the MHD system. The intertwining property of item 1 above is proved in [6] for quite general systems, though not in exactly the same topology as that used in Lemma 19. As for item 2, local well posedness for initial data in $\mathcal{B}_0$ is certainly not an issue; the only difficulty is to prove that the various norms of Definition 5 exhibit "paraboliclike" decay as $t \to \infty$. This is very likely to hold, particularly for systems satisfying item 1.

While the variables (a, b) are adapted to the study of the Cauchy problem because of the inherent asymmetry of spatial regularity in (1.1), they are not the best framework for studying the asymptotic structure of the solutions to (1.1). It turns out to be more convenient to change variables to quantities that move along the characteristics. We thus define

$$\begin{pmatrix} u(x,t) \\ v(x,t) \end{pmatrix} \equiv \begin{pmatrix} \mathcal{T}^{-1} & 0 \\ 0 & \mathcal{T} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} a(x,t) \\ b(x,t) \end{pmatrix} \equiv \begin{pmatrix} \mathcal{T}^{-1} & 0 \\ 0 & \mathcal{T} \end{pmatrix} \mathcal{S}\mathbf{z}(x,t),$$

where $\mathcal{T}$ is the translation operator defined by

$$(\mathcal{T}f)(x,t) = f(x+t,t) \quad \text{or, equivalently, by} \quad \widehat{\mathcal{T}f}(k,t) = \mathrm{e}^{ikt}\hat{f}(k,t). \tag{3.3}$$

Note in passing that

$$a(x,t) = \frac{1}{2}\Big(u(x+t,t) + v(x-t,t)\Big) \quad \text{and} \quad b(x,t) = \frac{1}{2}\Big(u(x+t,t) - v(x-t,t)\Big).$$

We then use the fact that $\mathbf{z}$ satisfies the integral equation

$$\begin{aligned} \mathcal{S}\mathbf{z}(t) &= \mathcal{S}\mathrm{e}^{\mathrm{L}t}\mathbf{z}_0 + \int_0^t \mathrm{d}s\; \mathcal{S}\mathrm{e}^{\mathrm{L}(t-s)} \begin{pmatrix} 0 \\ \partial_x h(\mathbf{z}(s)) \end{pmatrix} \\ &= \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}\mathbf{z}_0 + \int_0^t \mathrm{d}s\; \mathrm{e}^{\mathrm{L}_0(t-s)}\mathcal{S} \begin{pmatrix} 0 \\ \partial_x g_0(\mathbf{z}(s)) \end{pmatrix} + \mathcal{R}[\mathbf{z}](t), \end{aligned} \tag{3.4}$$

where

$$\begin{aligned} \mathcal{R}[\mathbf{z}](t) =& \Big(\mathcal{S}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}\Big)\mathbf{z}_0 \\ &+ \int_0^t \mathrm{d}s\; \left[\mathcal{S}\mathrm{e}^{\mathrm{L}(t-s)} \begin{pmatrix} 0 \\ \partial_x h(\mathbf{z}(s)) \end{pmatrix} - \mathrm{e}^{\mathrm{L}_0(t-s)}\mathcal{S} \begin{pmatrix} 0 \\ \partial_x g_0(\mathbf{z}(s)) \end{pmatrix} \right]. \end{aligned}$$

To justify the notation, which suggests that $\mathcal{R}[\mathbf{z}] = (\mathcal{R}_u[\mathbf{z}], \mathcal{R}_v[\mathbf{z}])$ is a remainder term, we will prove in section 7 that it satisfies the improved decay rates

$$\|\mathcal{R}_{\{u,v\}}[\mathbf{z}]\|_{2,\frac{3}{4}^\star} + \|\mathrm{D}\mathcal{R}_{\{u,v\}}[\mathbf{z}]\|_{2,\frac{5}{4}^\star} \le C\epsilon_0, \tag{3.5}$$

because of the intertwining relation $\mathcal{S}\mathrm{e}^{\mathrm{L}t} \approx \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}$ (see Lemma 19) and the fact that $h(\mathbf{z}) = g_0(\mathbf{z}) + \text{h.o.t.}$

Recalling that g_0 is quadratic (cf. Definition 1), we will write

$$\begin{aligned} g_0(\mathbf{z}) &= c_+(a+b)^2 - c_-(a-b)^2 + c_3(a+b)(a-b) \\ &= c_+(\mathcal{T}u)^2 - c_-(\mathcal{T}^{-1}v)^2 + c_3(\mathcal{T}u)(\mathcal{T}^{-1}v) \end{aligned}$$

for $\mathbf{z}=(a,b)$. We thus find from (3.4) that u and v satisfy

$$\begin{aligned} u(t) &= \mathrm{e}^{\partial_x^2 t}(a_0+b_0) + \partial_x \int_0^t \mathrm{d}s\ \mathrm{e}^{\partial_x^2(t-s)}\Big(c_+u(s)^2 - c_-\mathcal{T}^{-2}v(s)^2\Big) \\ &\qquad + \mathcal{T}^{-1}\mathcal{R}_u[\mathbf{z}](t) + c_3\partial_x \int_0^t \mathrm{d}s\ \mathrm{e}^{\partial_x^2(t-s)}\mathcal{T}^{-1}\Big((\mathcal{T}u(s))(\mathcal{T}^{-1}v(s))\Big), \end{aligned} \tag{3.6}$$

$$\begin{aligned} v(t) &= \mathrm{e}^{\partial_x^2 t}(a_0-b_0) + \partial_x \int_0^t \mathrm{d}s\ \mathrm{e}^{\partial_x^2(t-s)}\Big(c_-v(s)^2 - c_+\mathcal{T}^{2}u(s)^2\Big) \\ &\qquad + \mathcal{T}\mathcal{R}_v[\mathbf{z}](t) - c_3\partial_x \int_0^t \mathrm{d}s\ \mathrm{e}^{\partial_x^2(t-s)}\mathcal{T}\Big((\mathcal{T}u(s))(\mathcal{T}^{-1}v(s))\Big). \end{aligned} \tag{3.7}$$

Note that, but for the presence of the second lines in (3.6) and (3.7), these expressions are precisely Duhamel's formula for the solution of the model problem (1.7), written in terms of $u=\mathcal{T}^{-1}u_+$ and $v=\mathcal{T}u_-$. The next step is to write

$$u = u_\star + R_u^N = u_0 + u_1 + R_u^N \quad \text{and} \quad v = v_\star + R_v^N = v_0 + v_1 + R_v^N,$$

considering R_u^N and R_v^N as new "unknowns" and

$$\begin{aligned} u_0(x,t) &= \frac{1}{\sqrt{1+t}} g_0^+\left(\frac{x}{\sqrt{1+t}}\right), & u_1(x,t) &= \sum_{n=1}^N \frac{1}{(1+t)^{1-\frac{1}{2^{n+1}}}} d_n^+ g_n^+\left(\frac{x}{\sqrt{1+t}}\right), \\ v_0(x,t) &= \frac{1}{\sqrt{1+t}} g_0^-\left(\frac{x}{\sqrt{1+t}}\right), \quad \text{and} & v_1(x,t) &= \sum_{n=1}^N \frac{1}{(1+t)^{1-\frac{1}{2^{n+1}}}} d_n^- g_n^-\left(\frac{x}{\sqrt{1+t}}\right) \end{aligned} \tag{3.8}$$

for some coefficients $\{d_n^\pm\}_{n=1}^N$ and functions $\{g_n^\pm\}_{n=0}^N$ to be determined later.

We now use

$$\begin{aligned} u^2 &= (u-u_\star)(u+u_\star) + u_\star^2 = R_u^N(u+u_\star) + u_1^2 + 2u_0u_1 + u_0^2, \\ v^2 &= (v-v_\star)(v+v_\star) + v_\star^2 = R_v^N(v+v_\star) + v_1^2 + 2v_0v_1 + v_0^2, \\ (\mathcal{T}u)(\mathcal{T}^{-1}v) &= (\mathcal{T}R_u^N)\mathcal{T}^{-1}\Big(\frac{v+v_\star}{2}\Big) + (\mathcal{T}^{-1}R_v^N)\mathcal{T}\Big(\frac{u+u_\star}{2}\Big) + (\mathcal{T}u_\star)(\mathcal{T}^{-1}v_\star). \end{aligned}$$

Since

$$\begin{aligned} g_0^+(x) &= u_0(x,0), & u_1(x,0) &= \sum_{n=1}^N d_n^+ g_n^+(x), \\ g_0^-(x) &= v_0(x,0), \quad \text{and} & v_1(x,0) &= \sum_{n=1}^N d_n^- g_n^-(x), \end{aligned}$$

we find that R_u^N and R_v^N satisfy

$$
\begin{aligned}
R_u^N(t) &= \mathrm{e}^{\partial_x^2 t}(a_0+b_0-g_0^+) \\
&\quad + \left[\mathrm{e}^{\partial_x^2 t} u_0(0) + c_+\partial_x \int_0^t \mathrm{d}s\ \mathrm{e}^{\partial_x^2(t-s)} u_0(s)^2 \right] - u_0(t) \\
&\quad + \left[\mathrm{e}^{\partial_x^2 t} u_1(0) + 2c_+\partial_x \int_0^t \mathrm{d}s\ \mathrm{e}^{\partial_x^2(t-s)} u_0(s)u_1(s) \right] - u_1(t) \\
&\quad - c_-\left[\partial_x \int_0^t \mathrm{d}s\ \mathrm{e}^{\partial_x^2(t-s)} \mathcal{T}^{-2}\Big((v_0(s)^2 + 2v_0(s)v_1(s)) \Big) \right] - \sum_{n=1}^N \mathrm{e}^{\partial_x^2 t} d_n^+ g_n^+ \\
&\quad + \widetilde{\mathcal{R}}_u[\mathbf{z},\mathbf{R}^N](t) + \mathcal{T}^{-1}\mathcal{R}_u[\mathbf{z}](t),
\end{aligned} \tag{3.9}
$$

$$
\begin{aligned}
R_v^N(t) &= \mathrm{e}^{\partial_x^2 t}(a_0-b_0-g_0^-) \\
&\quad + \left[\mathrm{e}^{\partial_x^2 t} v_0(0) + c_-\partial_x \int_0^t \mathrm{d}s\ \mathrm{e}^{\partial_x^2(t-s)} v_0(s)^2 \right] - v_0(t) \\
&\quad + \left[\mathrm{e}^{\partial_x^2 t} v_1(0) + 2c_-\partial_x \int_0^t \mathrm{d}s\ \mathrm{e}^{\partial_x^2(t-s)} v_0(s)v_1(s) \right] - v_1(t) \\
&\quad - c_+\left[\partial_x \int_0^t \mathrm{d}s\ \mathrm{e}^{\partial_x^2(t-s)} \mathcal{T}^{2}\Big((u_0(s)^2 + 2u_0(s)u_1(s)) \Big) \right] - \sum_{n=1}^N \mathrm{e}^{\partial_x^2 t} d_n^- g_n^- \\
&\quad + \widetilde{\mathcal{R}}_v[\mathbf{z},\mathbf{R}^N](t) + \mathcal{T}\mathcal{R}_v[\mathbf{z}](t),
\end{aligned} \tag{3.10}
$$

where

$$
\begin{aligned}
\widetilde{\mathcal{R}}_u[\mathbf{z},\mathbf{R}^N](t) &= c_+\mathrm{E}_0[h_{1,u}+h_{3,u}](t) - c_-\mathrm{E}_{-2}[h_{1,v}+h_{3,v}](t) + c_3\mathrm{E}_{-1}[h_2+h_4](t), \\
\widetilde{\mathcal{R}}_v[\mathbf{z},\mathbf{R}^N](t) &= c_-\mathrm{E}_0[h_{1,v}+h_{3,v}](t) - c_+\mathrm{E}_2[h_{1,u}+h_{3,u}](t) - c_3\mathrm{E}_1[h_2+h_4](t),
\end{aligned}
$$

with $\mathbf{R}^N = (R_u^N, R_v^N)$, $h_{1,u} = R_u^N(u+u_\star)$, $h_{3,u} = u_1^2$, $h_{1,v} = R_v^N(v+v_\star)$, $h_{3,v} = v_1^2$, $h_4 = (\mathcal{T}u_\star)(\mathcal{T}^{-1}v_\star)$, and

$$
\begin{aligned}
h_2 &= (\mathcal{T}R_u^N)\mathcal{T}^{-1}\Big(\frac{v+v_\star}{2}\Big) + (\mathcal{T}^{-1}R_v^N)\mathcal{T}\Big(\frac{u+u_\star}{2}\Big), \\
\mathrm{E}_\sigma[h](t) &= \partial_x \int_0^t \mathrm{d}s\ \mathrm{e}^{\partial_x^2(t-s)}\ \mathcal{T}^\sigma h(s).
\end{aligned}
$$

Note that we can write (3.9) and (3.10) as $\mathbf{R}^N = \mathcal{F}[\mathbf{z},\mathbf{R}^N]$. If we now consider $\mathbf{z}$ fixed, we can interpret $\mathbf{R}^N = \mathcal{F}[\mathbf{z},\mathbf{R}^N]$ as an equation for $\mathbf{R}^N$ which can be solved via a contraction-mapping argument. Namely, we will prove that, if $\|\mathbf{z}\| \le C\epsilon_0$, $\mathbf{R}^N \mapsto \mathcal{F}[\mathbf{z},\mathbf{R}^N]$ defines a contraction map inside the ball

$$
\|R_u^N\|_{2,\frac{3}{4}-\epsilon} + \|\mathrm{D}R_u^N\|_{2,\frac{5}{4}-\epsilon} + \|R_v^N\|_{2,\frac{3}{4}-\epsilon} + \|\mathrm{D}R_v^N\|_{2,\frac{5}{4}-\epsilon} \le C \tag{3.11}
$$

for $\epsilon = 2^{-N-2}$, provided $\{g_n^\pm\}_{n=0}^N$ and $\{d_n^\pm\}_{n=1}^N$ are appropriately chosen.

Basically, we will choose u_0, v_0, u_1, and v_1 in such a way that the second and third lines of (3.9) and (3.10) vanish. Note that if, for instance, we set the second (respectively, third) lines of (3.9) and (3.10) equal to zero, the resulting equalities are nothing but Duhamel's formulas for Burger's equations for u_0 and v_0 (respectively, for linearized Burger's equations for u_1 and v_1). Properties of solutions to these types of equations are studied in detail in section 4 below.

Once u_0, v_0, u_1, and v_1 are fixed, the time convolutions in the fourth lines of (3.9) and (3.10) can then be viewed as the solution of inhomogeneous heat equations with very specific inhomogeneous terms. Properties of solutions to this type of equation are studied in detail in section 5 below.

Assuming all results of sections 4 and 5, we now explain how to proceed to prove that $\mathcal{F}[\mathbf{z}, \mathbf{R}^N]$ defines a contraction map.

Obviously, the requirement on $\{g_n^\pm\}_{n=0}^N$ and $\{d_n^\pm\}_{n=1}^N$ is that the first four lines in (3.9) and (3.10) satisfy (3.11). This is achieved in the following way:

1. The first line of (3.9) (respectively, of (3.10)) satisfies (3.11) for any $g_0^\pm$ such that the total mass of $g_0^\pm$ is equal to that of $a_0 \pm b_0$, provided $a_0 \pm b_0$ and $g_0^\pm$ satisfy $\| x^2(a_0 \pm b_0)\|_2 < \infty$ and $\| x^2 g_0^\pm\|_2 < \infty$. This fixes the total mass of $g_0^\pm$. Note also that we need the estimate $\| x^2(a_0 \pm b_0)\|_2 < \infty$. There is no smallness assumption here, which is to be expected since generically $\| x^2(a(\cdot,t) \pm b(\cdot,t))\|_2$ will grow as $t \to \infty$. Note, on the other hand, that Proposition 7 shows that $\| x^2(a(\cdot,t) \pm b(\cdot,t))\|_2$ remains finite for all $t < \infty$, so requiring $\| x^2(a_0 \pm b_0)\|_2 < \infty$ is acceptable.
2. We can set the second lines in (3.9) and (3.10) equal to zero by picking for u_0 and v_0 any solution of Burger's equations
$$\partial_t u_0 = \partial_x^2 u_0 + c_+ \partial_x (u_0)^2 \quad \text{and} \quad \partial_t v_0 = \partial_x^2 v_0 + c_- \partial_x (v_0)^2$$
(or of the corresponding heat equations if either c_+ or c_- happen to be zero). In Proposition 12, we will prove that there exist unique functions u_0 and v_0 of the form given in (3.8) that satisfy the conditions of item 1 above (total mass and decay properties). This uniquely determines u_0 and v_0.
3. We can also set the third lines in (3.9) and (3.10) equal to zero, by picking any solutions u_1 and v_1 of linearized Burger's equations
$$\partial_t u_1 = \partial_x^2 u_1 + 2c_+ \partial_x (u_0 u_1) \quad \text{and} \quad \partial_t v_1 = \partial_x^2 v_1 + 2c_- \partial_x (v_0 v_1). \tag{3.12}$$
In Proposition 12, we will also prove that there is a choice of functions $\{g_n^\pm\}_{n=1}^N$ such that u_1 and v_1 in (3.8) satisfy (3.12) for any choice of the coefficients $\{d_n^\pm\}_{n=1}^N$. Furthermore, in Proposition 12, we will prove that the choice of functions can be made in such a way that $g_n^\pm(x)$ have Gaussian tails as $x \to \mp\infty$ and algebraic tails as $x \to \pm\infty$. This actually completely determines $g_n^\pm(x)$ up to multiplicative constants (this last indeterminacy will be removed when the coefficients $\{d_n^\pm\}_{n=1}^N$ are fixed).
4. We then further decompose the terms involving $g_n^\pm$ in the fourth lines in (3.9) and (3.10) as $g_n^\pm(x) = f_n(\mp x) + R_n^\pm(x)$. The definition and properties of $f_n(x)$ are given in Lemma 10. In particular, in Proposition 12, we will prove that $R_n^\pm(x)$ have zero total mass and Gaussian tails as $|x| \to \infty$, which implies that $\mathrm{e}^{\partial_x^2 t} R_n^\pm$ also satisfy (3.11).
5. Finally, in section 5, we will prove that the time convolution part of the fourth lines in (3.9) and (3.10) can be split into linear combinations of $\mathrm{e}^{\partial_x^2 t} f_n(\mp x)$, with $n = 1 \dots N+1$ plus a remainder that satisfies (3.11). The coefficients $\{d_n^\pm\}_{n=1}^N$ can then be set recursively by requiring that all of the terms with $n = 1 \dots N$ coming from the time convolution are canceled by those coming from item 4 above. This can always be done because the coefficient of $\mathrm{e}^{\partial_x^2 t} f_m(\mp x)$ in the time convolution part of the fourth lines in (3.9) and (3.10) depends only on $g_0^\pm$ if $m = 1$ and on $d_{m-1}^\pm$ if $m > 1$. The only term that cannot be set

to zero is the last term in the linear combination (the one with $n = N+1$), which is the one that "drives" the equations and fixes $\epsilon = 2^{-N-2}$.

The procedure outlined in 1–5 takes care of the first four lines in (3.9) and (3.10). We will then prove in section 7 that the terms $\mathcal{R}_{\{u,v\}}[\mathbf{z}]$ satisfy (3.11) and that

$$\sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha \widetilde{\mathcal{R}}_{\{u,v\}}[\mathbf{z}, \mathbf{R}^N]\|_{2,\frac{3}{4}+\frac{\alpha}{2}-\epsilon} \le C\epsilon_0 \sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha \mathbf{R}^N\|_{2,\frac{3}{4}+\frac{\alpha}{2}-\epsilon} + C, \tag{3.13}$$

$$\sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha (\Delta\widetilde{\mathcal{R}}_{\{u,v\}})\|_{2,\frac{3}{4}+\frac{\alpha}{2}-\epsilon} \le C\epsilon_0 \sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha (\Delta\mathbf{R}^N)\|_{2,\frac{3}{4}+\frac{\alpha}{2}-\epsilon}, \tag{3.14}$$

where $\Delta\widetilde{\mathcal{R}}_{\{u,v\}} = \widetilde{\mathcal{R}}_{\{u,v\}}[\mathbf{z}, \mathbf{R}_1^N] - \widetilde{\mathcal{R}}_{\{u,v\}}[\mathbf{z}, \mathbf{R}_2^N]$ and $\Delta\mathbf{R} = \mathbf{R}_1^N - \mathbf{R}_2^N$. This finally proves that $\mathcal{F}[\mathbf{z}, \mathbf{R}^N]$ defines a contraction map and that the solution of $\mathbf{R}^N = \mathcal{F}[\mathbf{z}, \mathbf{R}^N]$ satisfies (3.11), which completes the proof of Theorems 2 and 8. ∎

4. Burger-type equations. In this section, we consider particular solutions of Burger-type equations

$$\partial_t u_0 = \partial_x^2 u_0 + \gamma \partial_x u_0^2, \tag{4.1}$$

$$\partial_t u_n^\pm = \partial_x^2 u_n^\pm + 2\gamma \partial_x (u_0 u_n^\pm) \tag{4.2}$$

of the form

$$u_0(x,t) = \frac{1}{\sqrt{1+t}} g_0\left(\frac{x}{\sqrt{1+t}}\right) \quad \text{and} \quad u_n^\pm(x,t) = \frac{1}{(1+t)^{1-\frac{1}{2^{n+1}}}} g_n^\pm\left(\frac{x}{\sqrt{1+t}}\right). \tag{4.3}$$

We will show that for fixed $\mathrm{M}(u_0) = \int_{-\infty}^{\infty} \mathrm{d}x\ u_0(x,t) = \int_{-\infty}^{\infty} \mathrm{d}x\ g_0(x)$ small enough, there is a unique choice of g_0 and $g_n^\pm$ such that $g_n^\pm(x) = f_n(\mp x) + R_n^\pm(x)$, where

$$f_n(z) = \int_z^\infty \mathrm{d}\xi\ \frac{\xi \mathrm{e}^{-\frac{\xi^2}{4}}}{(\xi - z)^{1-\frac{1}{2^n}}} \tag{4.4}$$

and $R_n^\pm$ has zero mean and Gaussian tails as $|x| \to \infty$. In particular, $g_n^\pm(x)$ decays algebraically as $x \to \pm\infty$, as is apparent from (4.4).

Before proceeding to our study of (4.1) and (4.2), we prove key properties of the functions f_n.

LEMMA 10. *Fix $1 \le n < \infty$. The function f_n is the unique solution of*

$$\begin{aligned} &\partial_z^2 f_n(z) + \frac{1}{2} z \partial_z f_n(z) + \left(1 - \frac{1}{2^{n+1}}\right) f_n(z) = 0, \qquad \textit{with} \\ &f_n(0) = 2^{\frac{1}{2^n}} \Gamma\left(\frac{1+2^{-n}}{2}\right) \qquad \textit{and} \qquad \lim_{z\to\infty} z^{-1+\frac{1}{2^n}} \mathrm{e}^{\frac{z^2}{4}} f_n(z) < \infty. \end{aligned} \tag{4.5}$$

It satisfies $\int_{-\infty}^{\infty} \mathrm{d}z\ f_n(z) = 0$, and there exists a constant $C(n)$ such that

$$\begin{aligned} &\sup_{z\in\mathbf{R}} \sum_{m=0}^{2} \rho_{\frac{1}{2^n}-m, 1+m-\frac{1}{2^n}}(z) \left|\partial_z^m\Big(z f_n(z) + 2\partial_z f_n(z)\Big)\right| \le C(n), \\ &\qquad\qquad \sup_{z\in\mathbf{R}} \sum_{m=0}^{3} \rho_{\frac{1}{2^n}-1-m, 2+m-\frac{1}{2^n}}(z) |\partial_z^m f_n(z)| \le C(n), \end{aligned} \tag{4.6}$$

where

$$\rho_{p,q}(z) = \begin{cases} (1+z^2)^{\frac{p}{2}} \mathrm{e}^{\frac{z^2}{4}} & \text{if } z \geq 0, \\ (1+z^2)^{\frac{q}{2}} & \text{if } z \leq 0. \end{cases}$$

Proof. We first note that f_n can be written as

$$f_n(z) = \int_0^\infty \mathrm{d}\xi \; \frac{(\xi+z)\mathrm{e}^{-\frac{(\xi+z)^2}{4}}}{\xi^{1-\frac{1}{2^n}}} = -2 \int_0^\infty \mathrm{d}\xi \; \xi^{\frac{1}{2^n}-1} \partial_\xi \Big(\mathrm{e}^{-\frac{(z+\xi)^2}{4}} \Big). \tag{4.7}$$

This shows that f_n solves (4.5) since, by defining $\mathcal{L}f \equiv \partial_z^2 f + \frac{1}{2} z \partial_z f + (1 - \frac{1}{2^{n+1}}) f$, we find

$$\mathcal{L}f_n(z) = \int_0^\infty \mathrm{d}\xi \; \Big[\xi^{\frac{1}{2^n}} \partial_\xi^2 \Big(\mathrm{e}^{-\frac{(z+\xi)^2}{4}} \Big) - \frac{1}{2^{n+1}} (-2) \xi^{\frac{1}{2^n}-1} \partial_\xi \Big(\mathrm{e}^{-\frac{(z+\xi)^2}{4}} \Big) \Big] = 0.$$

Obviously, $f_n(z)$ is finite for all finite z, so we need only to prove that f_n satisfies the correct decay properties as $|z| \to \infty$ so that (4.6) holds. It is apparent from (4.4) that f_n decays like a (modified) Gaussian as $z \to \infty$ and algebraically as $z \to -\infty$. Furthermore, substituting $f(z) = C|z|^{p_1}$ and $f(z) = C|z|^{p_2} \mathrm{e}^{-\frac{z^2}{4}}$ into $\mathcal{L}f = 0$ shows that the only decay rates compatible with $\mathcal{L}f = 0$ are $p_1 = -2 + \frac{1}{2^n}$ and $p_2 = 1 - \frac{1}{2^n}$.

We now complete the proof of the decay estimates (4.6). Let $F_{n,m}(\xi, z) = \partial_z^m ((\xi + z)\mathrm{e}^{-\frac{(\xi+z)^2}{4}})$ and $G_{n,m}(\xi, z) = \partial_z^m (z F_n(\xi, z) + 2 \partial_z F_n(\xi, z))$.

We first consider the case $z > 0$ and note that $F_{n,m}$ and $G_{n,m}$ satisfy

$$|F_{n,m}(\xi, z)| \leq |F_{n,m}(0, z)| \quad \text{and} \quad |G_{n,m}(\xi, z)| \leq |G_{n,m}(0, z)|,$$

respectively, for all $\xi \geq 0$ if $z \geq z_0$ for some z_0 large enough. We thus get, e.g.,

$$\begin{aligned} |f_n(z)| &= \Big| \int_0^\infty \mathrm{d}\xi \; F_{n,0}(\xi, z) \xi^{\frac{1}{2^n}-1} \Big| \\ &\leq |F_{n,0}(0, z)| \int_0^{z^{-1}} \mathrm{d}\xi \; \xi^{\frac{1}{2^n}-1} + z^{1-\frac{1}{2^n}} \int_{z^{-1}}^\infty \mathrm{d}\xi \; |F_{n,0}(\xi, z)| \leq C z^{1-\frac{1}{2^n}} \mathrm{e}^{-\frac{z^2}{4}}. \end{aligned}$$

The estimates on $|\partial_z^m (z f_n(z) + 2\partial_z f_n(z))|$ and $|\partial_z^{1+m} f_n(z)|$ when $z > 0$ and $m \geq 1$ can be done in exactly the same way; hence we omit the details.

We now consider the case $z < 0$ and note that $F_{n,m}$ and $G_{n,m}$ satisfy

$$|F_{n,m}(\xi, z)| \leq \Big| F_{n,m}\Big(-\frac{z}{2}, z\Big) \Big| \quad \text{and} \quad |G_{n,m}(\xi, z)| \leq \Big| G_{n,m}\Big(-\frac{z}{2}, z\Big) \Big|,$$

respectively, for all $0 \leq \xi \leq -\frac{z}{2}$ if $z \leq -z_0$ for some z_0 large enough. We thus find (integrating by parts in the second integral below)

$$\begin{aligned} |f_n(z)| &= \Big| \int_0^\infty \mathrm{d}\xi \; F_{n,0}(\xi, z) \xi^{\frac{1}{2^n}-1} \Big| \\ &\leq \Big| F_{n,0}\Big(-\frac{z}{2}, z\Big) \Big| \int_0^{-\frac{z}{2}} \mathrm{d}\xi \; \xi^{\frac{1}{2^n}-1} + \Big| \int_{-\frac{z}{2}}^\infty \mathrm{d}\xi \; F_{n,0}(\xi, z) \xi^{\frac{1}{2^n}-1} \Big| \\ &\leq C|z|^{\frac{1}{2^n}-1} \mathrm{e}^{-\frac{z^2}{16}} + 2\Big(1 - \frac{1}{2^n}\Big) \int_{-\frac{z}{2}}^\infty \mathrm{d}\xi \; \mathrm{e}^{-\frac{(\xi+z)^2}{4}} \xi^{\frac{1}{2^n}-2} \leq C|z|^{\frac{1}{2^n}-2}. \end{aligned}$$

Since the remaining estimates can again be done in exactly the same way, we omit the details. It remains only to show that $f_n(z)$ has zero total mass. This follows from

$$\int_{-\infty}^{\infty} \mathrm{d}z\ f_n(z) = \left(\frac{1}{2} - \frac{1}{2^{n+1}}\right)^{-1} \int_{-\infty}^{\infty} \mathrm{d}z\ \mathcal{L} f_n(z) = 0,$$

since $\partial_z^2 f_n, z\partial_z f_n$ and f_n are all integrable over $\mathbf{R}$. ∎

Remark 11. By using the representation (4.7), splitting the integration interval into $[0, 2^{-\frac{n}{2}})$ and $[2^{-\frac{n}{2}}, \infty)$, integrating by parts, and letting $n \to \infty$, one can prove that

$$\lim_{n\to\infty} 2^{-n} f_n(z) = z\mathrm{e}^{-\frac{z^2}{4}},$$

which shows that the constant $C(n)$ in (4.6) grows at most like 2^n.

We can now study in detail the solutions of (4.1) and (4.2) that are of the form (4.3).

PROPOSITION 12. *Fix $1 \le n < \infty$. For all $\alpha, \gamma \in \mathbf{R}$ with $|\alpha\gamma|$ small enough, there exist unique functions u_0 and $u_n^{\pm}$ of the form* (4.3) *that solve* (4.1) *and* (4.2), *with g_0 satisfying*

$$\int_{-\infty}^{\infty} \mathrm{d}z\ g_0(z) = \alpha, \qquad \sum_{m=0}^{3} \frac{\mathrm{e}^{\frac{z^2}{4}}}{(\sqrt{1+z^2})^m} |\partial_z^m g_0(z)| \le C|\alpha|$$

and with $g_n^{\pm}(z) = f_n(\mp z) + R_n^{\pm}(z)$, where $R_n^{\pm}$ satisfy

$$\int_{-\infty}^{\infty} \mathrm{d}z\ R_n^{\pm}(z) = 0 \quad \textit{and} \quad \sup_{z\in\mathbf{R}} \sum_{m=0}^{3} \frac{\mathrm{e}^{\frac{z^2}{4}}}{(\sqrt{1+z^2})^{1+m-\frac{1}{2^n}}} |\partial_z^m R_n^{\pm}(z)| \le C|\alpha\gamma|.$$

Proof. The (unique) solution of (4.1) of the form $u_0(x,t) = \frac{1}{\sqrt{1+t}} g_0(\frac{x}{\sqrt{1+t}})$ satisfying $\int_{-\infty}^{\infty} \mathrm{d}z\ g_0(z) = \alpha$ is given by

$$g_0(z) = \frac{\tanh(\frac{\alpha\gamma}{2})\mathrm{e}^{-\frac{z^2}{4}}}{\gamma\sqrt{\pi}(1 + \tanh(\frac{\alpha\gamma}{2})\mathrm{erf}(\frac{z}{2}))}.$$

In particular, we have

$$\sum_{m=0}^{3} \frac{\mathrm{e}^{\frac{z^2}{4}}}{(\sqrt{1+z^2})^m} |\partial_z^m g_0(z)| \le C|\alpha|. \tag{4.8}$$

We next note that substituting (4.3) into (4.2) gives

$$\begin{aligned} 0 &= \partial_z^2 g_n^{\pm}(z) + \frac{1}{2} z\partial_z g_n^{\pm}(z) + \left(1 - \frac{1}{2^{n+1}}\right) g_n^{\pm}(z) + 2\gamma\partial_z(g_0(z) g_n^{\pm}(z)) \\ &\equiv \mathcal{L} g_n^{\pm}(z) + 2\gamma\partial_z(u_0(z) g_n^{\pm}(z)). \end{aligned} \tag{4.9}$$

We formally have (using integration by parts)

$$\int_{-\infty}^{\infty} \mathrm{d}z\ g_n^{\pm}(z) = \left(\frac{1}{2} - \frac{1}{2^{n+1}}\right)^{-1} \int_{-\infty}^{\infty} \mathrm{d}z\ \mathcal{L} g_n^{\pm}(z) + 2\gamma\partial_z(u_0(z) g_n^{\pm}(z)) = 0, \tag{4.10}$$

which shows that $g_n^\pm$ have zero total mass, *provided the formal manipulations above are justified*, i.e., provided $g_n^\pm$ and its derivatives decay fast enough so that the integrals are convergent.

As is easily seen, $f_n(z)$ and $f_n(-z)$ are two linearly independent solutions of $\mathcal{L}f = 0$, whose general solution can thus be written as $c_1 f_n(z) + c_2 f_n(-z)$. By using the variation of constants formula, we get that the solution of (4.9) satisfies the integral equation

$$g_n^\pm(z) = f_n(z)\biggl(c_1^\pm + 2\gamma \int_0^z \mathrm{d}\xi \; \frac{f_n(-\xi)\partial_\xi(g_0(\xi)g_n^\pm(\xi))}{W(\xi)}\biggr) + f_n(-z)\biggl(c_2^\pm - 2\gamma \int_0^z \mathrm{d}\xi \; \frac{f_n(\xi)\partial_\xi(g_0(\xi)g_n^\pm(\xi))}{W(\xi)}\biggr),$$

where the Wronskian $W(z)$ is given by $W(z) = f_n(z)\partial_z f_n(-z) - f_n(-z)\partial_z f_n(z)$ and $c_1^\pm$ and $c_2^\pm$ are free parameters. Note that $W(z)$ satisfies $\partial_z W(z) = -\frac{z}{2}W(z)$ and hence $W(z) = W(0)\mathrm{e}^{-\frac{z^2}{4}}$ for some $W(0) \neq 0$. We now set $c_1^\pm$ and $c_2^\pm$ in such a way that (after integration by parts) we have

$$\begin{aligned} g_n^\pm(z) &= f_n(\mp z) + R[g_n^\pm](z), \\ R[g_n^\pm](z) &= \frac{\gamma}{W(0)} f_n(z) \int_{-\infty}^z \mathrm{d}\xi \; \mathrm{e}^{\frac{\xi^2}{4}}(\xi f_n(-\xi) + 2\partial_\xi f_n(-\xi))g_0(\xi)g_n^\pm(\xi) \\ &\quad + \frac{\gamma}{W(0)} f_n(-z) \int_z^\infty \mathrm{d}\xi \; \mathrm{e}^{\frac{\xi^2}{4}}(\xi f_n(\xi) + 2\partial_\xi f_n(\xi))g_0(\xi)g_n^\pm(\xi). \end{aligned} \tag{4.11}$$

By using Lemma 10 and (4.8), it is then easy to show that, for $|\alpha\gamma|$ small enough, (4.11) defines a contraction map in the norm

$$|f|_{2-\frac{1}{2^n}} \equiv \sup_{z\in\mathbf{R}} (\sqrt{1+z^2})^{2-\frac{1}{2^n}} |f(z)|.$$

Namely, we have the improved decay rates

$$\sup_{z\in\mathbf{R}} \sum_{m=0}^{1} \frac{\mathrm{e}^{\frac{z^2}{4}}}{(\sqrt{1+z^2})^{1+m-\frac{1}{2^n}}} |\partial_z^m R[g_n^\pm](z)| \leq C|\alpha\gamma| \; |g_n^\pm|_{2-\frac{1}{2^n}}.$$

This shows that (4.11) has a (locally) unique solution among functions with $|f|_{2-\frac{1}{2^n}} \leq c_0$ if $|\alpha\gamma|$ is small enough. In particular, there holds

$$\sup_{z\in\mathbf{R}} \sum_{m=0}^{1} \frac{\mathrm{e}^{\frac{z^2}{4}}}{(\sqrt{1+z^2})^{1+m-\frac{1}{2^n}}} |\partial_z^m R[g_n^\pm](z)| \leq C|\alpha\gamma|,$$

from which we deduce, by using again (4.11) and Lemma 10, that $|\mathrm{D}g_n^\pm|_{3-\frac{1}{2^n}} \leq c_1$ and thus

$$\sup_{z\in\mathbf{R}} \frac{\mathrm{e}^{\frac{z^2}{4}}}{(\sqrt{1+z^2})^{3-\frac{1}{2^n}}} |\partial_z^2 R[g_n^\pm](z)| \leq C|\alpha\gamma|.$$

Iterating this procedure shows that $|\mathrm{D}^m g_n^\pm|_{2+m-\frac{1}{2^n}} \leq c_m$ and that

$$\sup_{z\in\mathbf{R}} \sum_{m=0}^{3} \frac{\mathrm{e}^{\frac{z^2}{4}}}{(\sqrt{1+z^2})^{1+m-\frac{1}{2^n}}} |\partial_z^m R[g_n^\pm](z)| \leq C|\alpha\gamma|$$

as claimed. In turn, this proves that the formal manipulations in (4.10) are justified, so that the functions $g_n^{\pm}(z)$ have zero total mass, which shows that the remainders $R[g_n^{\pm}](z)$ have zero total mass as claimed, since $R[g_n^{\pm}](z) = g_n^{\pm}(z) - f_n(\pm z)$ and since both $g_n^{\pm}(z)$ and $f_n(z)$ have zero total mass. $\square$

5. Inhomogeneous heat equations. In this section, we consider solutions of inhomogeneous heat equations of the form

$$\partial_t u = \partial_x^2 u + \partial_x \left((1+t)^{\frac{1}{2^n} - \frac{3}{2}} f\left(\frac{x - 2\sigma t}{\sqrt{1+t}} \right) \right), \qquad u(x,0) = 0, \tag{5.1}$$

where f is a regular function having Gaussian decay at infinity. Solutions of (5.1) satisfy the following theorem.

THEOREM 13. *Let* $1 \leq n < \infty$, $\sigma = \pm 1$, $\Xi(x) = \mathrm{e}^{\frac{x^2}{8}}$, $\mathrm{M}(f) = \int_{-\infty}^{\infty} \mathrm{d}z\ f(z)$, *and*

$$u_n(x,t) = \frac{\sigma}{(1+t)^{1 - \frac{1}{2^{n+1}}}} \frac{2^{-1-\frac{1}{2^n}}}{\sqrt{4\pi}} f_n\left(\frac{-\sigma x}{\sqrt{1+t}} \right), \quad \textit{with} \quad f_n(z) = \int_z^{\infty} \mathrm{d}\xi\ \frac{\xi \mathrm{e}^{-\frac{\xi^2}{4}}}{(\xi - z)^{1 - \frac{1}{2^n}}}. \tag{5.2}$$

The solution u *of* (5.1) *satisfies*

$$\|u - \mathrm{M}(f)\, u_n\|_{2,\frac{3}{4}^\star} + \|\mathrm{D}(u - \mathrm{M}(f)\, u_n)\|_{2,\frac{5}{4}^\star} \leq C \sum_{m=0}^{2} \|\Xi \mathrm{D}^m f\|_\infty \tag{5.3}$$

for all f *such that the r.h.s. of* (5.3) *is finite.*

Remark 14. Note that, while $u \to \mathrm{M}(f) u_n$ as $t \to \infty$ in the Sobolev norm (5.3), it does not do so in spatially weighted norms such as $\mathrm{L}^2(\mathbf{R}, x^2 \mathrm{d}x)$, as u_n has infinite spatial moments for all times, while all moments of u are bounded for finite time.

Proof. We first define

$$F(\xi) = \int_{-\infty}^{\xi} \mathrm{d}z \left(f(z) - \mathrm{M}(f)\, \frac{\mathrm{e}^{-\frac{z^2}{4}}}{\sqrt{4\pi}} \right), \qquad \text{with} \qquad \mathrm{M}(f) = \int_{-\infty}^{\infty} \mathrm{d}z\ f(z), \tag{5.4}$$

and note that F satisfies

$$\|\mathrm{D}^3 F\|_1 + \sum_{m=0}^{2} \|\rho \mathrm{D}^m F\|_1 + \sum_{m=1}^{2} \|\mathrm{D}^m F\|_2 \leq C \sum_{m=0}^{2} \|\Xi \mathrm{D}^m f\|_\infty, \tag{5.5}$$

where $\rho(x) = \sqrt{1+x^2}$. Namely, we first note that $\|\rho F\|_1 \leq \|\hat{F}\|_2 + \|\hat{F}''\|_2$ and $\hat{F}(k) = (ik)^{-1}(\hat{f}(k) - \hat{f}(0)\mathrm{e}^{-k^2})$. Then, since $\|\Xi f\|_\infty < \infty$ implies that $\hat{f}$ is analytic, $\hat{F}$ is regular near $k = 0$. The proof of (5.5) now follows from elementary arguments.

We finally note that it follows from (5.4) that

$$\begin{aligned} (1+t)^{\frac{1}{2^n} - \frac{3}{2}} f\left(\frac{x - 2\sigma t}{\sqrt{1+t}} \right) &= \mathrm{M}(f) A(x,t) + \partial_x B(x,t), \qquad \text{where} \\ A(x,t) &= \frac{(1+t)^{\frac{1}{2^n} - \frac{3}{2}}}{\sqrt{4\pi}} \mathrm{e}^{-\frac{(x - 2\sigma t)^2}{4(1+t)}}, \\ B(x,t) &= (1+t)^{\frac{1}{2^n} - 1} \partial_x F\left(\frac{x - 2\sigma t}{\sqrt{1+t}} \right). \end{aligned} \tag{5.6}$$

The proof of (5.3) is then completed by considering separately the solutions of heat equations with inhomogeneous terms given by $\partial_x A(x,t)$ and $\partial_x^2 B(x,t)$. This is done in Propositions 15 and 16 below. ∎

PROPOSITION 15. *Let $\sigma = \pm 1$, $1 \le n < \infty$, and let u_n be defined as in* (5.2). *The solution u of*

$$\partial_t u = \partial_x^2 u + \partial_x A, \qquad u(x,0) = 0, \tag{5.7}$$

with A defined in (5.6), *satisfies*

$$\|u - u_n\|_{2,\frac{3}{4}} + \|\mathrm{D}(u - u_n)\|_{2,\frac{5}{4}} \le C. \tag{5.8}$$

Proof. The solution of (5.7) is given by

$$u(x,t) = \partial_x \int_0^t \mathrm{d}s \int_{-\infty}^{\infty} \mathrm{d}y \frac{\mathrm{e}^{-\frac{(x-y)^2}{4(t-s)}}}{\sqrt{4\pi(t-s)}} \frac{\mathrm{e}^{-\frac{(y-2\sigma s)^2}{4(1+s)}}}{\sqrt{4\pi}(1+s)^{\frac{3}{2}-\frac{1}{2^n}}}. \tag{5.9}$$

To motivate our result, we note that performing the y-integration and changing variables from s to $\xi \equiv \frac{2s - \sigma x}{\sqrt{1+t}}$ in (5.9) leads to

$$\begin{aligned}\lim_{t\to\infty}(1+t)^{1-\frac{1}{2^{n+1}}} u(-\sigma z\sqrt{1+t}, t) &= \lim_{t\to\infty} \frac{\sigma 2^{-1-\frac{1}{2^n}}}{\sqrt{4\pi}} \int_z^{\frac{2t}{\sqrt{1+t}}+z} \mathrm{d}\xi \; \frac{\xi \mathrm{e}^{-\frac{\xi^2}{4}}}{(\xi - z + \frac{2}{\sqrt{1+t}})^{1-\frac{1}{2^n}}} \\ &= \frac{\sigma 2^{-1-\frac{1}{2^n}}}{\sqrt{4\pi}} f_n(z).\end{aligned}$$

More formally, taking the Fourier transform of (5.9) gives

$$\hat{u}(k,t) = ik\mathrm{e}^{-k^2(1+t)} \int_0^t \mathrm{d}s \; \frac{\mathrm{e}^{2ik\sigma s}}{(1+s)^{1-\frac{1}{2^n}}}.$$

We now use that

$$\begin{aligned}\left| \int_0^t \mathrm{d}s \; \frac{\mathrm{e}^{2ik\sigma s}}{(1+s)^{1-\frac{1}{2^n}}} - \int_0^t \mathrm{d}s \; \frac{\mathrm{e}^{2ik\sigma s}}{s^{1-\frac{1}{2^n}}} \right| &\le C(n), \\ \int_0^t \mathrm{d}s \; \frac{\mathrm{e}^{2ik\sigma s}}{s^{1-\frac{1}{2^n}}} &= |k|^{-\frac{1}{2^n}} \Big(\theta(\sigma k) J_n(|k|t) + \theta(-\sigma k) \overline{J_n(|k|t)} \Big),\end{aligned}$$

where $\theta(k)$ is the Heaviside step function and we defined

$$J_n(z) = \int_0^z \mathrm{d}s \; \frac{\mathrm{e}^{2is}}{s^{1-\frac{1}{2^n}}}$$

for $z \ge 0$. This function satisfies

$$\sup_{z\ge 0} z^{1-\frac{1}{2^n}} |J_n(z) - J_{n,\infty}| \le \frac{1}{2} \quad \text{for} \quad J_{n,\infty} = \lim_{z\to\infty} J_n(z).$$

Now define

$$\widehat{u_n}(k,t) = ik\mathrm{e}^{-k^2(1+t)} |k|^{-\frac{1}{2^n}} \Big(\theta(\sigma k) J_{n,\infty} + \theta(-\sigma k) \overline{J_{n,\infty}} \Big). \tag{5.10}$$

We have

$$|\hat{u}(k,t) - \widehat{u_n}(k,t)| \le (C(n)|k| + t^{-1+\frac{1}{2^n}})\mathrm{e}^{-k^2(1+t)} \le \frac{\tilde{C}(n)}{\sqrt{t}}\mathrm{e}^{-\frac{k^2(1+t)}{2}}, \tag{5.11}$$

from which (5.8) follows by direct integration. We complete the proof by showing that the inverse Fourier transform of the function $\widehat{u_n}(k,t)$ defined in (5.10) satisfies

$$u_n(x,t) = \frac{\sigma}{(1+t)^{1-\frac{1}{2^{n+1}}}} \frac{2^{-1-\frac{1}{2^n}}}{\sqrt{4\pi}} f_n\left(\frac{-\sigma x}{\sqrt{1+t}}\right) \quad \text{for} \quad f_n(z) = \int_z^\infty \mathrm{d}\xi\; \frac{\xi \mathrm{e}^{-\frac{\xi^2}{4}}}{(\xi - z)^{1-\frac{1}{2^n}}}. \tag{5.12}$$

This follows easily from the fact that

$$\widehat{u_n}(k,t) = (1+t)^{-\frac{1}{2}+\frac{1}{2^{n+1}}} \widehat{u_n}(k\sqrt{1+t}, 0)$$

and that, since

$$f_n(z) = \int_0^\infty \mathrm{d}\xi\; \frac{(z+\xi)\mathrm{e}^{-\frac{(z+\xi)^2}{4}}}{\xi^{1-\frac{1}{2^n}}},$$

we get

$$\begin{aligned}\frac{\sigma 2^{-1-\frac{1}{2^n}}}{\sqrt{4\pi}} \widehat{f_n}(-\sigma k) &= 2^{-\frac{1}{2^n}} ik\mathrm{e}^{-k^2} \int_0^\infty \mathrm{d}\xi\; \frac{\mathrm{e}^{ik\sigma\xi}}{\xi^{1-\frac{1}{2^n}}} = ik\mathrm{e}^{-k^2}|k|^{-\frac{1}{2^n}} \int_0^\infty \mathrm{d}\xi\; \frac{\mathrm{e}^{2i\mathrm{sign}(k\sigma)\xi}}{\xi^{1-\frac{1}{2^n}}} \\ &= ik\mathrm{e}^{-k^2}|k|^{-\frac{1}{2^n}}\Big(\theta(k\sigma)J_{n,\infty} + \theta(-k\sigma)\overline{J_{n,\infty}}\Big) = \widehat{u_n}(k,0)\end{aligned}$$

as claimed. ◻

Proposition 16. *Let $\sigma = \pm 1$, $1 \le n < \infty$, and $\rho(x) = \sqrt{1+x^2}$. The solution u of*

$$\partial_t u = \partial_x^2 u + \partial_x^2 B, \qquad u(x,0) = 0, \tag{5.13}$$

with B defined in (5.6), *satisfies*

$$\|u\|_{2,\frac{3}{4}\star} + \|\mathrm{D}u\|_{2,\frac{5}{4}\star} \le C \left(\|\mathrm{D}^3 F\|_1 + \sum_{m=0}^{2} \|\rho \mathrm{D}^m F\|_1 + \sum_{m=1}^{2} \|\mathrm{D}^m F\|_2\right) \tag{5.14}$$

for all F for which the r.h.s. of (5.14) *is finite.*

Proof. We first note that the Fourier transform of u is given by

$$\hat{u}(k,t) = -k^2 \int_0^t \mathrm{d}s\; \mathrm{e}^{-k^2(t-s)-2ik\sigma s} \hat{F}(k\sqrt{1+s})(1+s)^{\frac{1}{2^n}-\frac{1}{2}},$$

which implies that

$$\|(1-\mathbb{Q})u\|_{2,\frac{3}{4}} + \|(1-\mathbb{Q})\mathrm{D}u\|_{2,\frac{5}{4}} \le C\Big(\|\mathrm{D}F\|_2 + \|\mathrm{D}^2 F\|_2\Big) \sup_{0\le t\le 1} \int_0^t \frac{\mathrm{d}s}{\sqrt{t-s}}.$$

Here $\mathbb{Q}$ is again defined as the characteristic function for $t \ge 1$. Next, by integrating by parts, we find

$$\hat{u}(k,t) = \frac{ik\hat{F}(k)\mathrm{e}^{-k^2 t}}{2\sigma} - \frac{ik\hat{F}(k\sqrt{1+t})\mathrm{e}^{-2ik\sigma t}}{2\sigma(1+t)^{\frac{1}{2}-\frac{1}{2^n}}} + \hat{N}(k,t),$$

$$\text{where}\qquad \hat{N}(k,t) = \frac{ik}{2\sigma}\int_0^t \mathrm{d}s\ \mathrm{e}^{-k^2(t-s)-2ik\sigma s}\Big(k^2+\partial_s\Big)\left(\frac{\hat{F}(k\sqrt{1+s})}{(1+s)^{\frac12-\frac{1}{2^n}}}\right).$$

We then note that

$$\|u-N\|_{2,\frac34}+\|\mathrm{D}(u-N)\|_{2,\frac54}\le C\Big(\|F\|_1+\|\mathrm{D}F\|_2+\|\mathrm{D}^2F\|_2\Big)$$

and that, by defining $\hat{G}(k)=\frac12\partial_k\hat{F}(k)$, we have $\hat{N}(k,t)=\hat{N}_0(k,t)+\hat{N}_1(k,t)+\hat{N}_2(k,t)$, where

$$\begin{aligned}
\hat{N}_0(k,t) &= \frac{ik^3}{2\sigma}\int_0^t \mathrm{d}s\ \mathrm{e}^{-k^2(t-s)-2ik\sigma s}\left(\frac{\hat{F}(k\sqrt{1+s})}{(1+s)^{\frac12-\frac{1}{2^n}}}\right),\\
\hat{N}_1(k,t) &= \frac{ik^2}{2\sigma}\int_0^t \mathrm{d}s\ \mathrm{e}^{-k^2(t-s)-2ik\sigma s}\left(\frac{\hat{G}(k\sqrt{1+s})}{(1+s)^{1-\frac{1}{2^n}}}\right),\\
\hat{N}_2(k,t) &= \frac{ik}{2\sigma}\left(\frac{1}{2^n}-\frac12\right)\int_0^t \mathrm{d}s\ \mathrm{e}^{-k^2(t-s)-2ik\sigma s}\left(\frac{\hat{F}(k\sqrt{1+s})}{(1+s)^{\frac32-\frac{1}{2^n}}}\right).
\end{aligned}$$

The procedure is now similar to that outlined in the proof of Theorem 6: Split the integration intervals into $[0,\frac{t}{2}]$ and $[\frac{t}{2},t]$, and distribute the derivatives (k-factors) either on the functions F and G or on the Gaussian. By introducing the notation

$$B_1\Big[\begin{smallmatrix}p_1,q_1\\ p_2,q_2\end{smallmatrix}\Big](t)\equiv\int_0^{\frac{t}{2}}\mathrm{d}s\ \frac{(1+s)^{-q_1}}{(t-s)^{p_1}}+\int_{\frac{t}{2}}^t\mathrm{d}s\ \frac{(1+s)^{-q_2}}{(t-s)^{p_2}}, \tag{5.15}$$

we then find that

$$\begin{aligned}
\|\mathbb{Q}\mathrm{D}^\alpha N_0\|_{2,\frac34+\frac{\alpha}{2}} &\le C(\|F\|_1+\|\mathrm{D}^{2+\alpha}F\|_1)\sup_{t\ge1}t^{\frac34+\frac{\alpha}{2}}\ B_1\Big[\begin{smallmatrix}\frac74+\frac{\alpha}{2},0\\ \frac34,1+\frac{\alpha}{2}\end{smallmatrix}\Big](t),\\
\|\mathbb{Q}\mathrm{D}^\alpha N_1\|_{2,\frac34+\frac{\alpha}{2}} &\le C(\|G\|_1+\|\mathrm{D}^{1+\alpha}G\|_1)\sup_{t\ge1}t^{\frac34+\frac{\alpha}{2}}\ B_1\Big[\begin{smallmatrix}\frac54+\frac{\alpha}{2},\frac12\\ \frac34,1+\frac{\alpha}{2}\end{smallmatrix}\Big](t),\\
\|\mathbb{Q}\mathrm{D}^\alpha N_2\|_{2,\frac34+\frac{\alpha}{2}\star} &\le C(\|F\|_1+\|\mathrm{D}^{\alpha}F\|_1)\sup_{t\ge1}\frac{t^{\frac34+\frac{\alpha}{2}}}{\ln(2+t)}\ B_1\Big[\begin{smallmatrix}\frac34+\frac{\alpha}{2},1\\ \frac34,1+\frac{\alpha}{2}\end{smallmatrix}\Big](t)
\end{aligned}$$

for $\alpha=0,1$. The proof is completed by a straightforward application of Lemma 18 below, where we consider generalizations of the function B_1 in (5.15), since those will occur later on in sections 6 and 7 (see Definition 17 below). ∎

6. Proof of Theorem 6, continued. In view of the estimates (2.6) and (2.8) on $\mathrm{e}^{\mathrm{L}t}$ and h, respectively, the estimates needed to conclude the proof of Theorem 6 will naturally involve the functions B_0 and B, which are defined as follows.

DEFINITION 17. *We define*

$$B_0[q](t)=\int_0^t\mathrm{d}s\frac{\mathrm{e}^{-\frac{t-s}{8}}}{\sqrt{t-s}(1+s)^q},$$

$$B\Big[\begin{smallmatrix}p_1,q_1,r_1\\ p_2,q_2,r_2,r_3\end{smallmatrix}\Big](t)=\int_0^{\frac{t}{2}}\mathrm{d}s\ \frac{(1+s)^{-q_1}}{(t-s)^{p_1}(t-1+s)^{r_1}}+\int_{\frac{t}{2}}^t\mathrm{d}s\ \frac{(1+s)^{-q_2}\ln(2+s)^{r_3}}{(t-s)^{p_2}(t-1+s)^{r_2}}. \tag{6.1}$$

These functions satisfy the following estimates.

LEMMA 18. *Let $0 \le p_2 < 1$, $0 \le r_2 \le 1 - p_2$, $p_1, q_1, q_2, r_1 \ge 0$, and $r_3 \in \{0, 1\}$. There exists a constant C such that for all $t \ge 0$ there holds*

$$B_0[q_1](t) \le C(1+t)^{-q_1},$$

$$B\Big[{}^{p_1,q_1,r_1}_{p_2,q_2,r_2,r_3}\Big](t) \le C \ \ln(2+t)^{\alpha} \begin{cases} \frac{1}{(1+t)^{\beta}} & \text{if} \quad 0 \le p_1 \le 1, \\ \frac{1}{t^{p_1-1}\ (1+t)^{\beta-p_1+1}} & \text{if} \quad p_1 > 1, \end{cases} \tag{6.2}$$

where $\beta = \min(p_1 + \min(q_1 - 1, 0) + r_1, p_2 + q_2 + r_2 - 1)$, $\alpha = \max(\delta_{q_1,1}, \delta_{p_2+r_2,1} + r_3)$, and $\delta_{i,j}$ is the Kronecker delta. Furthermore, since

$$B_1\Big[{}^{p_1,q_1}_{p_2,q_2}\Big](t) = B\Big[{}^{p_1,q_1,0}_{p_2,q_2,0,0}\Big](t),$$

the estimate in (6.2) *applies for B_1 as well.*

Proof. The proof follows immediately from

$$B_0[q_1](t) \le \mathrm{e}^{-\frac{t}{16}} \int_0^{\frac{t}{2}} \frac{\mathrm{d}s}{\sqrt{t-s}} + \frac{1}{(\frac{t}{2}+1)^{q_1}} \int_0^{\frac{t}{2}} \mathrm{d}s\ \frac{\mathrm{e}^{-\frac{s}{8}}}{\sqrt{s}},$$

$$B\Big[{}^{p_1,q_1,r_1}_{p_2,q_2,r_2,r_3}\Big](t) \le \frac{1}{(\frac{t}{2})^{p_1}(\frac{t}{2}+1)^{r_1}} \int_0^{\frac{t}{2}} \frac{\mathrm{d}s}{(1+s)^{q_1}} + \frac{\ln(2+t)^{r_3}}{(\frac{t}{2}+1)^{q_2}} \int_0^{\frac{t}{2}} \frac{\mathrm{d}s}{s^{p_2}(1+s)^{r_2}},$$

and straightforward integrations. ☐

We can now complete the proof of Theorem 6.

Proof of Theorem 6 *continued*. First, we recall that our goal is to prove that the map $\mathcal{N}$ defined by

$$\mathcal{N}[\mathbf{z}](t) = \int_0^t \mathrm{d}s\ \mathrm{e}^{\mathrm{L}(t-s)} \begin{pmatrix} 0 \\ \partial_x h(\mathbf{z}(s)) \end{pmatrix} \tag{6.3}$$

satisfies $\|\mathcal{N}[\mathbf{z}]\| \le C$ for all $\mathbf{z} \in \mathcal{B}$ with $\|\mathbf{z}\| = 1$. The estimate $\|\mathbb{P}\mathrm{D}\mathcal{N}[\mathbf{z}]\|_{2,\frac{3}{4}} \le C$ has already been proved. The other necessary estimates are done as follows:

$$\begin{aligned}
\|\widehat{\mathcal{N}[\mathbf{z}]}\|_{\infty,0} &\le C \sup_{t\ge 0} B_1\Big[{}^{\frac{1}{2},\frac{1}{2}}_{\frac{1}{2},\frac{1}{2}}\Big](t) \le C, \\
\|\mathcal{N}[\mathbf{z}]\|_{2,\frac{1}{4}} &\le C \sup_{t\ge 0} (1+t)^{\frac{1}{4}} B_1\Big[{}^{\frac{1}{2},\frac{3}{4}}_{\frac{1}{2},\frac{3}{4}}\Big](t) \le C, \\
\|\mathbb{P}\mathrm{D}\mathcal{N}[\mathbf{z}]\|_{2,\frac{3}{4}} &\le C \sup_{t\ge 0} (1+t)^{\frac{3}{4}} B_1\Big[{}^{1,\frac{3}{4}}_{\frac{1}{2},\frac{5}{4}}\Big](t) \le C, \\
\|(1-\mathbb{P})\mathrm{D}\mathcal{N}[\mathbf{z}]\|_{2,\frac{3}{4}} &\le \sup_{t\ge 0} (1+t)^{\frac{3}{4}} B_0[\tfrac{5}{4}](t) \le C, \\
\|(1-\mathbb{Q})\mathbb{P}\mathrm{D}^2\mathcal{N}[\mathbf{z}]_2\|_{2,\frac{5}{4}^\star} &\le C \|(1-\mathbb{Q})\mathbb{P}\mathrm{D}\mathcal{N}[\mathbf{z}]_2\|_{2,\frac{3}{4}} \\
&\le C \|\mathbb{P}\mathrm{D}\mathcal{N}[\mathbf{z}]_2\|_{2,\frac{3}{4}} \le C, && (6.4) \\
\|\mathbb{Q}\mathbb{P}\mathrm{D}^2\mathcal{N}[\mathbf{z}]_2\|_{2,\frac{5}{4}^\star} &\le C \sup_{t\ge 1} \frac{(1+t)^{\frac{5}{4}}}{\ln(2+t)}\ B\Big[{}^{\frac{3}{2},\frac{3}{4},0}_{\frac{1}{2},\frac{5}{4},\frac{1}{2},0}\Big](t) \le C, && (6.5) \\
\|(1-\mathbb{P})\mathrm{D}^2\mathcal{N}[\mathbf{z}]_2\|_{2,\frac{5}{4}^\star} &\le \sup_{t\ge 0} (1+t)^{\frac{5}{4}} B_0[\tfrac{5}{4}](t) \le C. && (6.6)
\end{aligned}$$

In (6.4), we used the obvious estimates $\|\mathbb{P}\mathrm{D}f\|_2 \le \|\mathbb{P}f\|_2$ and $\|(1-\mathbb{Q})f\|_{2,p} \le 2^{p-q}\|(1-\mathbb{Q})f\|_{2,q}$ if $q < p$, while in (6.5), we made use of $\sup_{|k|\le 1, t\ge 0}|k|\sqrt{1+t}\mathrm{e}^{-k^2 t} \le$

1, and finally in (6.6) we used $\sup_{k\in\mathbf{R}} |k|(1+k^2)^{-\frac{1}{2}} = 1$. Incidentally, (6.6) is the only place in the above estimates where the (crucial) presence of the extra factor $(1+k^2)^{-\frac{1}{2}}$ in the second component of the r.h.s. of (2.6) is used. This concludes the proof of Theorem 6. $\square$

7. Remainder estimates. We now make precise the sense in which the semigroup $\mathrm{e}^{\mathrm{L}t}$ is *close* to that of (2.3), whose Fourier transform is given by

$$\mathrm{e}^{\mathrm{L}_0 t} \equiv \begin{pmatrix} \mathrm{e}^{-k^2t+ikt} & 0 \\ 0 & \mathrm{e}^{-k^2t-ikt} \end{pmatrix}. \tag{7.1}$$

LEMMA 19. *Let $\mathbb{P}$ be the Fourier multiplier with the characteristic function on $[-1,1]$, and let $\mathrm{e}^{\mathrm{L}t}$ (respectively, $\mathrm{e}^{\mathrm{L}_0 t}$) be as in* (2.2) *(respectively,* (7.1)*) and $\mathcal{S}$ be as in* (2.4)*. Then one has the estimates*

$$\sup_{t\geq 0, k\in\mathbf{R}} \sqrt{1+t}\,\mathrm{e}^{\frac{k^2t}{2}} \left| \left(\mathbb{P}\mathcal{S}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}\right)_{i,j} \right| \leq C, \tag{7.2}$$

where $(\mathbb{P}\mathcal{S}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S})_{i,j}$ denotes the (i,j)-entry in the matrix $\mathbb{P}\mathcal{S}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}$.

Proof. The proof follows by considering separately $|k| \leq 1$ and $|k| > 1$. We first rewrite

$$\mathbb{P}\mathcal{S}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S} = \mathbb{P}\Big(\mathcal{S}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}\Big) + (1-\mathbb{P})\mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}.$$

We then have

$$\sup_{t\geq 0, k\in\mathbf{R}} \sqrt{1+t}\,\mathrm{e}^{\frac{k^2t}{2}} \left| \left((1-\mathbb{P})\mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}\right)_{i,j} \right| \leq \sup_{t\geq 0, |k|\geq 1} \sqrt{1+t}\,\mathrm{e}^{-\frac{k^2t}{2}} \leq C.$$

For $|k| \leq 1$, we first compute

$$\mathrm{e}^{\mathrm{L}_0 t}\mathcal{S} = \mathrm{e}^{-k^2t}\begin{pmatrix} \mathrm{e}^{ikt} & \mathrm{e}^{ikt} \\ \mathrm{e}^{-ikt} & -\mathrm{e}^{-ikt} \end{pmatrix},$$
$$\mathcal{S}\mathrm{e}^{\mathrm{L}t} = \mathrm{e}^{-k^2t}\begin{pmatrix} \cos(kt\Delta) + \frac{1-ik}{\Delta}\, i\sin(kt\Delta) & \cos(kt\Delta) + \frac{1+ik}{\Delta}\, i\sin(kt\Delta) \\ \cos(kt\Delta) - \frac{1+ik}{\Delta}\, i\sin(kt\Delta) & -(\cos(kt\Delta) - \frac{1-ik}{\Delta}\, i\sin(kt\Delta)) \end{pmatrix},$$

where we recall that $\Delta = \sqrt{1-k^2}$. We next note that

$$\begin{aligned}
\mathbb{P}|\sin(kt\Delta) - \sin(kt)| &\leq \mathbb{P}|\cos(kt(\Delta-1)) - 1| + \mathbb{P}|\sin(kt(\Delta-1))| \\
&\leq \mathbb{P}|\sqrt{1-k^2} - 1|\ |k|t \leq \mathbb{P}|k|^3 t, \\
\mathbb{P}|\cos(kt\Delta) - \cos(kt)| &\leq \mathbb{P}|\cos(kt(\Delta-1)) - 1| + \mathbb{P}|\sin(kt(\Delta-1))| \\
&\leq \mathbb{P}|\sqrt{1-k^2} - 1|\ |k|t \leq \mathbb{P}|k|^3 t, \\
\mathbb{P}\left|\left(\frac{1}{\Delta} - 1\right)\sin(kt\Delta)\right| &\leq \mathbb{P}|\sqrt{1-k^2} - 1|\ |k|t \leq \mathbb{P}|k|^3 t.
\end{aligned}$$

The proof is completed noting that

$$\sup_{|k|\leq 1, t\geq 0} t^{\frac{m}{2}} |k|^n \mathrm{e}^{-\frac{k^2t}{2}} \leq C(n)$$

for any (finite) $0 \leq m \leq n$. $\square$

We are now in a position to prove that the remainder

$$\mathcal{R}[\mathbf{z}](t) = \Big(\mathcal{S}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}\Big)\mathbf{z}_0 + \int_0^t \mathrm{d}s \;\left[\; \mathcal{S}\mathrm{e}^{\mathrm{L}(t-s)}\begin{pmatrix} 0 \\ \partial_x h(\mathbf{z}(s)) \end{pmatrix} - \mathrm{e}^{\mathrm{L}_0(t-s)}\mathcal{S}\begin{pmatrix} 0 \\ \partial_x g_0(\mathbf{z}(s)) \end{pmatrix} \right]$$

satisfies improved estimates as stated in (3.5).

THEOREM 20. *Let ϵ_0 be again the (small) constant provided by Theorem 6. Then for all $\mathbf{z}_0 \in \mathcal{B}_0$ with $|\mathbf{z}_0| \le \epsilon_0$, the solution $\mathbf{z}$ of (1.1) satisfies*

$$\|\mathcal{R}[\mathbf{z}]\|_{2,\frac{3}{4}^\star} + \|\mathrm{D}\mathcal{R}[\mathbf{z}]\|_{2,\frac{5}{4}^\star} \le C\epsilon_0. \tag{7.3}$$

Proof. We first note that

$$\Big(\mathcal{S}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}\Big)\mathbf{z}_0 = \Big(\mathcal{S}\mathbb{P}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}\Big)\mathbf{z}_0 + \mathcal{S}(1-\mathbb{P})\mathrm{e}^{\mathrm{L}t}\mathbf{z}_0 \equiv L_1[\mathbf{z}_0](t) + L_2[\mathbf{z}_0](t)$$

and then use the fact that by Lemma 19 we have

$$\|\mathrm{D}^\alpha L_1[\mathbf{z}_0]\|_{2,\frac{3}{4}+\frac{\alpha}{2}} \le C \sup_{t\ge 0}(1+t)^{\frac{1}{4}+\frac{\alpha}{2}} \min\Big(\|\mathrm{D}^\alpha \mathbf{z}_0\|_2,\; t^{-\frac{1}{4}-\frac{\alpha}{2}}\|\widehat{z_0}\|_\infty\Big) \le C|\mathbf{z}_0|$$

for $\alpha = 0, 1$, and finally

$$\|L_2[\mathbf{z}_0]\|_{2,\frac{3}{4}} + \|\mathrm{D}L_2[\mathbf{z}_0]\|_{2,\frac{5}{4}} \le C(\|\mathbf{z}_0\|_2 + \|\mathrm{D}\mathbf{z}_0\|_2) \sup_{t\ge 0}(1+t)^{\frac{5}{4}}\mathrm{e}^{-\frac{t}{4}} \le C|\mathbf{z}_0|.$$

This proves

$$\Big\|\Big(\mathcal{S}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}\Big)\mathbf{z}_0\Big\|_{2,\frac{3}{4}} + \Big\|\mathrm{D}\Big(\mathcal{S}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}\Big)\mathbf{z}_0\Big\|_{2,\frac{5}{4}} \le C|\mathbf{z}_0|$$

for all $\mathbf{z}_0 \in \mathcal{B}_0$. We then show that

$$\Big\|\mathcal{R}[\mathbf{z}](t) - \Big(\mathcal{S}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}\Big)\mathbf{z}_0\Big\|_{2,\frac{3}{4}^\star} + \Big\|\mathrm{D}\Big(\mathcal{R}[\mathbf{z}](t) - \Big(\mathcal{S}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}\Big)\mathbf{z}_0\Big)\Big\|_{2,\frac{5}{4}^\star} \le C\|\mathbf{z}\|^2$$

for all $\mathbf{z} \in \mathcal{B}$. We need only to prove the estimates for $\|\mathbf{z}\| = 1$. We first decompose

$$\mathcal{R}[\mathbf{z}](t) - \Big(\mathcal{S}\mathrm{e}^{\mathrm{L}t} - \mathrm{e}^{\mathrm{L}_0 t}\mathcal{S}\Big)\mathbf{z}_0 = \mathcal{S}\mathcal{N}_1[\mathbf{z}](t) + \mathcal{S}\mathcal{N}_2[\mathbf{z}](t) + \mathcal{N}_3[\mathbf{z}](t), \tag{7.4}$$

where

$$\mathcal{N}_1[\mathbf{z}](t) = (1-\mathbb{P})\int_0^t \mathrm{d}s\; \mathrm{e}^{\mathrm{L}(t-s)}\begin{pmatrix} 0 \\ \partial_x h(\mathbf{z}(s)) \end{pmatrix},$$

$$\mathcal{N}_2[\mathbf{z}](t) = \mathbb{P}\int_0^t \mathrm{d}s\; \mathrm{e}^{\mathrm{L}(t-s)}\begin{pmatrix} 0 \\ \partial_x h(\mathbf{z}(s)) - \partial_x g_0(\mathbf{z}(s)) \end{pmatrix},$$

$$\mathcal{N}_3[\mathbf{z}](t) = \int_0^t \mathrm{d}s\; \Big(\mathbb{P}\mathcal{S}\mathrm{e}^{\mathrm{L}(t-s)} - \mathrm{e}^{\mathrm{L}_0(t-s)}\mathcal{S}\Big)\begin{pmatrix} 0 \\ \partial_x g_0(\mathbf{z}(s)) \end{pmatrix}.$$

We then recall that $h(\mathbf{z})$ satisfies

$$\|h(\mathbf{z})\|_{2,\frac{3}{4}} + \|\mathrm{D}h(\mathbf{z})\|_{2,\frac{5}{4}} \le C\|\mathbf{z}\|^2,$$

which implies that

$$\|\mathcal{N}_1[\mathbf{z}]\|_{2,\frac{3}{4}} \leq C \sup_{t\geq 0}(1+t)^{\frac{3}{4}} B_0[\tfrac{3}{4}](t) \leq C, \quad \|\mathrm{D}\mathcal{N}_1[\mathbf{z}]\|_{2,\frac{5}{4}} \leq C \sup_{t\geq 0}(1+t)^{\frac{5}{4}} B_0[\tfrac{5}{4}](t) \leq C.$$

Moreover $h_0(a,b) \equiv f(a,b)\partial_x b + g(a,b) - g_0(a,b)$ satisfies

$$\|h_0(\mathbf{z})\|_{1,1} + \|\mathrm{D}h_0(\mathbf{z})\|_{1,\frac{3}{2}^\star} \leq C\|\mathbf{z}\|^2.$$

Here we need to consider separately $t \in [0,1]$ and $t \geq 1$ when estimating $\|\mathbb{P}\mathrm{D}\mathcal{N}_2[\mathbf{z}]\|_{2,\frac{5}{4}^\star}$. By writing again $\mathbb{Q}$ for the characteristic function for $t \geq 1$, we find that

$$\begin{aligned}
\|\mathbb{P}\mathcal{N}_2[\mathbf{z}]\|_{2,\frac{3}{4}^\star} &\leq C \sup_{t\geq 0} \frac{(1+t)^{\frac{3}{4}}}{\ln(2+t)} B_1\Big[\begin{smallmatrix}\frac{3}{4},1\\ \frac{3}{4},1\end{smallmatrix}\Big](t) \leq C,\\
\|(1-\mathbb{Q})\mathbb{P}\mathrm{D}\mathcal{N}_2[\mathbf{z}]\|_{2,\frac{5}{4}^\star} &\leq C \sup_{0\leq t\leq 1} (1+t)^{\frac{5}{4}}\ B_1\Big[\begin{smallmatrix}\frac{3}{4},\frac{3}{2}\\ \frac{3}{4},\frac{3}{2}\end{smallmatrix}\Big](t) \leq C,\\
\|\mathbb{Q}\mathbb{P}\mathrm{D}\mathcal{N}_2[\mathbf{z}]\|_{2,\frac{5}{4}^\star} &\leq C \sup_{t\geq 1} \frac{(1+t)^{\frac{5}{4}}}{\ln(2+t)}\ B\Big[\begin{smallmatrix}\frac{5}{4},1,0\\ \frac{3}{4},\frac{3}{2},0,1\end{smallmatrix}\Big](t) \leq C.
\end{aligned}$$

We finally note that

$$\|g_0(\mathbf{z})\|_{2,\frac{3}{4}} + \|\mathrm{D}g_0(\mathbf{z})\|_{2,\frac{5}{4}} \leq C\|\mathbf{z}\|^2,$$

and so, by using Lemma 19, we find

$$\begin{aligned}
\|\mathcal{N}_3[\mathbf{z}]\|_{2,\frac{3}{4}^\star} &\leq \sup_{t\geq 0} \frac{(1+t)^{\frac{3}{4}}}{\ln(2+t)}\ B\Big[\begin{smallmatrix}\frac{1}{2},\frac{3}{4},\frac{1}{2}\\ \frac{1}{2},\frac{3}{4},\frac{1}{2},0\end{smallmatrix}\Big](t) \leq C,\\
\|\mathrm{D}\mathcal{N}_3[\mathbf{z}]\|_{2,\frac{5}{4}^\star} &\leq \sup_{t\geq 0} \frac{(1+t)^{\frac{5}{4}}}{\ln(2+t)}\ B\Big[\begin{smallmatrix}1,\frac{3}{4},\frac{1}{2}\\ \frac{1}{2},\frac{5}{4},\frac{1}{2},0\end{smallmatrix}\Big](t) \leq C.
\end{aligned}$$

This completes the proof. ∎

It now remains only to prove the estimates (3.13) and (3.14) on the maps $\widetilde{\mathcal{R}}_{\{u,v\}}$, where we recall that

$$\begin{aligned}
\widetilde{\mathcal{R}}_u[\mathbf{z},\mathbf{R}^N](t) &= c_+\mathrm{E}_0[h_{1,u}+h_{3,u}](t) - c_-\mathrm{E}_{-2}[h_{1,v}+h_{3,v}](t) + c_3\mathrm{E}_{-1}[h_2+h_4](t),\\
\widetilde{\mathcal{R}}_v[\mathbf{z},\mathbf{R}^N](t) &= c_-\mathrm{E}_0[h_{1,v}+h_{3,v}](t) - c_+\mathrm{E}_2[h_{1,u}+h_{3,u}](t) - c_3\mathrm{E}_1[h_2+h_4](t),
\end{aligned}$$

with $h_{1,u} = R_u^N(u+u_\star)$, $h_{3,u} = u_1^2$, $h_{1,v} = R_v^N(v+v_\star)$, $h_{3,v} = v_1^2$, $h_4 = (\mathcal{T}u_\star)(\mathcal{T}^{-1}v_\star)$, and

$$\begin{aligned}
h_2 &= (\mathcal{T}R_u^N)\mathcal{T}^{-1}\Big(\frac{v+v_\star}{2}\Big) + (\mathcal{T}^{-1}R_v^N)\mathcal{T}\Big(\frac{u+u_\star}{2}\Big),\\
\mathrm{E}_\sigma[h](t) &= \partial_x \int_0^t \mathrm{d}s\ \mathrm{e}^{\partial_x^2(t-s)}\ \mathcal{T}^\sigma h(s).
\end{aligned}$$

Here we will prove only that

$$\sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha \widetilde{\mathcal{R}}_{\{u,v\}}[\mathbf{z},\mathbf{R}^N]\|_{2,\frac{3}{4}+\frac{\alpha}{2}-\epsilon} \leq C\epsilon_0 \sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha \mathbf{R}^N\|_{2,\frac{3}{4}+\frac{\alpha}{2}-\epsilon} + C. \tag{7.5}$$

It is then straightforward to show (3.14), namely, that the maps $\widetilde{\mathcal{R}}_{\{u,v\}}$ are Lipschitz in their second argument; we omit the details.

To prove (7.5), we first need estimates on $\mathbf{h}_1 = (h_{1,u}, h_{1,v})$, h_2, $\mathbf{h}_3 = (h_{3,u}, h_{3,v})$, and h_4. We note that $\mathbf{u}_0 = (u_0, v_0)$ and $\mathbf{u}_1 = (u_1, v_1)$ satisfy

$$\|\mathbf{u}_0\|_{1,0} + \|\mathbf{u}_1\|_{1,0} + \|\mathrm{D}\mathbf{u}_0\|_{1,\frac{1}{2}} + \|\mathrm{D}\mathbf{u}_1\|_{1,\frac{1}{2}} \le C,$$
$$\sup_{t\ge 0}(1+t)^{\frac{3}{2}}\Big(|\mathbf{u}_0(\pm t,t)| + |\mathbf{u}_1(\pm t,t)|\Big) + (1+t)^2\Big(|\mathrm{D}\mathbf{u}_0(\pm t,t)| + |\mathrm{D}\mathbf{u}_1(\pm t,t)|\Big) \le C$$

for some constant C; see Proposition 12. We thus find that

$$\begin{aligned}
\|\mathbf{h}_1\|_{1,1-\epsilon} + \|\mathrm{D}\mathbf{h}_1\|_{1,\frac{3}{2}-\epsilon} &\le C\epsilon_0 \sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha \mathbf{R}^N\|_{2,\frac{3}{4}+\frac{\alpha}{2}-\epsilon},\\
\|h_2\|_{1,1-\epsilon} + \|\mathrm{D}h_2\|_{1,\frac{3}{2}-\epsilon} &\le C\epsilon_0 \sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha \mathbf{R}^N\|_{2,\frac{3}{4}+\frac{\alpha}{2}-\epsilon},\\
\|\mathbf{h}_3\|_{1,1} + \|\mathrm{D}\mathbf{h}_3\|_{1,\frac{3}{2}} + \|h_4\|_{1,\frac{3}{2}} + \|\mathrm{D}h_4\|_{2,2} &\le C.
\end{aligned}\tag{7.6}$$

The proof of (7.5) then follows from Proposition 21, which implies that

$$\sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha \mathrm{E}_\sigma[\mathbf{h}_1]\|_{2,\frac{3}{4}+\frac{\alpha}{2}-\epsilon} + \|\mathrm{D}^\alpha \mathrm{E}_\sigma[h_2]\|_{2,\frac{3}{4}+\frac{\alpha}{2}-\epsilon} \le C\epsilon_0 \sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha \mathbf{R}^N\|_{2,\frac{3}{4}+\frac{\alpha}{2}-\epsilon},$$
$$\sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha \mathrm{E}_\sigma[\mathbf{h}_3]\|_{2,\frac{3}{4}+\frac{\alpha}{2}^\star} + \|\mathrm{D}^\alpha \mathrm{E}_\sigma[h_4]\|_{2,\frac{3}{4}+\frac{\alpha}{2}^\star} \le C$$

for any $\sigma \in \{-2,-1,0,1,2\}$ if the estimates in (7.6) are satisfied.

PROPOSITION 21. *Let $\epsilon > 0$ and $\sigma \in \{-2,-1,0,1,2\}$. Then there holds*

$$\sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha \mathrm{E}_\sigma[h_1]\|_{2,\frac{3}{4}+\frac{\alpha}{2}-\epsilon} \le C \sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha h_1\|_{1,1+\frac{\alpha}{2}-\epsilon},$$
$$\sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha \mathrm{E}_\sigma[h_2]\|_{2,\frac{3}{4}+\frac{\alpha}{2}^\star} \le C \sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha h_2\|_{1,1+\frac{\alpha}{2}}.$$

Proof. Let $u_i = \mathrm{E}_\sigma[h_i]$. By taking the Fourier transform, we find that

$$\widehat{u_i}(k,t) = ik\int_0^t \mathrm{d}s\, \mathrm{e}^{-k^2(t-s)+i\sigma k s}\widehat{h_i}(k,s).$$

We can restrict ourselves to $\sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha h_1\|_{1,1+\frac{\alpha}{2}-\epsilon} = 1$ and $\sum_{\alpha=0}^{1} \|\mathrm{D}^\alpha h_2\|_{1,1+\frac{\alpha}{2}} = 1$. Then it follows that

$$\|\mathrm{D}^\alpha u_1\|_{2,\frac{3}{4}+\frac{\alpha}{2}-\epsilon} \le C\sup_{t\ge 0}(1+t)^{\frac{3}{4}+\frac{\alpha}{2}-\epsilon}\, B_1\Big[\begin{smallmatrix}\frac{3}{4}+\frac{\alpha}{2},1-\epsilon\\ \frac{3}{4},1+\frac{\alpha}{2}-\epsilon\end{smallmatrix}\Big](t) \le C,$$
$$\|\mathrm{D}^\alpha u_2\|_{2,\frac{3}{4}+\frac{\alpha}{2}^\star} \le C\sup_{t\ge 0}\frac{(1+t)^{\frac{3}{4}+\frac{\alpha}{2}}}{\ln(2+t)}\, B_1\Big[\begin{smallmatrix}\frac{3}{4}+\frac{\alpha}{2},1\\ \frac{3}{4},1+\frac{\alpha}{2}\end{smallmatrix}\Big](t) \le C$$

for $\alpha = 0, 1$ as claimed. $\square$

REFERENCES

[1] J. Bricmont, A. Kupiainen, and G. Lin, *Renormalization group and asymptotics of solutions of nonlinear parabolic equations*, Comm. Pure Appl. Math., 47 (1994), pp. 893–922.

[2] I.-L. Chern and T.-P. Liu, *Convergence to diffusion waves of solutions for viscous conservation laws*, Comm. Math. Phys., 110 (1987), pp. 503–517, and *Erratum: "Convergence to diffusion waves of solutions for viscous conservation laws"* Comm. Math. Phys., 120 (1989), pp. 525–527.

[3] T. Gallay and G. Raugel, *Scaling variables and asymptotic expansions in damped wave equations*, J. Differential Equations, 150 (1998), pp. 42–97.

[4] S. Kawashima, *Large-time behaviour of solutions to hyperbolic-parabolic systems of conservation laws and applications*, Proc. Roy. Soc. Edinburgh Sect. A, 106 (1987), pp. 169–194.

[5] T.-P. Liu, *Hyperbolic and Viscous Conservation Laws*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 72, SIAM, Philadelphia, PA, 2000.

[6] T.-P. Liu and Y. Zeng, *Large time behavior of solutions for general quasilinear hyperbolic-parabolic systems of conservation laws*, Mem. Amer. Math. Soc., 125 (1997), pp. viii–120.

[7] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci. 68, 2nd ed., Springer-Verlag, New York, 1997.

[8] G. van Baalen, *Downstream asymptotics in exterior domains: From stationary wakes to time periodic flows*, J. Math. Fluid Mech., 9 (2007), pp. 295–342.

[9] C. E. Wayne, *Invariant manifolds for parabolic partial differential equations on unbounded domains*, Arch. Ration. Mech. Anal., 138 (1997), pp. 279–306.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 1978–1995

LARGE TIME BEHAVIOR OF RADIALLY SYMMETRIC SURFACES IN THE MEAN CURVATURE FLOW*

MITSUNORI NARA†

Abstract. The large time behavior of radially symmetric surfaces in $\mathbb{R}^{n+1}$ moving by the mean curvature flow is studied. By studying a Cauchy problem, we deal with moving surfaces represented by entire graphs on a hyperplane. Here an initial surface is given by a function that is bounded and radially symmetric. It is proved that the solution converges uniformly to the solution of the Cauchy problem of the heat equation with the same initial value. The difference is of order $O(t^{-1/2})$ as time goes to infinity. The proof is based on the construction of the Green's function and the decay estimates for the derivatives of the solution. By virtue of the stability results for the heat equation, our result gives the sufficient and necessary conditions on the asymptotic stability of constant functions that represent stationary hyperplanes in the mean curvature flow.

Key words. large time behavior, mean curvature flow, heat equation, radially symmetric solution

AMS subject classifications. 35K15, 35B40, 53C44

DOI. 10.1137/070686330

1. Introduction.

In this paper, we study *the mean curvature flow* in $\mathbb{R}^{n+1}$. Especially we study the large time behavior of the solution to the Cauchy problem of a scalar parabolic equation

$$\frac{u_t}{\sqrt{1+|\nabla u|^2}} = \operatorname{div}\left(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\right), \quad x \in \mathbb{R}^n,\ t > 0, \tag{1}$$

$$u(x,0) = u_0(x), \qquad x \in \mathbb{R}^n. \tag{2}$$

Here we assume that $n \geq 2$ and that the initial value $u_0(x)$ is bounded and radially symmetric. Throughout this paper, we say that a function $f(x)$ is *radially symmetric* when it depends only on $r = |x|$.

The mean curvature flow is a mathematical model that describes motion of oriented hypersurfaces. Let $S(t)$ be a moving hypersurface in $\mathbb{R}^{n+1}$. We are interested in the dynamics of $S(t)$ given by $V = -\kappa$, where V and κ are the outward normal velocity and the mean curvature of $S(t)$, respectively. This model appears in several fields. For example, this describes the motion of phase boundaries in the Allen–Cahn equation [2] and also in the reaction-diffusion systems of a competition type [7].

The mean curvature flow of compact surfaces is extensively studied. It is well known that a compact surface $S(t)$ in $\mathbb{R}^{n+1}$ shrinks to a single point when $n = 1$ or $S(0)$ is convex. This fact has been proved by Gage and Hamilton [11], Grayson [12], and Huisken [13]. When $n \geq 2$, smooth initial surfaces may develop geometric singularities. Evans and Spruck [9] and Chen, Giga, and Goto [1] developed a level set approach for handling such geometric singularities.

On the other hand, some researchers focus on the mean curvature flow of *graphical surfaces*. In this case, an initial surface $S(0)$ is given by a scalar function $y = u_0(x), x \in$

*Received by the editors March 26, 2007; accepted for publication (in revised form) September 25, 2007; published electronically March 18, 2008.
http://www.siam.org/journals/sima/39-6/68633.html
†Graduate School of Mathematical Sciences, The University of Tokyo, Komaba 3-8-1, Meguro-ku, Tokyo, 153-8914, Japan (nara@ms.u-tokyo.ac.jp).

$\mathbb{R}^n$, and a moving surface $S(t), t > 0$, is expressed by a scalar function $y = u(x,t), x \in \mathbb{R}^n, t > 0$. Under these assumptions, the mean curvature flow $V = -\kappa$ is rewritten as the Cauchy problem (1)–(2).

A pioneering work for the mean curvature flow of graphical surfaces is given by Ecker and Huisken [6]. They showed the existence of expanding self-similar solutions and the large time behavior of solutions. It is proved that a graphical solution $u(x,t)$ of (1)–(2) converges to an expanding self-similar solution, if an unbounded initial graph $u_0(x)$ satisfies linear growth and additional assumptions. Ishimura [14] also studied the same problem in detail.

Our aim in this paper is to study the large time behavior of graphical surfaces with bounded initial graphs, i.e., the large time behavior of the solution $u(x,t)$ to the Cauchy problem (1)–(2) with bounded initial values $u_0(x)$. The major interest is the asymptotic stability of the hyperplane $u(x,t) = 0$, which is the simplest stationary surface in the mean curvature flow.

Now we briefly mention the stability results for the heat equation. For the Cauchy problem of the heat equation

$$H_t = \Delta H, \qquad x \in \mathbb{R}^n,\ t > 0, \tag{3}$$

$$H(x,0) = \Phi(x), \quad x \in \mathbb{R}^n, \tag{4}$$

the necessary and sufficient conditions for the asymptotic stability of $H(x,t) = 0$ are obtained as follows. This result is shown in [5, 8, 15, 18], for example.

Proposition 1.1. *Let $\Phi(x)$ be a bounded function. Then the solution $H(x,t)$ of (3)–(4) satisfies $\lim_{t\to\infty} H(0,t) = 0$ if and only if $\Phi(x)$ satisfies*

$$\lim_{R\to\infty} \frac{1}{(2R)^n} \int_{y\in V_R} \Phi(y)\, dy = 0.$$

Moreover it satisfies $\lim_{t\to\infty} \sup_{x\in\mathbb{R}^n} |H(x,t)| = 0$ if and only if $\Phi(x)$ satisfies

$$\lim_{R\to\infty} \sup_{x\in\mathbb{R}} \frac{1}{(2R)^n} \left| \int_{y\in V_R} \Phi(y-x)\, dy \right| = 0,$$

where V_R is the cube with the center at the origin and the side $2R$.

Here we note that an initial value $\Phi(x)$ does not need to decay to zero as $|x| \to \infty$. It may oscillate with some amplitude. Collet and Eckmann [4] showed a remarkable example where an initial value does not satisfy the criterion for the asymptotic stability of $H(x,t) = 0$ and the solution $H(x,t)$ does not converge to any fixed constant as $t \to \infty$.

Proposition 1.2 (see [4]). *Let $H(x,t)$ be the solution of the Cauchy problem of the heat equation*

$$\begin{aligned} &H_t = H_{xx}, && x \in \mathbb{R},\ t > 0, \\ &H(x,0) = \Phi^*(x), && x \in \mathbb{R}. \end{aligned}$$

Suppose that $\Phi^*(x)$ is a smooth even function that satisfies $|\Phi^*(x)| \le 1$ for $x \in \mathbb{R}$ and

$$\Phi^*(x) = (-1)^m, \quad x \in [\, m! + 2^m,\ (m+1)! - 2^{m+1}\,]$$

for $m \ge 5$. Then it holds that

$$\liminf_{t\to\infty} H(0,t) = -1, \quad \limsup_{t\to\infty} H(0,t) = 1.$$

This implies that the solution $H(x,t)$ oscillates forever as $t \to \infty$ with an amplitude of 2. Here the initial value $\Phi^*(x)$ does not decay but oscillates slower and slower as $|x| \to \infty$. This example shows the difficulty in considering the asymptotic stability of the zero solution for spatially nondecaying initial values.

In the case of $n = 1$, the asymptotic stability of $u(x,t) = 0$ of the problem (1)–(2) with bounded initial values is obtained in [17]. In this case, the problem (1)–(2) is rewritten as follows:

$$u_t = \frac{u_{xx}}{1+u_x^2}, \qquad x \in \mathbb{R},\ t > 0, \tag{5}$$

$$u(x,0) = u_0(x), \quad x \in \mathbb{R}. \tag{6}$$

This describes the motion of curves in $\mathbb{R}^2$ expressed by the function $y = u(x,t), x \in \mathbb{R}, t > 0$. The constant solution $u(x,t) = 0$ means the stationary line in $\mathbb{R}^2$. Equation (5) is rewritten as $u_t = (\arctan u_x)_x = u_{xx} - (u_x - \arctan u_x)_x = u_{xx} + F_x(x,t)$, which is considered to be the heat equation with the flux $F_x(x,t)$. Then $u(x,t)$ is given by

$$u(x,t) = \int_{\mathbb{R}} \Gamma(x-y,t)\, u_0(y)\, dy + \int_0^t \int_{\mathbb{R}} \Gamma(x-y,t-s)\, F_y(y,s)\, dy\, ds,$$

where $\Gamma(\xi,\tau)$ is the heat kernel. In [17], under the assumption of $u_0 \in C^{2+\alpha}(\mathbb{R})$, it is proved that

$$\left| \int_0^t \int_{\mathbb{R}} \Gamma(x-y,t-s)\, F_y(y,s)\, dy\, ds \right| \le C\, t^{-\frac{1}{2}}, \quad t > 0.$$

This means that the contribution of the flux F_x can be ignored as $t \to \infty$, and the large time behavior of $u(x,t)$ of (5)–(6) is derived directly from that of the Cauchy problem of the heat equation with the same initial value $u_0(x)$. Here the initial value $u_0(x)$ does not need to decay to zero as $|x| \to \infty$. Consequently, the asymptotic stability of the stationary line $u(x,t) = 0$ of (5)–(6) is obtained as in Proposition 1.1. In addition, a similar result to Proposition 1.2 holds true for the problem (5)–(6). We briefly mention why such a result holds true. One reason is that the mean value of $F_x(x,t)$ is zero at each $t \ge 0$ when $|u_x|$ is bounded. Thus Proposition 1.1 gives that, for each $s \ge 0$,

$$\lim_{t\to\infty} \int_{\mathbb{R}} \Gamma(x-y,t-s)\, F_y(y,s)\, dy = 0.$$

This means that the contribution of the flux $F_y(y,s)$ at each $s \ge 0$ may be ignored as $t \to \infty$. Another reason is that the flux $F_y(y,s)$ itself decays fast enough as $s \to \infty$. Then the whole contribution of the flux may be neglected in considering the large time behavior.

Our motivation in this paper is to answer the question of whether similar results hold true for the problem (1)–(2) with $n \ge 2$. The following is the main result in this paper.

THEOREM 1.3. *Suppose that $n \ge 2$ and that $u_0 \in C(\mathbb{R}^n) \cap W^{1,\infty}(\mathbb{R}^n)$ is radially symmetric. Then there exists a classical solution $u(x,t)$ to the Cauchy problem* (1)–(2) *up to $t = +\infty$. Moreover it satisfies*

$$\sup_{x\in\mathbb{R}^n} \left| u(x,t) - \int_{\mathbb{R}^n} \Gamma(x-y,t)\, u_0(y)\, dy \right| \le C\, t^{-\frac{1}{2}}, \quad t \ge 2,$$

for a constant $C > 0$ depending only on n and u_0. Here $\Gamma(\xi, \tau)$ is the heat kernel given by $\Gamma(\xi, \tau) = 1/(4\pi\tau)^{n/2} \exp(-|\xi|^2/(4\tau))$.

Thus the large time behavior of the solution of (1)–(2) with $n \geq 2$ is also derived directly from that of the Cauchy problem of the heat equation with the same initial value $u_0(x)$ if it is radially symmetric. The difference is of order $O(t^{-1/2})$. We also note that the initial value $u_0(x)$ does not need to decay to zero as $|x| \to \infty$ and that it does not need to converge to any fixed constant as $|x| \to \infty$. It may oscillate with some amplitude.

From Proposition 1.1 and Theorem 1.3, we obtain the necessary and sufficient conditions for the asymptotic stability of the hyperplane $u(x, t) = 0$ of (1)–(2) with radially symmetric initial values as follows.

COROLLARY 1.4. *Suppose that $n \geq 2$ and that $u_0 \in C(\mathbb{R}^n) \cap W^{1,\infty}(\mathbb{R}^n)$ is radially symmetric. Then the solution $u(x, t)$ to the Cauchy problem* (1)–(2) *satisfies* $\lim_{t\to\infty} u(0, t) = 0$ *if and only if $u_0(x)$ satisfies*

$$\lim_{R\to\infty} \frac{1}{(2R)^n} \int_{y\in V_R} u_0(y)\, dy = 0. \tag{7}$$

Moreover it satisfies $\lim_{t\to\infty} \sup_{x\in\mathbb{R}^n} |u(x, t)| = 0$ *if and only if $u_0(x)$ satisfies*

$$\lim_{R\to\infty} \sup_{x\in\mathbb{R}} \frac{1}{(2R)^n} \left| \int_{y\in V_R} u_0(y - x)\, dy \right| = 0, \tag{8}$$

where V_R is the cube with the center at the origin and the side $2R$.

To prove Theorem 1.3, we will extend the idea given by [17] to the case where an initial value is radially symmetric and $n \geq 2$. If the initial value is radially symmetric, the solution of (1)–(2) is also radially symmetric and satisfies

$$u_t = \frac{u_{rr}}{1 + u_r^2} + \frac{n-1}{r} u_r = u_{rr} + \frac{n-1}{r} u_r + F_r(r, t), \quad r > 0,\ t > 0,$$

where $r = |x|$ and $F_r(r, t) = -(u_r - \arctan u_r)_r$. An extra difficulty in most of the radial problems is to handle a singularity appearing at the origin properly. Our idea in this paper is to construct the Green's function $Z^{(n)}$ of the equation $u_t = u_{rr} + (n - 1)u_r/r$, which is the heat equation for radially symmetric solutions. In section 3, we give an explicit expression for the Green's function $Z^{(n)}$ and its estimates. Once we obtain the Green's function $Z^{(n)}$, we will have the expression

$$u(r, t) = \int_0^\infty Z^{(n)}(r, \xi, t)\, u_0(\xi)\, d\xi + \int_0^t \int_0^\infty Z^{(n)}(r, \xi, t - s)\, F_\xi(\xi, s)\, d\xi\, ds. \tag{9}$$

The first term of the right-hand side is the radially symmetric solution of the heat equation with the same initial value. The remaining problem is to estimate $F_r(r, t)$. For this purpose, we show the decay estimates for the derivatives of $u(r, t)$ in section 4. In section 5, we give the proof of Theorem 1.3. We estimate the second term of the right-hand side of (9) by making use of the estimates for the Green's function $Z^{(n)}$ and the derivatives of $u(r, t)$ obtained in sections 3 and 4, respectively.

Here we introduce the notation. $L^\infty(\mathbb{R}^n)$ and $W^{1,\infty}(\mathbb{R}^n)$ denote the Lebesgue and the Sobolev spaces, respectively. For $\alpha \in (0, 1)$, $C^\alpha(\mathbb{R}^n)$ denotes the Hölder space, that is, the space of functions that are bounded and uniformly Hölder continuous with exponent α on $\mathbb{R}^n$. $C^{2+\alpha}(\mathbb{R}^n)$ means the space of functions with $u, u_{x_i}, u_{x_i x_j} \in$

$C^\alpha(\mathbb{R}^n)$. For a domain $R_T = \mathbb{R}^n \times [0,T]$, $C^{\alpha,\alpha/2}(R_T)$ denotes the space of functions that are bounded and uniformly Hölder continuous with exponent α and $\alpha/2$ with respect to x_i and t, respectively, on R_T. $C^{2+\alpha,1+\alpha/2}(R_T)$ means the space of functions that satisfy $u, u_{x_i}, u_{x_i x_j}, u_t \in C^{\alpha,\alpha/2}(R_T)$. The norm in the space $C^{2+\alpha,1+\alpha/2}(R_T)$ is denoted by $\|\cdot\|_{C^{2+\alpha,1+\alpha/2}(R_T)}$.

2. Preliminaries. Before proving our results for large time behavior, we briefly mention the existence and some estimates of the solutions $u(x,t)$ to the problem (1)–(2). Throughout this paper, we always assume that $n \geq 2$ and that an initial value belongs to $C(\mathbb{R}^n) \cap W^{1,\infty}(\mathbb{R}^n)$ even if it is not mentioned specifically. First we introduce the existence result given by Ecker and Huisken [6].

PROPOSITION 2.1 (see [6]). *Suppose that $u_0(x)$ is uniform Lipschitz continuous on $\mathbb{R}^n$. Then there exists a solution $u(x,t)$ of (1)–(2) globally in time. Moreover it satisfies*

$$|\nabla u(x,t)| \leq \|\nabla u_0\|_{L^\infty(\mathbb{R}^n)}, \quad x \in \mathbb{R}^n,\ t > 0. \tag{10}$$

They assumed only uniform Lipschitz continuity of initial surfaces to study the large time behavior of solutions with linear growth. In the present paper, we focus on the large time behavior of bounded solutions and assume that $u_0 \in C(\mathbb{R}^n) \cap W^{1,\infty}(\mathbb{R}^n)$. In this case, the existence of the solution follows directly from Proposition 2.1. In addition, the following estimate holds true.

PROPOSITION 2.2. *Suppose that $u_0(x)$ belongs to $C(\mathbb{R}^n) \cap W^{1,\infty}(\mathbb{R}^n)$. Then there exists a solution $u(x,t)$ of (1)–(2) globally in time. Moreover it belongs to $C^{2+\alpha,1+\alpha/2}(\mathbb{R}^n \times [1,T])$ for any $T > 0$ and satisfies the estimate*

$$\|u\|_{C^{2+\alpha,1+\alpha/2}(\mathbb{R}^n \times [1,T])} \leq C_1, \tag{11}$$

where $C_1 > 0$ is a constant that depends only on n and u_0 and is independent of $T > 0$.

Proof. Proposition 2.1 implies that the solution exists globally in time, and the estimate (10) holds true. The estimate (10) implies that (1) is considered to be a uniformly parabolic equation, and hence the comparison principle is valid. Thus the L^∞-bound for u is obtained, since the constants $\pm\|u_0\|_{L^\infty(\mathbb{R}^n)}$ are both stationary solutions of (1)–(2). The L^∞-bound for u and the estimate (10) imply that

$$\|u_{x_i}\|_{C^{\alpha,\alpha/2}(\mathbb{R}^n \times [1,T])} \leq C \quad \text{for } 1 \leq i \leq n,$$

where $C > 0$ is a constant that depends only on n and u_0 and is independent of $T > 0$. This fact follows from the general theory for quasi-linear parabolic equations. See Chapter XII of [16], for instance. Finally we apply the Schauder theory of linear parabolic equations and hence obtain (11). Thus the proof is completed. ∎

All discussions in this paper will construct on the basis of Propositions 2.1 and 2.2. The existence of solutions to the problem (1)–(2) is obtained also in other papers. For example, Chou and Kwong [3] proved the existence of solutions for smooth initial values without the restriction of growth order.

3. Green's function for radially symmetric solutions. As is mentioned in the introduction, our idea in this paper is to express radially symmetric solutions of (1)–(2) by using the Green's function of the heat equation for radially symmetric solutions. For this purpose, we begin with the Cauchy problem of the heat equation

$$\begin{aligned} &H_t = \Delta H, && x \in \mathbb{R}^n,\ t > 0, \\ &H(x,0) = \Phi(x), && x \in \mathbb{R}^n. \end{aligned}$$

It is well known that the solution $H(x,t)$ is expressed by the heat kernel as follows:

$$H(x,t) = \int_{\mathbb{R}^n} \frac{1}{(4\pi\, t)^{n/2}} \exp\left(-\frac{|x-y|^2}{4t}\right) \Phi(y)\, dy. \tag{12}$$

When the initial value $\Phi(x)$ is radially symmetric, the solution $H(x,t)$ is also radially symmetric. Here we denote $H(x,t)$ and $\Phi(x)$ by $h(r,t)$ and $\varphi(r)$ for $r = |x|$, respectively. Then we have the initial value problem of the heat equation for radially symmetric solutions, that is,

$$h_t = h_{rr} + \frac{n-1}{r} h_r, \quad r > 0,\ t > 0, \tag{13}$$

$$h(r,0) = \varphi(r), \qquad r > 0. \tag{14}$$

By changing the variables in (12) to the polar coordinate system, we obtain an explicit expression for $h(r,t)$, which is useful to our discussions. We give the following proposition.

PROPOSITION 3.1. *The solution $h(r,t)$ of* (13)–(14) *is given by*

$$h(r,t) = \int_0^\infty Z^{(n)}(r,\xi,t)\, \varphi(\xi)\, d\xi, \tag{15}$$

where $Z^{(n)}(r,\xi,t)$ is the function defined by

$$Z^{(n)}(r,\xi,t) = \frac{K_n\, \xi^{n-1}}{(4\pi\, t)^{n/2}} \int_0^\pi \exp\left(-\frac{(r-\xi\cos\theta)^2 + (\xi\sin\theta)^2}{4t}\right) \sin^{n-2}\theta\, d\theta. \tag{16}$$

The constant K_n is given by $K_2 = 2$, $K_3 = 2\pi$, and

$$K_n = \frac{2(2\pi)^{m+1}}{1\cdot 3\cdot 5 \ldots (2m+1)} \qquad \text{if} \quad n = 2m+4 \geq 4, \tag{17}$$

$$K_n = \frac{(2\pi)^{m+2}}{2\cdot 4\cdot 6 \ldots 2(m+1)} \qquad \text{if} \quad n = 2m+5 \geq 5. \tag{18}$$

Proof. First we consider the case of $n = 2$. In this case, we use the polar coordinate system with $x_1 = r\cos\Theta, x_2 = r\sin\Theta$ and denote $H(x,t)$ and $\Phi(x)$ by $\widetilde{H}(r,\Theta,t)$ and $\widetilde{\Phi}(r,\Theta)$, respectively. Then by changing the variables in (12), we have

$$\widetilde{H}(r,\Theta,t) = \int_0^\infty \int_0^{2\pi} \frac{1}{4\pi\, t} \exp\left(-\frac{d(r,\Theta,\xi,\theta)}{4t}\right) \widetilde{\Phi}(\xi,\theta)\, \xi\, d\theta\, d\xi, \tag{19}$$

where $d(r,\Theta,\xi,\theta) = (r\cos\Theta - \xi\cos\theta)^2 + (r\sin\Theta - \xi\sin\theta)^2$. If the initial value $\widetilde{\Phi}(r,\Theta)$ is radially symmetric, the solution $\widetilde{H}(r,\Theta,t)$ is also radially symmetric; that is, $\widetilde{H}(r,\Theta,t)$ is independent of Θ. Thus the solution $h(r,t)$ of (13)–(14) is given by $h(r,t) = \widetilde{H}(r,0,t)$, for example. By setting $\Theta = 0$ in (19), we obtain

$$\begin{aligned} h(r,t) = \widetilde{H}(r,0,t) &= \int_0^\infty \int_0^{2\pi} \frac{1}{4\pi\, t} \exp\left(-\frac{(r-\xi\cos\theta)^2 + (\xi\sin\theta)^2}{4t}\right) \varphi(\xi)\, \xi\, d\theta\, d\xi \\ &= \int_0^\infty \frac{\xi}{2\pi\, t} \int_0^\pi \exp\left(-\frac{(r-\xi\cos\theta)^2 + (\xi\sin\theta)^2}{4t}\right) d\theta\ \varphi(\xi)\, d\xi. \end{aligned}$$

This implies (15)–(16) with $K_2 = 2$. Thus the proof for the case $n = 2$ is completed.

Next we consider the case of $n \geq 3$. In this case, we use the polar coordinate system with

$$\begin{aligned} x_1 &= r\cos\Theta_1, \\ x_2 &= r\sin\Theta_1\cos\Theta_2, \\ x_3 &= r\sin\Theta_1\sin\Theta_2\cos\Theta_3, \\ &\cdots \\ x_{n-1} &= r\sin\Theta_1\sin\Theta_2\ldots\sin\Theta_{n-2}\cos\Theta_{n-1}, \\ x_n &= r\sin\Theta_1\sin\Theta_2\ldots\sin\Theta_{n-2}\sin\Theta_{n-1}, \end{aligned}$$

where

$$(r, \Theta_1, \ldots, \Theta_{n-2}, \Theta_{n-1}) \in [0,\infty) \times [0,\pi] \times \cdots \times [0,\pi] \times [0,2\pi].$$

We denote $H(x,t)$ and $\Phi(x)$ by $\widetilde{H}(r,\mathbf{\Theta},t)$ and $\widetilde{\Phi}(r,\mathbf{\Theta})$ with $\mathbf{\Theta} = (\Theta_1,\ldots,\Theta_{n-1})$, respectively. Then by changing the variables in (12), we have

$$\begin{aligned} \widetilde{H}(r,\mathbf{\Theta},t) = \int_0^\infty \int_0^{2\pi} \int_0^\pi \cdots \int_0^\pi & \frac{1}{(4\pi\, t)^{n/2}} \exp\left(-\frac{d(r,\mathbf{\Theta},\xi,\boldsymbol{\theta})}{4t}\right) \widetilde{\Phi}(\xi,\boldsymbol{\theta}) \\ & \times \xi^{n-1} \sin^{n-2}\theta_1 \,\ldots\, \sin\theta_{n-2}\;\; d\theta_1 \,\ldots\, d\theta_{n-1}\, d\xi, \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_{n-1})$ and the function $d(r,\mathbf{\Theta},\xi,\boldsymbol{\theta})$ is given by

$$\begin{aligned} d(r,\mathbf{\Theta},\xi,\boldsymbol{\theta}) = {} & (r\cos\Theta_1 - \xi\cos\theta_1)^2 \\ & + (r\sin\Theta_1\cos\Theta_2 - \xi\sin\theta_1\cos\theta_2)^2 \\ & + \cdots \\ & + (r\sin\Theta_1\ldots\sin\Theta_{n-2}\cos\Theta_{n-1} - \xi\sin\theta_1\ldots\sin\theta_{n-2}\cos\theta_{n-1})^2 \\ & + (r\sin\Theta_1\ldots\sin\Theta_{n-2}\sin\Theta_{n-1} - \xi\sin\theta_1\ldots\sin\theta_{n-2}\sin\theta_{n-1})^2. \end{aligned}$$

Similarly to the case of $n = 2$, the solution $h(r,t)$ of (13)–(14) is given by $h(r,t) = \widetilde{H}(r,\mathbf{0},t)$, with $\mathbf{0} = (0,\ldots,0)$. By setting $\mathbf{\Theta} = \mathbf{0}$, we have

$$\begin{aligned} d(r,\mathbf{0},\xi,\boldsymbol{\theta}) &= (r - \xi\cos\theta_1)^2 + (\xi\sin\theta_1\cos\theta_2)^2 \\ & \quad + \cdots \\ & \quad + (\xi\sin\theta_1\ldots\sin\theta_{n-2}\cos\theta_{n-1})^2 + (\xi\sin\theta_1\ldots\sin\theta_{n-2}\sin\theta_{n-1})^2 \\ &= (r - \xi\cos\theta_1)^2 + (\xi\sin\theta_1)^2. \end{aligned}$$

Thus $h(r,t)$ is given as follows:

$$\begin{aligned} h(r,t) &= \int_0^\infty \int_0^\pi \frac{1}{(4\pi\, t)^{n/2}} \exp\left(-\frac{d_0(r,\xi,\theta_1)}{4t}\right) \varphi(\xi)\, \xi^{n-1} \sin^{n-2}\theta_1\, d\theta_1\, d\xi \\ & \quad \times \int_0^\pi \sin^{n-3}\theta_2\, d\theta_2 \times \cdots \times \int_0^\pi \sin\theta_{n-2}\, d\theta_{n-2} \times \int_0^{2\pi} d\theta_{n-1} \\ &= \int_0^\infty \frac{\xi^{n-1}}{(4\pi\, t)^{n/2}} \int_0^\pi \exp\left(-\frac{d_0(r,\xi,\theta_1)}{4t}\right) \sin^{n-2}\theta_1\, d\theta_1\, \varphi(\xi)\, d\xi \\ & \quad \times \prod_{i=1}^{n-3} \int_0^\pi \sin^i\theta\, d\theta \times 2\pi, \end{aligned}$$

where $d_0(r,\xi,\theta_1) = (r-\xi\cos\theta_1)^2 + (\xi\sin\theta_1)^2$. By setting $K_n = 2\pi\prod_{i=1}^{n-3}\int_0^\pi \sin^i\theta\,d\theta$ for $n \geq 4$ and $K_3 = 2\pi$, we obtain (15)–(16). Here K_n for $n \geq 4$ is given by (17)–(18), since we have

$$\int_0^\pi \sin^i\theta\,d\theta = \pi\cdot\frac{1\cdot 3\cdot 5\ldots(2m-1)}{2\cdot 4\cdot 6\ldots 2m} \quad \text{for} \quad i = 2m,$$
$$\int_0^\pi \sin^i\theta\,d\theta = 2\cdot\frac{2\cdot 4\cdot 6\ldots 2m}{1\cdot 3\cdot 5\ldots(2m+1)} \quad \text{for} \quad i = 2m+1.$$

This completes the proof. $\square$

For later discussions, we need some properties of the Green's function $Z^{(n)}(r,\xi,t)$. These properties look like that of the usual heat kernel.

LEMMA 3.2. *The function $Z^{(n)}(r,\xi,r)$ defined in Proposition 3.1 has the following properties:*

(i) *For every $r \geq 0$, it satisfies*

$$\int_0^\infty |Z^{(n)}(r,\xi,t)|\,d\xi = \int_0^\infty Z^{(n)}(r,\xi,t)\,d\xi = 1, \quad t > 0.$$

(ii) *For every $r \geq 0$, it satisfies*

$$\int_0^\infty |Z_\xi^{(n)}(r,\xi,t)|\,d\xi \leq C_2\,t^{-\frac{1}{2}}, \quad t > 0.$$

(iii) *For every ξ with $0 \leq \xi \leq 1$, it satisfies*

$$\sup_{r\geq 2}|Z^{(n)}(r,\xi,t)| \leq C_3\,(1+t)^{-\frac{n}{2}}, \quad t > 0,$$

where $C_2 > 0$ and $C_3 > 0$ are constants depending only on n.

Proof. First we show property (i). By the definition of $Z^{(n)}$, it is obvious that

$$\int_0^\infty Z^{(n)}(r,\xi,t)\,d\xi = \int_{\mathbb{R}^n}\frac{1}{(4\pi\,t)^{n/2}}\exp\left(-\frac{|x-y|^2}{4t}\right)dy = 1, \quad t > 0,$$

holds true for $|x| = r$. In addition, $Z^{(n)}(r,\xi,t) > 0$ also holds true. Thus we obtain property (i).

Next we show property (iii). By the definition of $Z^{(n)}$, we have

$$|Z^{(n)}(r,\xi,t)| \leq \frac{K_n}{(4\pi\,t)^{n/2}}\int_0^\pi \exp\left(-\frac{1}{4t}\right)\sin^{n-2}\theta\,d\theta, \quad t > 0,$$

for $0 \leq \xi \leq 1$ and $r \geq 2$. This implies that

$$\sup_{r\geq 2}|Z^{(n)}(r,\xi,t)| \leq \min\left\{\frac{K_n}{(4\pi\,t)^{n/2}},\ C\right\}, \quad t > 0,$$

for every ξ with $0 \leq \xi \leq 1$, where C is a constant depending only on n. Thus we obtained property (iii).

Finally we show property (ii). For simplicity, we define a function $E^{(n)}$ as

$$E^{(n)}(r,\xi,\theta,t) = \frac{K_n}{(4\pi\,t)^{n/2}}\exp\left(-\frac{(r-\xi\cos\theta)^2+(\xi\sin\theta)^2}{4t}\right).$$

Then by definition, we have

$$\int_0^\infty |Z_\xi^{(n)}|\,d\xi \le \int_0^\infty \int_0^\pi \left| \left(\xi^{n-1} E^{(n)} \sin^{n-2}\theta\right)_\xi \right| d\xi\, d\theta$$
$$\le \int_0^\pi \int_0^\infty \left[(n-1)\xi^{n-2}E^{(n)} + \xi^{n-1}|E_\xi^{(n)}|\right] \sin^{n-2}\theta\, d\xi\, d\theta.$$

Noting that $E_\xi^{(n)} = (r\cos\theta - \xi)/2t \cdot E^{(n)}$ and $E^{(n)} > 0$, we divide it into four terms as follows:

$$\int_0^\infty |Z_\xi^{(n)}|\,d\xi \le \int_0^\pi \int_0^\infty (n-1)\xi^{n-2}E^{(n)} \sin^{n-2}\theta\, d\xi\, d\theta$$
$$+ \left(-\int_{\frac{\pi}{2}}^\pi \int_0^\infty - \int_0^{\frac{\pi}{2}} \int_{r\cos\theta}^\infty + \int_0^{\frac{\pi}{2}} \int_0^{r\cos\theta} \right) \xi^{n-1} E_\xi^{(n)} \sin^{n-2}\theta\, d\xi\, d\theta$$
$$= \int_0^\pi \int_0^\infty (n-1)\xi^{n-2}E^{(n)} \sin^{n-2}\theta\, d\xi\, d\theta + J_1 + J_2 + J_3.$$

By integrating by parts, we have

$$J_1 = \int_{\frac{\pi}{2}}^\pi \int_0^\infty (n-1)\, \xi^{n-2} E^{(n)} \sin^{n-2}\theta\, d\xi\, d\theta,$$
$$J_2 = \int_0^{\frac{\pi}{2}} \left[\xi^{n-1}E^{(n)}\right]_{\xi = r\cos\theta} \sin^{n-2}\theta\, d\theta + \int_0^{\frac{\pi}{2}} \int_{r\cos\theta}^\infty (n-1)\, \xi^{n-2} E^{(n)} \sin^{n-2}\theta\, d\xi\, d\theta,$$
$$J_3 = \int_0^{\frac{\pi}{2}} \left[\xi^{n-1}E^{(n)}\right]_{\xi = r\cos\theta} \sin^{n-2}\theta\, d\theta - \int_0^{\frac{\pi}{2}} \int_0^{r\cos\theta} (n-1)\, \xi^{n-2} E^{(n)} \sin^{n-2}\theta\, d\xi\, d\theta.$$

By combining these results, we have

$$\int_0^\infty |Z_\xi^{(n)}|\,d\xi \le 2 \int_0^{\frac{\pi}{2}} \left[\xi^{n-1}E^{(n)}\right]_{\xi = r\cos\theta} \sin^{n-2}\theta\, d\theta$$
$$+ 2 \int_0^\pi \int_0^\infty (n-1)\xi^{n-2}E^{(n)} \sin^{n-2}\theta\, d\xi\, d\theta = J_4 + J_5.$$

Here J_4 is estimated as

$$J_4 = 2\int_0^{\frac{\pi}{2}} (r\cos\theta)^{n-1} \frac{K_n}{(4\pi\, t)^{n/2}} \exp\left(-\frac{(r\sin\theta)^2}{4t}\right) \sin^{n-2}\theta\, d\theta$$
$$\le \frac{K_n}{\pi^{n/2}} \cdot t^{-\frac{1}{2}} \int_0^\infty z^{n-2} \exp\left(-z^2\right) dz = C\, t^{-\frac{1}{2}},$$

where C is a constant depending only on n. For J_5, we have

$$J_5 = \frac{2\,K_n}{\sqrt{4\pi t}\, K_{n-1}} \int_0^\pi \int_0^\infty (n-1)\xi^{n-2}E^{(n-1)} \sin^{n-2}\theta\, d\xi\, d\theta$$
$$= \widehat{C}\, t^{-\frac{1}{2}} \int_0^\pi \int_0^\infty \xi^{n-2} E^{(n-1)} \sin^{n-3}\theta \cdot \sin\theta\, d\xi\, d\theta$$
$$\le \widehat{C}\, t^{-\frac{1}{2}} \int_0^\infty Z^{(n-1)}\, d\xi = \widehat{C}\, t^{-\frac{1}{2}} \quad \text{if } n \ge 3.$$

If $n = 2$, we have

$$\begin{aligned} J_5 &= 2\int_0^{\pi}\int_0^{\infty} E^{(2)}\,d\xi\,d\theta = 2\int_0^{\pi}\int_0^{\infty} \frac{K_2}{4\pi\,t}\exp\left(-\frac{(r-\xi\cos\theta)^2+(\xi\sin\theta)^2}{4t}\right)d\xi\,d\theta \\ &\le \int_0^{\pi}\int_0^{\infty}\frac{K_2}{2\pi\,t}\exp\left(-\frac{(\xi-r\cos\theta)^2}{4t}\right)d\xi\,d\theta \\ &= \frac{K_2}{\pi}\,t^{-\frac{1}{2}}\int_0^{\pi}\int_{-\infty}^{\infty}\exp\left(-z^2\right)\,dz\,d\theta = \widetilde{C}\,t^{-\frac{1}{2}}, \end{aligned}$$

where $\widetilde{C}$ is a constant depending only on n. Thus we obtain property (ii). This completes the proof. ∎

Finally we give an expression for radially symmetric solutions of (1)–(2) by using the Green's function $Z^{(n)}(r,\xi,t)$ obtained above. This expression plays an essential role in our discussion.

Lemma 3.3. *Let $u(r,t)$ be the solution to the initial value problem of the form*

$$u_t = \frac{u_{rr}}{1+u_r^2} + \frac{n-1}{r}u_r, \quad r>0,\ t>0, \tag{20}$$

$$u(r,0) = u_0(r), \qquad r>0. \tag{21}$$

Then the solution $u(r,t)$ of (20)–(21) *is expressed by*

$$u(r,t) = \int_0^{\infty} Z^{(n)}(r,\xi,t)\,u_0(\xi)\,d\xi - \int_0^t\int_0^{\infty} Z^{(n)}(r,\xi,t-s)\,\frac{u_{\xi\xi}\,u_\xi^2}{1+u_\xi^2}\,d\xi\,ds, \tag{22}$$

where the function $Z^{(n)}(r,\xi,t)$ is defined as in Proposition 3.1.

Proof. Since (1) is rewritten as

$$u_t = \Delta u - \sum_{i,j=1}^{n}\frac{u_{x_i}u_{x_j}u_{x_ix_j}}{1+|\nabla u|^2},$$

we have the expression

$$\begin{aligned} u(x,t) &= \int_{\mathbb{R}^n}\frac{1}{(4\pi\,t)^{n/2}}\exp\left(-\frac{|x-y|^2}{4t}\right)u_0(y)\,dy \\ &\quad - \sum_{i,j=1}^{n}\int_0^t\int_{\mathbb{R}^n}\frac{1}{(4\pi(t-s))^{n/2}}\exp\left(-\frac{|x-y|^2}{4(t-s)}\right)\frac{u_{y_i}u_{y_j}u_{y_iy_j}}{1+|\nabla u|^2}\,dy\,ds. \end{aligned}$$

If the solution $u(x,t)$ is radially symmetric, we have $\partial/\partial x_i = x_i/r\cdot\partial/\partial r$ and hence obtain

$$\sum_{i,j=1}^{n}\frac{u_{x_i}u_{x_j}u_{x_ix_j}}{1+|\nabla u|^2} = \frac{u_{rr}u_r^2}{1+u_r^2}.$$

Thus we obtain (22) by changing the variables to the polar coordinate system similarly to the proof of Proposition 3.1. This completes the proof. ∎

4. Decay estimates for the derivatives. In this section, we show the decay estimates for derivatives of radially symmetric solutions of (1)–(2). When we consider radially symmetric solutions, the problem (1)–(2) is rewritten as follows:

$$u_t = \frac{u_{rr}}{1+u_r^2} + \frac{n-1}{r}u_r, \qquad r>0,\ t>0, \tag{23}$$

$$u(r,0) = u_0(r), \qquad r>0, \tag{24}$$

$$u_r(0,t)=0,\ u_{tr}(0,t)=0, \quad t>0, \tag{25}$$

where $r=|x|$. The boundary condition (25) is necessary to derive the decay estimates for the derivatives. In what follows, we focus on this problem to study the large time behavior of radially symmetric solutions.

First we prepare maximum principles. In Chapter 2 of [10], the maximum principle for the Cauchy problem of usual parabolic equations is given. As a modification of discussions in [10], we provide the maximum principles on the half-space with the Neumann boundary condition. We need to take care of the coefficient $(n-1)/r$ in (23).

LEMMA 4.1. *Let $g=g(r,t)$ be a continuous and bounded function for $r\ge 0$ and $t\ge 0$. Suppose that a function $w=w(r,t)$ is continuous and bounded for $r\ge 0$ and $t\ge 0$, has continuous derivatives w_t, w_r, w_{rr} for $r\ge 0$ and $t>0$, and satisfies*

$$\begin{aligned} &L[w] = w_t - \frac{w_{rr}}{1+g^2} - \frac{n-1}{r}w_r \le 0, \quad && r>0,\ t>0,\\ &w_r(0,t)=0, && t>0. \end{aligned}$$

Then $w(r,t)$ satisfies

$$\sup_{r>0,\ t>0} w(r,t) \le \|w(r,0)\|_{L^\infty(0,\infty)}.$$

Proof. We define a function

$$E(r,t) = \exp\left(\frac{r^2}{1-\mu t}+\nu t\right) \quad \text{on } R_0 = [0,\infty)\times\left[0,\frac{1}{2\mu}\right].$$

By taking constants $\mu=16$ and $\nu>4n$, we have $L[E]/E>0$ on R_0. Indeed, because of $1/2\le 1-\mu t\le 1$, it follows that

$$\begin{aligned} \frac{L[E]}{E} &= \frac{1}{E}\left(E_t - \frac{E_{rr}}{1+g^2} - \frac{n-1}{r}E_r\right)\\ &= \frac{\mu r^2}{(1-\mu t)^2} + \nu - \frac{1}{1+g^2}\left(\frac{2}{1-\mu t}+\frac{4r^2}{(1-\mu t)^2}\right) - \frac{n-1}{r}\cdot\frac{2r}{1-\mu t}\\ &\ge \mu r^2 + \nu - (4+16r^2) - 4(n-1) = \nu - 4n. \end{aligned}$$

For the constant $M=\|w(r,0)\|_{L^\infty(0,\infty)}$, we consider the function $v=(w-M)/E$ on R_0. Then v satisfies

$$v_t - \frac{v_{rr}}{1+g^2} - \frac{n-1}{r}v_r - \frac{2E_r}{E(1+g^2)}v_r + \frac{L[E]}{E}v \le 0, \tag{26}$$

$$v(r,0)\le 0. \tag{27}$$

Moreover $\lim_{r\to\infty} v(r,t)=0$ holds true for $0\le t\le 1/2\mu$, since w is bounded by the assumption. Thus the inequality (27) implies that v satisfies one of the following:

(i) v takes the positive maximum on R_0 at a point $(0, t_0)$ with $0 < t_0 \leq 1/2\mu$.

(ii) v takes the positive maximum on R_0 at a point (r_0, t_0) with $r_0 > 0$ and $0 < t_0 \leq 1/2\mu$.

(iii) $v(x, t) \leq 0$ on R_0.

Case (i) cannot happen. Indeed, by passing to the limit as $r \to 0$ in (26), we have

$$v_t(0,t) - \left(\frac{1}{1+g^2} + (n-1)\right) v_{rr}(0,t) + \frac{L[E]}{E} v(0,t) \leq 0, \tag{28}$$

by the boundary condition $v_r(0,t) = 0$. On the other hand, we have $v_t(0, t_0) \geq 0$, $v_{rr}(0, t_0) \leq 0$, and $v(0, t_0) > 0$, since $v(0, t_0)$ is the positive maximum. Because of $L[E]/E > 0$, these values do not satisfy (28). This is a contradiction.

Case (ii) also leads to a contradiction. Indeed, for any $R > 0$, v takes the positive maximum on $[r_0/2, r_0 + R] \times [0, 1/2\mu]$ at the interior point (r_0, t_0). Thus the strong maximum principle implies that v is positive and constant on $[r_0/2, r_0 + R] \times [0, 1/2\mu]$. Since $R > 0$ is arbitrary, this contradicts $\lim_{r\to\infty} v(r,t) = 0$.

Consequently, case (iii) holds true. Namely, we have $v = (w - M)/E \leq 0$ on R_0. Since E is positive, this implies that $w \leq M$ on R_0. By repeating a similar argument for $t \geq 1/2\mu$, we obtain $w \leq M$ for $r \geq 0$ and $t \geq 0$. This completes the proof. ∎

Similarly to the above, we obtain the following lemma. Here the equation differs slightly from that of Lemma 4.1.

LEMMA 4.2. *Let $g = g(r, t)$ be a given function such that g and g_r are continuous and bounded for $r \geq 0$ and $t \geq 0$. Suppose that a function $w = w(r, t)$ is continuous and bounded for $r \geq 0$ and $t \geq 0$, has continuous derivatives w_t, w_r, w_{rr} for $r \geq 0$ and $t > 0$, and satisfies*

$$\begin{aligned} &L[w] = w_t - \left(\frac{w_r}{1+g^2}\right)_r - \frac{n-1}{r} w_r \leq 0, && r > 0,\ t > 0, \\ &w_r(0,t) = 0, && t > 0. \end{aligned}$$

Then $w(r,t)$ satisfies

$$\sup_{r>0,\, t>0} w(r,t) \leq \|w(r,0)\|_{L^\infty(0,\infty)}.$$

Proof. We define a function

$$E(r,t) = \exp\left(\frac{r^2}{1-\mu t} + \nu t\right) \quad \text{on } R_0 = [0,\infty) \times \left[0, \frac{1}{2\mu}\right].$$

By taking constants $\mu \geq 16 + 8\|gg_r\|_{L^\infty((0,\infty)\times(0,\infty))}$ and $\nu > 4n + 8\|gg_r\|_{L^\infty((0,\infty)\times(0,\infty))}$, we have $L[E]/E > 0$ on R_0. Indeed, because of $1/2 \leq 1 - \mu t \leq 1$, we have

$$\begin{aligned} \frac{L[E]}{E} &= \frac{1}{E}\left(E_t - \left(\frac{E_r}{1+g^2}\right)_r - \frac{n-1}{r}E_r\right) \\ &= \frac{\mu r^2}{(1-\mu t)^2} + \nu - \frac{1}{1+g^2}\left(\frac{2}{1-\mu t} + \frac{4r^2}{(1-\mu t)^2}\right) + \frac{2gg_r}{(1+g^2)^2}\frac{2r}{1-\mu t} - \frac{n-1}{r}\frac{2r}{1-\mu t} \\ &\geq \mu r^2 + \nu - (4 + 16r^2) - 8\|gg_r\|_{L^\infty((0,\infty)\times(0,\infty))} \cdot r - 4(n-1) \\ &\geq \left(\mu - 16 - 8\|gg_r\|_{L^\infty((0,\infty)\times(0,\infty))}\right) r^2 + \nu - 4n - 8\|gg_r\|_{L^\infty((0,\infty)\times(0,\infty))}. \end{aligned}$$

For the constant $M = \|w(r,0)\|_{L^\infty(0,\infty)}$, we consider the function $v = (w-M)/E$ on R_0. Then v satisfies

$$\begin{aligned}&v_t - \left(\frac{v_r}{1+g^2}\right)_r - \frac{n-1}{r}v_r - \frac{2E_r}{E(1+g^2)}v_r + \frac{L[E]}{E}v \le 0,\\ &v(r,0) \le 0.\end{aligned}$$

Moreover $\lim_{r\to\infty} v(r,t) = 0$ holds true for $0 \le t \le 1/2\mu$. Thus, similarly to the proof of Lemma 4.1, we obtain $w \le M$ on R_0. By repeating a similar argument for $t \ge 1/2\mu$, we obtain $w \le M$ for $r \ge 0$ and $t \ge 0$. This completes the proof. □

Now we show the decay estimates for derivatives of the solution of (23)–(25). The following estimate for $|u_t|$ is derived by using the idea of Ecker and Huisken [6].

PROPOSITION 4.3. *The solution $u(r,t)$ of (23)–(25) satisfies the following estimate:*

$$\sup_{r>0} |u_t(r,t)| \le C_4\, t^{-\frac{1}{2}}, \quad t \ge 1, \tag{29}$$

where $C_4 > 0$ is a constant depending only on n and u_0.

Proof. For the solution $u(r,t)$ of (23)–(25), we define an operator L as follows:

$$L[w] = w_t - \left(\frac{w_r}{1+u_r^2}\right)_r - \frac{n-1}{r}w_r.$$

We consider a function $V(r,t)$ defined by

$$V(r,t) = u_r^2 + \frac{t}{n-1}u_t^2.$$

Here we note that $V(r,t)$ is continuous and bounded for $r \ge 0$ and $t \ge 1$ by virtue of the estimate (11) in Proposition 2.2. Then we obtain

$$L[V] = L[u_r^2] + \frac{1}{n-1}\,u_t^2 + \frac{t}{n-1}\,L[u_t^2] \le -\frac{u_t^2}{n-1} + \frac{1}{n-1}u_t^2 = 0.$$

Indeed, we have

$$\begin{aligned}L[u_r^2] &= -\frac{2u_{rr}^2}{1+u_r^2} - \frac{2(n-1)}{r^2}u_r^2 \le -\frac{2}{n-1}\left[\left(\frac{u_{rr}}{1+u_r^2}\right)^2 + \left(\frac{n-1}{r}u_r\right)^2\right]\\ &\le -\frac{1}{n-1}\left(\frac{u_{rr}}{1+u_r^2} + \frac{n-1}{r}u_r\right)^2 = -\frac{u_t^2}{n-1},\\ L[u_t^2] &= -\frac{2u_{rt}^2}{1+u_r^2} \le 0.\end{aligned}$$

Since the boundary condition (25) gives

$$V_r(0,t) = \left.\left(2u_r u_{rr} + \frac{2t}{n-1}u_t u_{tr}\right)\right|_{r=0} = 0, \quad t > 0,$$

we can apply Lemma 4.2 and hence obtain

$$\sup_{r>0} V(r,t) = \sup_{r>0}\left(u_r^2 + \frac{t}{n-1}u_t^2\right) \le \|V(r,1)\|_{L^\infty(0,\infty)}.$$

The estimate (11) in Proposition 2.2 implies that $\|V(r,1)\|_{L^\infty(0,\infty)}$ depends only on n and u_0. Thus the decay estimate (29) follows. This completes the proof. $\square$

Next we use a similar technique for deriving the decay estimate for $|u_r|$. Note that the definition of the operator L differs slightly from that in Proposition 4.3.

Proposition 4.4. *The solution $u(r,t)$ of (23)–(25) satisfies the following estimate:*

$$\sup_{r>0} |u_r(r,t)| \le C_5\,(1+t)^{-\frac{1}{2}}, \quad t \ge 0, \tag{30}$$

where $C_5 > 0$ is a constant depending only on $\|u_0\|_{W^{1,\infty}(0,\infty)}$.

Proof. For the solution $u(r,t)$ of (23)–(25), we define an operator L as follows:

$$L[w] = w_t - \frac{w_{rr}}{1+u_r^2} - \frac{n-1}{r} w_r.$$

We denote the derivative of $u_0(r)$ with respect to r by $u_0'(r)$. For the constant $M = \|u_0'\|_{L^\infty(0,\infty)}$, we consider a function $V(r,t)$ defined by

$$V(r,t) = u^2 + \frac{1+t}{1+M^2} u_r^2.$$

Since Proposition 2.1 gives $|u_r| \le M$, we have

$$L[V] = L[u^2] + \frac{1}{1+M^2}\, u_r^2 + \frac{1+t}{1+M^2}\, L[u_r^2] \le -\frac{2u_r^2}{1+u_r^2} + \frac{1}{1+M^2} u_r^2 \le 0.$$

Indeed, we have

$$L[u^2] = -\frac{2u_r^2}{1+u_r^2} \quad \text{and} \quad L[u_r^2] = -\frac{4u_r^2 u_{rr}^2}{(1+u_r^2)^2} - \frac{2(n-1)}{r^2} u_r^2 - \frac{2u_{rr}^2}{1+u_r^2} \le 0.$$

Noting that

$$V_r(0,t) = \left(2uu_r + \frac{2(1+t)}{1+M^2} u_r u_{rr}\right)\bigg|_{r=0} = 0, \quad t > 0,$$

we apply Lemma 4.1 and hence obtain

$$\sup_{r>0} V(r,t) = \sup_{r>0} \left(u^2 + \frac{1+t}{1+M^2} u_r^2\right) \le \|u_0\|^2_{L^\infty(0,\infty)} + \frac{\|u_0'\|^2_{L^\infty(0,\infty)}}{1+M^2}, \quad t \ge 0.$$

This implies the estimate (30). This completes the proof. $\square$

Now we give the decay estimate for $|u_{rr}|$ by combining the estimates established above. The following estimate may have room for improvement. But this suffices to prove Theorem 1.3.

Proposition 4.5. *The solution $u(r,t)$ of (23)–(25) satisfies the following estimate:*

$$|u_{rr}(r,t)| \le C_6 \cdot \min\left\{1, \frac{1}{r}\, t^{-\frac{1}{2}}\right\}, \quad r > 0,\ t \ge 1, \tag{31}$$

where $C_6 > 0$ is a constant depending only on n and u_0.

Proof. The estimate (11) in Proposition 2.2 implies that $|u_{rr}|$ is bounded uniformly for $t \ge 1$. On the other hand, by substituting the decay estimates for $|u_t|$ and

$|u_r|$ to (23), we find that $|u_{rr}|$ decays with the rate of $1/r \times t^{-1/2}$. Thus we obtain the estimate (31). ∎

Theorem 1.3 will be proved in section 5 by using the decay estimates for $|u_r|$ and $|u_{rr}|$ given by Propositions 4.4 and 4.5. Finally, we modify the proof of Proposition 4.3 and improve the estimate for $|u_t|$ which, however, will not be used in the later discussions.

Proposition 4.6. *The solution $u(r,t)$ of (23)–(25) satisfies the following estimate:*

$$\sup_{r>0} |u_t(r,t)| \le C_7\, t^{-1}\sqrt{\log t}, \quad t > e^2, \tag{32}$$

where $C_7 > 0$ is a constant depending only on n and u_0.

Proof. Similarly to the proof of Proposition 4.3, we define an operator L as follows:

$$L[w] = w_t - \left(\frac{w_r}{1+u_r^2}\right)_r - \frac{n-1}{r} w_r.$$

Let σ be any constant with $0 < \sigma < 1/2$. We consider a function $V(r,t)$ defined by

$$V(r,t) = t^{1-\sigma} u_r^2 + \frac{t^{2-\sigma}}{(2-\sigma)(n-1)} u_t^2,$$

where $V(r,t)$ is continuous and bounded for $r \ge 0$ and $t \ge 1$ from Proposition 2.2. By noting that the inequalities $L[u_r^2] \le -u_t^2/(n-1)$ and $L[u_t^2] \le 0$ hold true as in the proof of Proposition 4.3, we have

$$\begin{aligned} L[V] &= (1-\sigma)\, t^{-\sigma} u_r^2 + t^{1-\sigma} L[u_r^2] + \frac{t^{1-\sigma}}{n-1}\, u_t^2 + \frac{t^{2-\sigma}}{(2-\sigma)(n-1)}\, L[u_t^2] \\ &\le (1-\sigma)\, t^{-\sigma} u_r^2. \end{aligned}$$

By substituting the decay estimate for $|u_r|$ established in Proposition 4.4, we have

$$L[V] \le C(1-\sigma)\, t^{-(1+\sigma)},$$

where $C > 0$ is a constant depending only on u_0. Now we define

$$P(t) = \|V(r,1)\|_{L^\infty(0,\infty)} + \frac{C(1-\sigma)}{\sigma}\left(1 - t^{-\sigma}\right)$$

for $t \ge 1$. Then we find that $\widehat{V}(r,t) = V(r,t) - P(t)$ satisfies $L[\widehat{V}] \le 0$ and $\widehat{V}(r,1) \le 0$. Thus Lemma 4.2 gives $\widehat{V}(r,t) \le 0$ for $t \ge 1$, and thus

$$V(r,t) \le P(t) \le P(\infty) = \|V(r,1)\|_{L^\infty(0,\infty)} + \frac{C(1-\sigma)}{\sigma}, \quad t \ge 1.$$

Since $0 < \sigma < 1/2$ by the definition, this implies that

$$\begin{aligned} t^{2-\sigma} u_t^2 &\le (2-\sigma)(n-1)\left(\|V(r,1)\|_{L^\infty(0,\infty)} + \frac{C(1-\sigma)}{\sigma}\right) \\ &\le \frac{2(n-1)}{\sigma}\left(\|V(r,1)\|_{L^\infty(0,\infty)} + C\right), \quad t \ge 1. \end{aligned}$$

Here $\|V(r,1)\|_{L^\infty(0,\infty)}$ depends only on n and u_0 from Proposition 2.2. Consequently, there exists a constant $\widetilde{C} > 0$ depending only on n and u_0 such that, for any constant $0 < \sigma < 1/2$, it holds that

$$|u_t| \le \frac{\widetilde{C}}{\sqrt{\sigma}}\, t^{-1+\frac{\sigma}{2}}, \quad t \ge 1.$$

Now we set $\sigma = 1/\log t$ for each $t \ge 1$. Here we assume that $t > e^2$ for $\sigma < 1/2$. Then we have

$$|u_t| \le \frac{\widetilde{C}}{\sqrt{\sigma}} \cdot t^{-1} \exp\left(\log t \cdot \frac{\sigma}{2}\right) = \widetilde{C}\sqrt{\log t} \cdot t^{-1}\sqrt{e}, \quad t > e^2.$$

This implies the estimate (32). The proof is completed. ∎

5. Proof of Theorem 1.3. In this section, we give a proof of Theorem 1.3. We analyze the expression for $u(r,t)$ obtained in Lemma 3.3 by using the properties of the Green's function $Z^{(n)}(r,\xi,r)$ and the decay estimates for the derivatives u_r and u_{rr}.

Proof of Theorem 1.3. Proposition 2.1 gives the existence of the solution globally in time. It suffices to show large time behavior of the solution. In this proof, we always write C or $\widetilde{C}$ for positive constants depending only on n and u_0, which may take a different value in each context. From Lemma 3.3, the solution $u(r,t)$ of (23)–(25) is given by

$$\begin{aligned} u(r,t) &= \int_0^\infty Z^{(n)}(r,\xi,t)\, u_0(\xi)\, d\xi - \int_0^t \int_0^\infty Z^{(n)}(r,\xi,t-s)\, \frac{u_{\xi\xi}u_\xi^2}{1+u_\xi^2}\, d\xi\, ds \\ &= h(r,t) - \int_0^t \int_0^\infty Z^{(n)}(r,\xi,t-s)\, \frac{u_{\xi\xi}u_\xi^2}{1+u_\xi^2}\, d\xi\, ds. \end{aligned}$$

Since the function $h(r,t)$ is the solution of the Cauchy problem of the heat equation, it suffices to show the decay estimate for the second term of the right-hand side. For this purpose, we divide it into two terms and define I, I_1, and I_2 as follows:

$$I = -\left(\int_0^{t/2} + \int_{t/2}^t\right) \int_0^\infty Z^{(n)}(r,\xi,t-s)\, \frac{u_{\xi\xi}u_\xi^2}{1+u_\xi^2}\, d\xi\, ds = -(I_1 + I_2).$$

First we evaluate I_1. By noting that $u_{\xi\xi}u_\xi^2/(1+u_\xi^2) = (u_\xi - \arctan u_\xi)_\xi$ and integrating I_1 by parts, we have

$$|I_1| \le \int_0^{t/2} \int_0^\infty |Z_\xi^{(n)}|\, |u_\xi - \arctan u_\xi|\, d\xi\, ds.$$

Here the Taylor expansion for $f(p) = \arctan p$ implies that

$$\begin{aligned} |p - f(p)| &= \left| p - \left(f(0) + f'(0)\,p + \frac{f''(0)}{2}\, p^2 + \frac{f'''(\delta p)}{3!}\, p^3 \right) \right| \\ &= \left| \frac{3(\delta p)^2 - 1}{3(1+(\delta p)^2)^3}\, p^3 \right| \le |p|^3 \end{aligned}$$

for a constant δ with $0<\delta<1$. Thus by virtue of property (ii) of $Z^{(n)}$ in Lemma 3.2 and the decay estimate for $|u_r|$ in Proposition 4.4, we obtain

$$\begin{aligned}|I_1| &\le \int_0^{t/2}\int_0^\infty |Z_\xi^{(n)}|\,|u_\xi|^3\,d\xi\,ds \le C\int_0^{t/2}(t-s)^{-\frac12}(1+s)^{-\frac32}\,ds\\ &\le \sqrt{2}\,C\,t^{-\frac12}\int_0^{t/2}(1+s)^{-\frac32}\,ds \le 2\sqrt{2}\,C\,t^{-\frac12},\quad t>0.\end{aligned}$$

Next we evaluate $|I_2|$. We divide I_2 into three terms as follows:

$$\begin{aligned}I_2 &= -\int_{t/2}^t\left(\int_0^{s^{-\frac12}}+\int_{s^{-\frac12}}^1+\int_1^\infty\right)Z^{(n)}(r,\xi,t-s)\,\frac{u_{\xi\xi}u_\xi^2}{1+u_\xi^2}\,d\xi\,ds\\ &= -(I_{21}+I_{22}+I_{23}).\end{aligned}$$

Here we assume that $t\ge 2$ for $s^{-1/2}\le 1$. To evaluate I_{21}, we use the decay estimate for $|u_r|$ given by Proposition 4.4 and the L^∞-bounds for $Z^{(n)}$ and $|u_{rr}|$ given by property (iii) of Lemma 3.2 and Proposition 4.5. Then, for $r\ge 2$, we have

$$\begin{aligned}|I_{21}| &\le \int_{t/2}^t\int_0^{s^{-\frac12}}\left|Z^{(n)}(r,\xi,t-s)\right|\left|\frac{u_{\xi\xi}u_\xi^2}{1+u_\xi^2}\right|d\xi\,ds \le C\int_{t/2}^t\int_0^{s^{-\frac12}} s^{-1}\,d\xi\,ds\\ &= C\int_{t/2}^t s^{-\frac32}\,ds \le 2\sqrt{2}\,C\,t^{-\frac12},\quad t\ge 2.\end{aligned}$$

Next, for I_{22}, we use property (iii) of $Z^{(n)}$ and the decay estimates for $|u_{rr}|$ and $|u_r|$. Then, for $r\ge 2$, we have

$$\begin{aligned}|I_{22}| &\le \int_{t/2}^t\int_{s^{-\frac12}}^1\left|Z^{(n)}(r,\xi,t-s)\right|\left|\frac{u_{\xi\xi}u_\xi^2}{1+u_\xi^2}\right|d\xi\,ds\\ &\le C\int_{t/2}^t\int_{s^{-\frac12}}^1(1+t-s)^{-\frac n2}\cdot\frac{1}{\xi}\,s^{-\frac32}\,d\xi\,ds\\ &\le \frac{C}{2}\int_{t/2}^t(1+t-s)^{-\frac n2}\,s^{-\frac32}\,\log s\,ds\\ &\le \sqrt{2}\,C\,t^{-\frac32}\,\log t\int_{t/2}^t(1+t-s)^{-\frac n2}\,ds\\ &\le \sqrt{2}\,C\,t^{-1}\int_{t/2}^t(1+t-s)^{-\frac n2}\,ds \le \widetilde{C}\,t^{-\frac12},\quad t\ge 2.\end{aligned}$$

Finally, for I_{23}, we use property (i) of $Z^{(n)}$ and the decay estimates for $|u_{rr}|$ and $|u_r|$. Then we have

$$|I_{23}| \le C\int_{t/2}^t\int_1^\infty\left|Z^{(n)}(r,\xi,t-s)\right|s^{-\frac32}\,d\xi\,ds \le C\int_{t/2}^t s^{-\frac32}\,ds \le 2\sqrt{2}\,C\,t^{-\frac12}$$

for $t\ge 2$. By combining all estimates established above, we obtain the following estimate:

$$\sup_{r\ge 2}|u(r,t)-h(r,t)| \le C\,t^{-\frac12},\quad t\ge 2.$$

Thus it remains to show the large time behavior of $u(r,t)$ for $r \in [0,2]$. If $0 \le r \le 2$, we have

$$|u(r,t) - h(r,t)| \le |u(r,t) - u(2,t)| + |u(2,t) - h(2,t)| + |h(2,t) - h(r,t)|$$
$$\le 2 \sup_{r>0} |u_r(r,t)| + 2 \sup_{r>0} |h_r(r,t)| + C\, t^{-\frac{1}{2}}, \quad t \ge 2.$$

For the heat equation, it is well known that $\sup_{r>0} |h_r(r,t)| \le t^{-\frac{1}{2}} \|u_0'\|_{L^\infty(0,\infty)}$ for $t > 0$. By this fact and the decay estimate for $|u_r|$ in Proposition 4.4, we obtain the estimate

$$\sup_{r>0} |u(r,t) - h(r,t)| \le C\, t^{-\frac{1}{2}}, \quad t \ge 2.$$

This completes the proof of Theorem 1.3. $\square$

Acknowledgment. The author thanks Prof. Masaharu Taniguchi at Tokyo Institute of Technology who gave me valuable comments and suggestions.

REFERENCES

[1] Y.-G. Chen, Y. Giga, and S. Goto, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, J. Differential Geom., 33 (1991), pp. 749–786.
[2] X. Chen, *Generation and propagation of interfaces for reaction-diffusion equations*, J. Differential Equations, 96 (1992), pp. 116–141.
[3] K.-S. Chou and Y.-C. Kwong, *On quasilinear parabolic equations which admit global solutions for initial data with unrestricted growth*, Calc. Var., 12 (2001), pp. 281–315.
[4] P. Collet and J.-P. Eckmann, *Space-time behaviour in problems of hydrodynamic type: A case study*, Nonlinearity, 5 (1992), pp. 1265–1302.
[5] V. N. Denisov and V. D. Repnikov, *The stabilization of a solution of a Cauchy problem for parabolic equations*, Differ. Equ., 20 (1984), pp. 16–33.
[6] K. Ecker and G. Huisken, *Mean curvature evolution of entire graphs*, Ann. of Math., 130 (1989), pp. 453–471.
[7] S.-I. Ei and E. Yanagida, *Dynamics of interfaces in competition-diffusion systems*, SIAM J. Appl. Math., 54 (1994), pp. 1355–1373.
[8] S. D. Eidel'man, *Parabolic Systems*, North-Holland, Amsterdam, 1969.
[9] L. C. Evans and J. Spruck, *Motion of level sets by mean curvature*, J. Differential Geom., 33 (1991), pp. 635–681.
[10] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[11] M. Gage and R. Hamilton, *The shrinking of convex plane curves by the heat equation*, J. Differential Geom., 23 (1986), pp. 69–96.
[12] M. Grayson, *The heat equation shrinks embedded plane curves to points*, J. Differential Geom., 26 (1987), pp. 285–314.
[13] G. Huisken, *Flow by mean curvature of convex surface into surface*, J. Differential Geom., 20 (1984), pp. 237–266.
[14] N. Ishimura, *Curvature evolution of plane curves with prescribed opening angle*, Bull. Austral. Math. Soc., 52 (1995), pp. 287–296.
[15] S. Kamin, *On stabilisation of solutions of the Cauchy problem for parabolic equations*, Proc. Roy. Soc. Edinburgh, 76A (1976), pp. 43–53.
[16] G. M. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, Singapore, 2005.
[17] M. Nara and M. Taniguchi, *The condition on the stability of stationary lines in a curvature flow in the whole plane*, J. Differential Equations, 237 (2007), pp. 61–76.
[18] V. D. Repnikov and S. D. Eidel'man, *A new proof of the theorem on the stabilization of the solution of the Cauchy problem for the heat equation*, Mat. Sb., 2 (1967), pp. 135–139.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 1996–2015

THE DERRIDA–LEBOWITZ–SPEER–SPOHN EQUATION: EXISTENCE, NONUNIQUENESS, AND DECAY RATES OF THE SOLUTIONS*

ANSGAR JÜNGEL† AND DANIEL MATTHES‡

Abstract. The logarithmic fourth-order equation $\partial_t u + \frac{1}{2}\sum_{i,j=1}^d \partial_{ij}^2(u\partial_{ij}^2 \log u) = 0$, called the Derrida–Lebowitz–Speer–Spohn equation, with periodic boundary conditions is analyzed. The global-in-time existence of weak nonnegative solutions in space dimensions $d \leq 3$ is shown. Furthermore, a family of entropy-entropy dissipation inequalities is derived in arbitrary space dimensions, and rates of the exponential decay of the weak solutions to the homogeneous steady state are estimated. The proofs are based on the algorithmic entropy construction method developed by the authors and on an exponential variable transformation. Finally, an example for nonuniqueness of the solution is provided.

Key words. logarithmic fourth-order equation, entropy-entropy dissipation method, existence of weak solutions, long-time behavior of solutions, decay rates, nonuniqueness of solutions

AMS subject classifications. 35K30, 35B40, 35Q40, 35Q99

DOI. 10.1137/060676878

1. Introduction.

The logarithmic fourth-order equation

$$\partial_t u + \frac{1}{2}\partial_{ij}^2(u\partial_{ij}^2 \log u) = 0, \quad u(0,\cdot) = u_0 \geq 0, \tag{1.1}$$

appears in various places in mathematical physics (notice that we employed the summation convention). It has been first derived by Derrida, Lebowitz, Speer, and Spohn [10, 11], and we shall therefore refer to (1.1) as the *DLSS equation*. Derrida et al. studied in [10, 11] interface fluctuations in a two-dimensional spin system, the so-called (time-discrete) Toom model. In a suitable scaling limit, a random variable u related to the deviation of the interface from a straight line satisfies the one-dimensional equation (1.1). The multidimensional DLSS equation appears in quantum semiconductor modeling as the zero-temperature, zero-field limit of the quantum drift-diffusion model [18]. The variable u describes the electron density in a microelectronic device or in a quantum plasma. In both applications, the variable u is a nonnegative quantity.

In fact, the proof of positivity or nonnegativity of solutions constitutes the main analytical difficulty in rigorous studies of (1.1). There is generally no maximum principle available for fourth-order equations, which would allow to conclude from $u_0 \geq 0$ that also $u(t,\cdot) \geq 0$ at later times $t > 0$. Consequently, one has to rely on suitable regularization techniques and a priori estimates. The latter are difficult to

*Received by the editors December 6, 2006; accepted for publication (in revised form) August 30, 2007; published electronically March 21, 2008. This research was partially supported by the Deutsche Forschungsgemeinschaft (DFG), grants JU359/5 (Priority Program "Multiscale Problems") and JU359/7. This research is also part of the ESF program "Global and geometrical aspects of nonlinear partial differential equations (GLOBAL)."

http://www.siam.org/journals/sima/39-6/67687.html

†Institute for Analysis and Scientific Computing, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Wien, Austria (juengel@anum.tuwien.ac.at). This author was partially supported by the Wissenschaftskolleg "Differential Equations," funded by the Fonds zur Förderung der wissenschaftlichen Forschung (FWF).

‡Department of Mathematics, University of Pavia, Via Ferrata 1, 27100 Pavia, Italy (matthes@mathematik.uni-mainz.de).

obtain because of the highly nonlinear structure of the equation. We remark that similar difficulties appear in studies of the thin-film equation

$$\partial_t u + \operatorname{div}(u^\alpha \nabla \Delta u) = 0, \quad u(0,\cdot) = u_0 \geq 0.$$

For this equation it is well known that preservation of positivity strongly depends on the parameter $\alpha > 0$. For a certain range of α's, there are solutions which are strictly positive initially, but which vanish at certain points after finite time [2].

In the present paper, we prove global-in-time existence of nonnegative weak solutions to (1.1) on the d-dimensional torus $\mathbb{T}^d$, and we calculate rates for the exponential decay of the solutions to the homogeneous steady state. Moreover, we provide a family of initial data u_0 for which there exist at least two solutions. These results are new in the literature (also see below). Our method of proof is based on the entropy construction method recently developed in [16] to derive a priori estimates and an exponential transformation of variables to prove the nonnegativity of solutions.

The first mathematically rigorous treatment of (1.1) is due to [4]. There, *local-in-time* existence of classical solutions for strictly positive $W^{1,p}(\mathbb{T}^d)$ initial data with $p > d$ was proven. The existence result is obtained by means of classical semigroup theory applied to the equation

$$2\partial_t \sqrt{u} + \Delta^2 \sqrt{u} - \frac{(\Delta\sqrt{u})^2}{\sqrt{u}} = 0, \quad \sqrt{u}(0,x) = \sqrt{u_0(x)} > 0, \quad x \in \mathbb{T}^d,$$

which is equivalent to (1.1) as long as u remains bounded away from zero. Lacking suitable a priori estimates, existence was proven only locally in time (for $d > 1$), even for strictly positive solutions.

More information is available in dimension $d = 1$ because (1.1) is then well posed in H^1. The *Fisher information*

$$F = \int_{\mathbb{T}} (\sqrt{u})_x^2 dx$$

is a Lyapunov functional, $dF/dt \leq 0$, which allows to relate global existence of solutions to strict positivity: if a classical solution breaks down at $t = t^*$, then the limit profile $\lim_{t \nearrow t^*} u(t,x)$ is still in H^1 but vanishes at some point $x \in \mathbb{T}$.

This observation has motivated the study on *nonnegative weak* solutions instead of positive classical solutions. In [17] global existence for the one-dimensional DLSS equation was shown in the class of functions with finite generalized entropy, $\widetilde{E}_0 = \int_{\mathbb{T}} (u - \log u) dx < \infty$ and with physically motivated boundary conditions. The key ingredient in the proof is the observation that $\widetilde{E}_0$ constitutes another Lyapunov functional for (1.1), providing nonnegativity of the solutions. The restriction to one spatial dimension is essential, since $\widetilde{E}_0$ is seemingly *not* a Lyapunov functional in dimensions $d > 1$.

In the following years, the one-dimensional DLSS equation with (mainly) periodic boundary conditions was extensively studied in the context of entropy-entropy production methods, and the exponentially fast decay of the solutions to the steady state has been proved [6, 7, 12, 14, 16, 19]. A numerical study of the long-time asymptotics for various boundary conditions can be found in [8].

Concerning the multidimensional problem, we remark that an independent investigation of (generalizations to) the DLSS equation has been just finished [13]. There, it is proven that (1.1) constitutes the gradient flow for the Fisher information with respect to the Wasserstein distance. The resulting existence theorem is more general than ours as it also holds in the nonphysical dimensions $d \geq 4$ and on unbounded domains. Clearly, the treatment of the DLSS equation as a gradient flow promotes a deeper understanding of its nature. Our approach in the present note is complementary as it is very direct and much simpler (and also much shorter). Furthermore, we point out that the decay estimates derived by our methods are slightly sharper than those of [13], and we are able to present an example of nonuniqueness of solutions.

In the following we describe our results in more detail. The global existence result is based on the fact that the *physical entropy*

$$\widetilde{E}_1 = \int_{\mathbb{T}^d} u \log\left(\frac{u}{\int u dx}\right) dx \geq 0 \tag{1.2}$$

is a Lyapunov functional in any space dimension $d \geq 1$. In fact, multiplying (1.1) formally by $\log u$, integrating over $\mathbb{T}^d$, and integrating by parts lead to

$$\frac{d\widetilde{E}_1}{dt} + \frac{1}{2}\int_{\mathbb{T}^d} u\|\nabla^2 \log u\|^2 dx = 0,$$

where $\nabla^2 \log u$ is the Hessian of $\log u$ and $\|\cdot\|$ is the Euclidean norm. We call the above integral the *entropy production*. Since there is no lower bound for u available, this does not yield an H^2 bound on $\log u$. However, we are able to show that, for periodic functions with positive lower bound,

$$\frac{1}{4}\int_{\mathbb{T}^d} u\|\nabla^2 \log u\|^2 dx \geq \kappa_1 \int_{\mathbb{T}^d} \|\nabla^2\sqrt{u}\|^2 dx, \quad \text{where } \kappa_1 = \frac{4d-1}{d(d+2)}, \tag{1.3}$$

leading to an H^2 bound for $\sqrt{u}$. This inequality was proved independently in [13, equation (1.82)]; we will show a more general version below (see Lemma 2.2). This motivates rewriting the nonlinearity in (1.1) in terms of $\sqrt{u}$, yielding the following (formally) equivalent equation:

$$\partial_t u + \partial^2_{ij}\big(\sqrt{u}\partial^2_{ij}\sqrt{u} - \partial_i\sqrt{u}\partial_j\sqrt{u}\big) = 0, \quad x \in \mathbb{T}^d,\ t > 0, \tag{1.4}$$

$$u(0,x) = u_0(x), \quad x \in \mathbb{T}^d. \tag{1.5}$$

The other crucial idea, which eventually provides nonnegativity of u, is an exponential variable transformation. To be more precise, for the fixed-point argument leading to the existence result, we also work in the original formulation (1.1),

$$\partial_t u + \frac{1}{2}\partial^2_{ij}(u\partial^2_{ij} y) = 0$$

with the exponential variable $y = \log u$. In a suitable regularization regime, y is bounded in modulus, and hence $u = \exp(y)$ is strictly positive.

Our proof of the crucial inequality (1.3) and its generalizations presented below is inspired by the algorithmic entropy construction method of [16]; there, the task of deriving inequalities like (1.3) is reformulated as a decision problem for polynomial systems. The latter can be solved (at least in principle) by computer algebra systems. The solution of the decision problem determines how to perform integration by parts in a way which leads to the desired inequality. By this method, the proof of (1.3) becomes quite short and elementary. Positivity of u is needed to make the computer-aided manipulations mathematically rigorous.

Our existence result reads as follows.

THEOREM 1.1. *Let $T > 0$ and $d \leq 3$. Furthermore, let u_0 be a nonnegative measurable function on $\mathbb{T}^d$ with finite physical entropy $E_1(u_0) = \int_T (u_0(\log u_0 - 1) + 1)dx < +\infty$. Then there exists a weak solution u to (1.4)–(1.5) satisfying*

$$u(t,\cdot) \geq 0 \ a.e., \quad u \in W^{1,1}(0,T;H^{-2}(\mathbb{T}^d)), \quad \sqrt{u} \in L^2(0,T;H^2(\mathbb{T}^d)),$$

and for all $z \in L^\infty(0,T;H^2(\mathbb{T}^d))$,

$$\int_0^T \langle \partial_t u, z\rangle_{H^{-2},H^2} dt + \int_0^T \int_{\mathbb{T}^d} \big(\sqrt{u}\partial_{ij}^2 \sqrt{u} - \partial_i \sqrt{u}\partial_j \sqrt{u}\big)\partial_{ij}^2 z dx = 0.$$

The theorem is valid in the physically relevant dimensions $d \leq 3$. This restriction is related to the lack of certain Sobolev embeddings in higher dimensions $d \geq 4$. Most prominently, the fixed-point argument exploits the continuous embedding $H^2(\mathbb{T}^d) \hookrightarrow L^\infty(\mathbb{T}^d)$ to conclude absolute boundedness of y and hence strict positivity of $u = \exp(y)$. We have chosen periodic boundary conditions in order to avoid boundary integrals. For the treatment of nonhomogeneous boundary conditions of Dirichlet–Neumann-type in one space dimension, we refer the reader to [14].

Our second result concerns the long-time behavior of weak solutions to the homogeneous steady state u_∞, and generally the systematic investigation of Lyapunov functionals. More specifically, we determine a range of parameters $\gamma > 0$, for which the *entropies*

$$\widetilde{E}_\gamma = \frac{1}{\gamma(\gamma-1)} \int_{\mathbb{T}^d} \big(u(t,\cdot)^\gamma - u_\infty^\gamma\big) dx$$

monotonically decay to zero. (Recall that $\widetilde{E}_1$ is the physical entropy from (1.2).) Starting from the results of [4], Lyapunov functionals of this (and more general) type have been investigated for $d = 1$. Here, we extend the entropy construction method developed in [16] to the multidimensional case.

To prove entropy decay, we multiply (1.4) formally by $v^{2(\gamma-1)}/(\gamma-1)$, where $v = \sqrt{u}$, integrate over the torus, and integrate by parts. This leads to

$$\frac{d\widetilde{E}_\gamma}{dt} + \frac{1}{\gamma-1} \int_{\mathbb{T}^d} v^2 \partial_{ij}^2(\log v)\partial_{ij}^2(v^{2(\gamma-1)}) dx = 0.$$

Next, we need to relate the entropy production to the entropy itself. For this, inequality (1.3) is generalized. We show, by using the method of [16], that if $0 < \gamma < 2(d+1)/(d+2)$, then

$$\frac{1}{2(\gamma-1)} \int_{\mathbb{T}^d} v^2 \partial_{ij}^2(\log v)\partial_{ij}^2(v^{2(\gamma-1)}) dx \geq \kappa_\gamma \int_{\mathbb{T}^d} (\Delta v^\gamma)^2 dx, \tag{1.6}$$

where

$$\kappa_\gamma = \frac{-(d+2)^2\gamma^2 + 2(d+1)(d+2)\gamma - (d-1)^2}{\gamma^2(-(d+2)^2\gamma^2 + 2(d+1)(d+2)\gamma)}.$$

The constant κ_γ is positive if and only if $(\sqrt{d}-1)^2/(d+2) < \gamma < (\sqrt{d}+1)^2/(d+2)$. By a Beckner-type inequality, we can relate the integral of Δv^γ to the entropy itself, giving $d\widetilde{E}_\gamma/dt + c\widetilde{E}_\gamma \leq 0$ for some $c > 0$. A similar strategy works for the physical entropy, $\gamma = 1$. Eventually, Gronwall's lemma yields the following exponential decay estimates.

THEOREM 1.2. *Assume that u is either a positive classical solution to* (1.4)–(1.5) *or the weak solution constructed in the proof of Theorem* 1.1. *Let*

$$u_\infty \equiv \mathrm{meas}(\mathbb{T}^d)^{-1}\int_{\mathbb{T}^d} u_0\,dx > 0.$$

Then the entropies decay exponentially fast,

$$\widetilde{E}_\gamma(u(t,\cdot)) \le \widetilde{E}_\gamma(u_0)\exp(-16\pi^4\gamma^2\kappa_\gamma t) \quad \textit{for} \quad 1 \le \gamma < \frac{(\sqrt{d}+1)^2}{d+2},$$

and the solution itself decays exponentially in the L^1 norm,

$$\|u(t,\cdot)-u_\infty\|_{L^1(\mathbb{T}^d)} \le (2\widetilde{E}_1(u_0))^{1/2}\exp(-8\pi^4\kappa_1 t).$$

In order to make the above inequalities rigorous, we consider a regularized semidiscrete version of (1.4) for which we obtain positive H^2 solutions. Since the fourth-order differential operator in (1.1) is not strictly elliptic in $y = \log u$ ($u = 0$ may be possible), we add the regularization $-\varepsilon(\Delta^2 y + y)$ for $\varepsilon > 0$ to the right-hand side of (1.1). Unfortunately, this regularization destroys the dissipative structure of the DLSS equation, and we cannot prove anymore the entropy-entropy production inequality for $\gamma \neq 1$. To cure this problem, we need to add the expression

$$\varepsilon\mathrm{div}(|\nabla \log\max\{v,\mu\}|^2\nabla y) \quad \text{for some } \mu > 0.$$

The third main result of this paper concerns the nonuniqueness of solutions. We show that, for a family of particular initial data, there exist at least two solutions to (1.4)–(1.5) in the class of nonnegative functions in $L^1(0,T;H^2(\mathbb{T}^d))$ with finite physical entropy $\widetilde{E}_1$. Recall that uniqueness holds in the class of positive smooth functions [4]. Here, the initial data are chosen in such a way that they vanish on a set of measure zero and that they represent classical solutions to the stationary and, hence, to the transient equation. On the other hand, our existence result provides a solution which converges to the homogeneous positive steady state u_∞. Therefore, this solution is not equal to the first one. This observation may give a criterium how to choose the physically relevant solution: it should dissipate the physical entropy.

Throughout this paper, we make the following simplification. Due to the scaling invariance of (1.5) with respect to $x \to \xi x$, $t \to \xi^4 t$, and $u \to \eta u$ for ξ, $\eta > 0$, we may assume that the torus $\mathbb{T}^d$ is normalized, $\mathbb{T}^d \cong [0,1]^d$. We further assume that the initial datum has unit mass, $\int_{\mathbb{T}^d} u_0 dx = 1$; notice that the DLSS equation is mass preserving.

The paper is organized as follows. In section 2 we show some inequalities needed for the analysis of the DLSS equation. In particular we prove (1.3) and (1.6). Theorem 1.2 is proved in section 3 for smooth positive solutions. Then the existence of solutions is shown in section 4. Section 5 is devoted to the proof of Theorem 1.2 for weak solutions. Finally, in section 6 the nonuniqueness result is presented.

2. Some inequalities. We collect some inequalities which are needed in the following sections. We start with a lower bound on the Euclidean norm of a matrix. Let $A = (a_{ij}) \in \mathbb{R}^{d\times d}$ be a matrix and let $a \in \mathbb{R}^d$ be a vector. We define the Euclidean norm of A and a, respectively, by $\|A\|^2 = \sum_{i,j} a_{ij}^2$ and $\|a\|^2 = \sum_j a_j^2$. Furthermore, $\mathrm{tr}\, A = \sum_j a_{jj}$ is the trace of A and

$$A : (a)^2 = \sum_{i,j=1}^d a_{ij}a_i a_j.$$

LEMMA 2.1. *Let $A \in \mathbb{R}^{d\times d}$ be a real symmetric matrix and let $a \in \mathbb{R}^d$ be a nonzero vector. Then*

$$\|A\|^2 \geq \frac{1}{d}(\operatorname{tr} A)^2 + \frac{d}{d-1}\left(\frac{A:(a)^2}{\|a\|^2} - \frac{\operatorname{tr} A}{d}\right)^2. \tag{2.1}$$

Proof. Since A is real and symmetric, one can assume (by the spectral theorem), without loss of generality, that A is a diagonal matrix, $A = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$; recall that norms and traces are invariant under orthogonal transformations. Furthermore, one can also assume, by homogeneity of (2.1), that $a = (a_1, \dots, a_d)^\top$ is a unit vector, $\sum_j a_j^2 = 1$. Thus, inequality (2.1) becomes

$$\frac{1}{d}\sum_{j=1}^d \lambda_j^2 - \left(\frac{1}{d}\sum_{j=1}^d \lambda_j\right)^2 \geq \frac{1}{d-1}\left(\sum_{j=1}^d \lambda_j a_j^2 - \frac{1}{d}\sum_{j=1}^d \lambda_j\right)^2. \tag{2.2}$$

Set $\sigma = \sum_{j=1}^d \lambda_j/d$ and $\rho_j = \lambda_j - \sigma$. Then (2.2) is equivalent to

$$\frac{1}{d}\sum_{j=1}^d \rho_j^2 \geq \frac{1}{d-1}\left(\sum_{j=1}^d \rho_j a_j^2\right)^2, \tag{2.3}$$

where $\sum_{j=1}^d \rho_j = 0$. Without loss of generality, we assume that ρ_d is the term with maximal modulus. We employ the identity $\sum_{j=1}^{d-1}\rho_j = -\rho_d$ and the elementary inequalities

$$\sum_{j=1}^{d-1}\rho_j^2 \geq \frac{1}{d-1}\left(\sum_{j=1}^{d-1}\rho_j\right)^2 \quad \text{and} \quad \rho_d^2 \geq \left(\sum_{j=1}^d \rho_j a_j^2\right)^2$$

to obtain

$$\begin{aligned}\frac{1}{d}\sum_{j=1}^d \rho_j^2 &= \frac{1}{d}\rho_d^2 + \frac{1}{d}\sum_{j=1}^{d-1}\rho_j^2 \geq \frac{1}{d}\rho_d^2 + \frac{1}{d(d-1)}\left(\sum_{j=1}^{d-1}\rho_j\right)^2 \\ &= \frac{1}{d}\rho_d^2 + \frac{1}{d(d-1)}\rho_d^2 = \frac{1}{d-1}\rho_d^2 \geq \frac{1}{d-1}\left(\sum_{j=1}^d \rho_j a_j^2\right)^2.\end{aligned}$$

This shows (2.3) and finishes the proof. $\square$

The main result of this section is the following inequality.

LEMMA 2.2. *Let $v \in H^2(\mathbb{T}^d) \cap W^{1,4}(\mathbb{T}^d) \cap L^\infty(\mathbb{T}^d)$ in dimension $d \geq 2$, and assume that $\inf_{\mathbb{T}^d} v > 0$. Then, for any $0 < \gamma < 2(d+1)/(d+2)$,*

$$\frac{1}{2(\gamma-1)}\int_{\mathbb{T}^d} v^2 \partial_{ij}^2(\log v)\partial_{ij}^2(v^{2(\gamma-1)})dx \geq \kappa_\gamma \int_{\mathbb{T}^d}(\Delta v^\gamma)^2 dx \tag{2.4}$$

if $\gamma \neq 1$, or

$$\int_{\mathbb{T}^d} v^2 \partial_{ij}^2(\log v)^2 dx \geq \kappa_1 \int_{\mathbb{T}^d}(\Delta v)^2 dx \tag{2.5}$$

if $\gamma = 1$*, respectively, where*

$$\kappa_\gamma = \frac{p(\gamma)}{\gamma^2(p(\gamma)-p(0))} \quad \textit{and} \quad p(\gamma) = -\gamma^2 + \frac{2(d+1)}{d+2}\gamma - \Big(\frac{d-1}{d+2}\Big)^2. \tag{2.6}$$

The function $\nabla^2 v$ denotes the Hessian of v. By Sobolev embedding, it is sufficient to assume $v \in H^2(\mathbb{T}^d)$ in space dimensions $d \le 3$. The condition $0 < \gamma < 2(d+1)/(d+2)$ ensures that $p(\gamma) > p(0)$ so that κ_γ is well defined; if the stronger condition $(\sqrt{d}-1)^2/(d+2) < \gamma < (\sqrt{d}+1)^2/(d+2)$ holds, then $\kappa_\gamma > 0$. Finally, we remark that the method of [16] directly applies to the one-dimensional situation, yielding (2.4) and (2.5), respectively, for $0 \le \gamma \le \frac{3}{2}$, with $\kappa_\gamma = \min(\gamma, 12-8\gamma)/\gamma^3$.

Proof. In order to simplify the computations, we introduce the functions θ, λ, and μ, respectively, by (recall that $v > 0$)

$$\theta = \frac{|\nabla v|}{v}, \quad \lambda = \frac{1}{d}\frac{\Delta v}{v}, \quad (\lambda+\mu)\theta^2 = \frac{1}{v^3}\nabla^2 v : (\nabla v)^2,$$

and $\rho \ge 0$ by

$$\|\nabla^2 v\|^2 = \Big(d\lambda^2 + \frac{d}{d-1}\mu^2 + \rho^2\Big)v^2.$$

We need to show that ρ is well defined. But this is clear since

$$\|\nabla^2 v\|^2 \ge \Big(d\lambda^2 + \frac{d}{d-1}\mu^2\Big)v^2$$

follows directly from (2.1) after taking $A = \nabla^2 v$ and $a = \nabla v$.

We compute the left-hand side of (2.4),

$$\begin{aligned} J &= \frac{1}{2(\gamma-1)}\int_{\mathbb{T}^d} v^2 \partial_{ij}^2(\log v)\partial_{ij}^2(v^{2(\gamma-1)})dx \\ &= \frac{1}{2(\gamma-1)}\int_{\mathbb{T}^d} (v\partial_{ij}^2 v - \partial_i v \partial_j v)\partial_{ij}^2(v^{2(\gamma-1)})dx \\ &= \int_{\mathbb{T}^d} (v\partial_{ij}^2 v - \partial_i v\partial_j v)v^{2(\gamma-2)}\big(v\partial_{ij}^2 v + (2\gamma-3)\partial_i v\partial_j v\big)dx \\ &= \int_{\mathbb{T}^d} v^{2\gamma}\left(\frac{\|\nabla^2 v\|^2}{v^2} - 2(2-\gamma)\frac{\nabla^2 v}{v^2} : \Big(\frac{\nabla v}{v}\Big)^2 + (3-2\gamma)\frac{|\nabla v|^4}{v^4}\right)dx, \end{aligned}$$

and express it in terms of the functions θ, λ, μ, and ρ defined above,

$$J = \int_{\mathbb{T}^d} v^{2\gamma}\Big(d\lambda^2 + \frac{d}{d-1}\mu^2 + \rho^2 - 2(2-\gamma)(\lambda+\mu)\theta^2 + (3-2\gamma)\theta^4\Big)dx.$$

This integral is compared to

$$\begin{aligned} K &= \frac{1}{\gamma^2}\int_{\mathbb{T}^d}(\Delta v^\gamma)^2 dx = \int_{\mathbb{T}^d} v^{2(\gamma-2)}\big(v\Delta v + (\gamma-1)|\nabla v|^2\big)^2 dx \\ &= \int_{\mathbb{T}^d} v^{2\gamma}\big(d\lambda + (\gamma-1)\theta^2\big)^2 dx. \end{aligned}$$

More precisely, we shall determine a constant $c_0 > 0$ independent of v such that $J - c_0 K \ge 0$ for all (positive) functions v. Our strategy is an adaption of the method developed in [16]. We formally perform integration by parts in the expression $J - c_0K$ by adding a linear combination of certain "dummy" integrals—which are actually zero and hence do not change the value of $J - c_0K$. The coefficients in the linear combination are determined in such a way that makes the resulting integrand pointwise nonnegative. The latter is a decision problem from real algebraic geometry, and it is solved with computer aid.

We shall rely on the following two "dummy" integral expressions:

$$J_1 = \int_{\mathbb{T}^d} \operatorname{div}\big(v^{2\gamma-2}(\nabla^2 v - \Delta v \mathbb{I}) \cdot \nabla v\big)dx,$$
$$J_2 = \int_{\mathbb{T}^d} \operatorname{div}\big(v^{2\gamma-3}|\nabla v|^2 \nabla v\big)dx,$$

where $\mathbb{I}$ is the unit matrix in $\mathbb{R}^{d\times d}$. Clearly, in view of the periodic boundary conditions, $J_1 = J_2 = 0$. The goal is to find constants c_0, c_1, and c_2 such that $J - c_0K = J - c_0K + c_1J_1 + c_2J_2 \geq 0$; moreover, c_0 should be as large as possible. Since, with the above notations,

$$\begin{aligned} J_1 &= \int_{\mathbb{T}^d} v^{2(\gamma-2)}\big(v^2(\|\nabla^2 v\|^2 - (\Delta v)^2) + 2(\gamma-1)v(\nabla^2 v - \Delta v\mathbb{I}) : (\nabla v)^2\big)dx \\ &= \int_{\mathbb{T}^d} v^{2\gamma}\Big(- d(d-1)\lambda^2 + \frac{d}{d-1}\mu^2 + \rho^2 + 2(\gamma-1)(-(d-1)\lambda\theta^2 + \mu\theta^2)\Big)dx \end{aligned}$$

and

$$\begin{aligned} J_2 &= \int_{\mathbb{T}^d} v^{2(\gamma-2)}\big((2\nabla^2 v + \Delta v\mathbb{I}) : (\nabla v)^2 + (2\gamma-3)|\nabla v|^4\big)dx \\ &= \int_{\mathbb{T}^d} v^{2\gamma}\big((d+2)\lambda\theta^2 + 2\mu\theta^2 + (2\gamma-3)\theta^4\big)dx, \end{aligned}$$

we obtain

$$\begin{aligned} J - c_0K + c_1J_1 + c_2J_2 = \int_{\mathbb{T}^d} & v^{2\gamma}\big\{d\lambda^2\big[1 - dc_0 - (d-1)c_1\big] \\ & + \lambda\theta^2\big[2(\gamma-1)(1 - dc_0 - (d-1)c_1) + (d+2)c_2 - 2\big] + Q(\theta,\mu,\rho)\big\}dx, \end{aligned} \tag{2.7}$$

where Q is a polynomial in θ, μ, and ρ with coefficients depending on c_0, c_1, and c_2 but not on λ. We choose to eliminate λ from the above integrand by defining c_1 and c_2 appropriately. The linear system

$$\begin{aligned} 1 - dc_0 - (d-1)c_1 &= 0, \\ 2(\gamma-1)(1 - dc_0 - (d-1)c_1) + (d+2)c_2 - 2 &= 0 \end{aligned}$$

has the solution $c_1 = (1 - dc_0)/(d-1)$ and $c_2 = 2/(d+2)$. With this choice, the polynomial Q in (2.7) reads as

$$Q(\theta,\mu,\rho) = \frac{1}{(d-1)^2(d+2)}\big(b_1\mu^2 + 2b_2\mu\theta^2 + b_3\theta^4 + b_4\rho^2\big),$$

where

$$\begin{aligned} b_1 &= d^2(d+2)(1-c_0), \\ b_2 &= d(d-1)\big((d+2)(\gamma + c_0(1-\gamma)) - 2d - 1\big), \\ b_3 &= (d-1)^2\big(d(3-2\gamma) - c_0(d+2)(\gamma-1)^2\big), \\ b_4 &= d(d+2)(d-1)(1-c_0). \end{aligned}$$

If $c_0 \le 1$, then $b_4 \ge 0$. We wish to choose $c_0 \le 1$ in such a way that the remaining sum $b_1\mu^2 + 2b_2\mu\theta^2 + b_3\theta^4$ is nonnegative as well for any μ and θ. This is the case if (i) $b_1 > 0$ and (ii) $b_1 b_3 - b_2^2 \ge 0$. Condition (ii) is equivalent to

$$\begin{aligned} 0 &\le (1-c_0)(d+2)\big(2(d+1)\gamma - (d+2)\gamma^2\big) - (d-1)^2 \\ &= (1-c_0)(d+2)^2(p(\gamma)-p(0)) - (d-1)^2, \end{aligned}$$

which is further equivalent to (recall that $p(\gamma) > p(0)$ on the considered range of γ's)

$$c_0 \le \frac{p(\gamma)}{p(\gamma)-p(0)}.$$

The best choice for c_0 is obviously to make it equal to the right-hand side. As $p(0) < 0$, one has in particular that $c_0 < 1$, so condition (i) is satisfied as well. Thus we have found constants c_0, c_1, and c_2 for which the expression $J - c_0 K + c_1 J_1 + c_2 J_2$ is nonnegative. With $\kappa_\gamma = c_0/\gamma^2$, Lemma 2.2 is proven. ∎

Remark 2.3. Elimination of λ from the integrand in (2.7) is clearly not the only strategy to initiate the polynomial reduction process. However, from numerical studies of the multivariate polynomial, there is strong evidence that this strategy leads to the optimal values for c_0, at least for γ close to one.

As a consequence of Lemma 2.2 for $v = \sqrt{u}$ and $\gamma = 1$, we obtain inequality (1.3) which connects the entropy production of (1.4) to the smoothness of its solution.

LEMMA 2.4. *For all $d \ge 1$ and all strictly positive functions u such that $\sqrt{u} \in H^2(\mathbb{T}^d) \cap L^\infty(\mathbb{T}^d)$ it holds that*

$$\frac{1}{4}\int_{\mathbb{T}^d} u\|\nabla^2 \log u\|^2 dx \ge \kappa_1 \int_{\mathbb{T}^d} \|\nabla^2\sqrt{u}\|^2 dx, \quad \textit{where } \kappa_1 = \frac{4d-1}{d(d+2)}.$$

We also need the following generalized convex Sobolev inequalities.

LEMMA 2.5. *Let $f \in H^2(\mathbb{T}^d)$ be nonnegative. Then, for $1 < p \le 2$,*

$$\frac{p}{p-1}\left(\int_{\mathbb{T}^d} f^2 dx - \left(\int_{\mathbb{T}^d} f^{2/p} dx\right)^p\right) \le \frac{1}{8\pi^4}\int_{\mathbb{T}^d} (\Delta f)^2 dx. \tag{2.8}$$

Furthermore,

$$\int_{\mathbb{T}^d} f^2 \log\big(f^2/\|f\|_{L^2}^2\big) dx \le \frac{1}{8\pi^4}\int_{\mathbb{T}^d} (\Delta f)^2 dx. \tag{2.9}$$

Inequality (2.9) represents the limit of (2.8) as $p \searrow 1$. Unfortunately, (2.8) does seemingly *not* generalize for parameters $0 < p < 1$ in dimensions $d > 1$. The reason is that the functional on the left-hand side is convex in f only if $1 \le p \le 2$; see the discussion in [5]. This limits our decay estimates to entropies E_γ with $\gamma \ge 1$.

Proof. We only prove inequality (2.8) as (2.9) follows in a completely analogous manner. The estimate is a consequence of the Beckner-type inequality and the Poincaré inequality. For the one-dimensional torus $\mathbb{T}$, the former reads as (see [3])

$$\frac{p}{p-1}\left(\int_{\mathbb{T}} f^2 dx_j - \left(\int_{\mathbb{T}} f^{2/p} dx_j\right)^p\right) \le \frac{1}{2\pi^2}\int_{\mathbb{T}} |\partial_j f|^2 dx_j$$

(see, e.g., [12] for an easy proof). In several space dimensions, we obtain the same result since the above inequality *tensorizes*. Indeed, by employing the relation

$$\int_{\mathbb{T}^d} f^2 dx - \left(\int_{\mathbb{T}^d} f^{2/p} dx\right)^p \le \sum_{j=1}^d \int_{\mathbb{T}^d}\left(\int_0^1 f^2 dx_j - \left(\int_0^1 f^{2/p} dx_j\right)^p\right) dx$$

from Proposition 4.1 in [21], it follows that

$$\frac{p}{p-1}\left(\int_{\mathbb{T}^d} f^2 dx - \left(\int_{\mathbb{T}^d} f^{2/p} dx\right)^p\right) \le \frac{1}{2\pi^2}\sum_{j=1}^d \int_{\mathbb{T}^d}\left(\int_0^1 (\partial_j f)^2 dx_j\right) dx$$
$$= \frac{1}{2\pi^2}\int_{\mathbb{T}^d} |\nabla f|^2 dx.$$

Now, Poincaré's inequality for multiperiodic functions with zero mean,

$$\int_{\mathbb{T}^d} |\nabla f|^2 dx \le \frac{1}{4\pi^2}\int_{\mathbb{T}^d} \|\nabla^2 f\|^2 dx = \frac{1}{4\pi^2}\int_{\mathbb{T}^d} (\Delta f)^2 dx,$$

gives the assertion. ◻

In section 4 we frequently refer to the Gagliardo–Nirenberg inequalities which we recall for convenience [15].

Lemma 2.6. *Let m, $k \in \mathbb{N}_0$ with $0 \le k \le m$, $0 \le \theta < 1$, and $1 \le p, q, r \le \infty$. If both*

$$k - \frac{d}{p} \le \theta\left(m - \frac{d}{q}\right) + (1-\theta)\left(-\frac{d}{r}\right) \quad \text{and} \quad \frac{1}{p} \le \frac{\theta}{q} + \frac{1-\theta}{r},$$

then any function $f \in W^{m,q}(\mathbb{T}^d) \cap L^r(\mathbb{T}^d)$ belongs to $W^{k,p}(\mathbb{T}^d)$, and there exists a constant $C > 0$ independent of f such that

$$\|f\|_{W^{k,p}} \le C\|f\|_{W^{m,q}}^{\theta}\|f\|_{L^r}^{1-\theta}. \tag{2.10}$$

If additionally $k \ge 1$, we conclude from (2.10) that

$$\|\nabla^k f\|_{L^p} \le C\|\nabla^m f\|_{L^q}^{\theta}\|f\|_{L^r}^{1-\theta} \tag{2.11}$$

by means of the Poincaré inequality.

3. Decay rates for smooth positive solutions. We show Theorem 1.2 first for smooth positive solutions by using Lemmas 2.2 and 2.5. The proof for weak solutions is based on estimates for the semidiscrete, regularized problem and is therefore presented later in section 5. Both proofs are identical in their structure, but the proof for smooth solutions is stripped of the technicalities that are introduced by the regularization process.

The essential tool to derive the a priori estimates are the so-called *relative entropies*, as introduced in [1],

$$E_\gamma(u_1|u_2) = \int_{\mathbb{T}^d} \phi_\gamma\Big(\frac{u_1}{u_2}\Big) u_2 dx, \quad \gamma \notin \{0,1\},$$

where u_1 and u_2 are nonnegative functions on $\mathbb{T}^d$ with unit mean value, and ϕ_γ is given by

$$\phi_\gamma(s) = \frac{1}{\gamma(\gamma-1)}\big(s^\gamma - \gamma s + \gamma - 1\big), \quad s \ge 0. \tag{3.1}$$

The natural continuation for $\gamma = 1$ is $\phi_1(s) = s(\log s - 1) + 1$; the functional E_1 corresponds to the physical entropy. The functions ϕ_γ are nonnegative and convex and attain their minimal value at $s = 1$. Consequently, E_γ is nonnegative (possibly $+\infty$) and vanishes if and only if $u_1 = u_2$.

To obtain the a priori estimates (and the decay rates), we consider entropies of solutions $u_1 = u$ relative to the spatial homogeneous steady state $u_2 \equiv 1$:

$$E_\gamma(u(t,\cdot)) = \frac{1}{\gamma(\gamma-1)}\left(\int_{\mathbb{T}^d} u(t,x)^\gamma dx - 1\right), \quad \gamma \geq 1. \tag{3.2}$$

We obtain the following entropy-entropy production estimate.

PROPOSITION 3.1. *Assume that* u *is a smooth positive solution to* (1.4)–(1.5) *in dimensions* $d \geq 2$. *Then*

$$\frac{dE_\gamma}{dt} + 2\kappa_\gamma \int_{\mathbb{T}^d} (\Delta u^{\gamma/2})^2 dx \leq 0 \quad \textit{for} \quad 0 < \gamma < \frac{2(d+1)}{d+2}, \tag{3.3}$$

where $\kappa_\gamma > 0$ *is defined in* (2.6).

Proof. For convenience, we work with the function $v = \sqrt{u}$ instead of u. If $\gamma \neq 1$, we integrate the DLSS equation (1.4) against the test function $v^{2(\gamma-1)}/2(\gamma-1)$. Then we obtain for the time derivative

$$\frac{1}{2(\gamma-1)}\int_{\mathbb{T}^d} \partial_t(v^2)v^{2(\gamma-1)}dx = \frac{1}{2\gamma(\gamma-1)}\int_{\mathbb{T}^d}\partial_t(v^{2\gamma})dx = \frac{1}{2}\,\frac{dE_\gamma}{dt}.$$

In combination with Lemma 2.2,

$$\frac{1}{2}\,\frac{dE_\gamma}{dt} = -\frac{1}{2(\gamma-1)}\int_{\mathbb{T}^d} v^2\partial_{ij}^2(\log v)\partial_{ij}^2(v^{2(\gamma-1)})dx \leq -\kappa_\gamma\int_{\mathbb{T}^d}(\Delta v^\gamma)^2 dx.$$

If $\gamma = 1$, we use the test function $\log v$ instead. $\square$

Remark 3.2. As pointed out before, the coefficient function κ_γ follows a different law in dimension $d = 1$; see the remarks after Lemma 2.2. In conclusion, the entropy production is estimated as

$$\frac{dE_\gamma}{dt} + \frac{2\mu_\gamma}{\gamma^2}\int_{\mathbb{T}} |(u^{\gamma/2})_{xx}|^2 dx \leq 0,$$

where

$$\mu_\gamma = \begin{cases} 1 & \text{for } 0 < \gamma < 4/3, \\ 12/\gamma - 8 & \text{for } 4/3 < \gamma < 3/2. \end{cases}$$

Estimates for the limiting case $\gamma = 0$ are also available (see Theorem 3 in [7]).

The proof of Theorem 1.2 in the case of smooth positive solutions is immediate: applying Lemma 2.5 to $f = u^{\gamma/2}$ with $p = \gamma$ and taking into account that u has unit mass, we obtain from Proposition 3.1

$$\frac{dE_\gamma}{dt} + (2\pi)^4\gamma^2\kappa_\gamma E_\gamma \leq 0.$$

Gronwall's lemma shows the entropy decay. The decay in the L^1 norm is a straightforward consequence of the Csiszár–Kullback inequality [9, 20].

The values of κ_γ as a function of γ are plotted in Figure 3.1 (left). For $\gamma \geq 1$, these correspond to exponential decay rates of the respective entropy E_γ. (There is no immediate interpretation of κ_γ for $0 < \gamma < 1$.) The right figure shows the decay

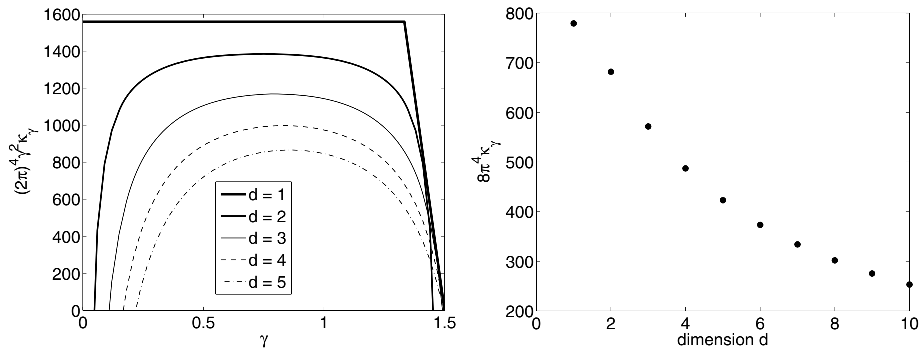


FIG. 3.1. *Decay rates for the entropy E_γ (left) and in the L^1 norm for $\gamma = 1$ (right) depending on the dimension d.*

rate $8\pi^4\kappa_1$ in the L^1 norm for $\gamma = 1$ as a function of the dimension d. This rate is given by

$$8\pi^4\kappa_1 = 8\pi^4\frac{4d-1}{d(d+2)};$$

this is slightly better than the rate obtained in [13], which amounts to $24\pi^4/(d+2)$ for (1.1).

4. Existence of solutions. In this section we prove Theorem 1.1. The proof is divided into a series of lemmas. We continue to use $v = \sqrt{u}$ for easier notation.

4.1. Existence of a time-discrete solution. Let $T > 0$ be a terminal time and $\tau > 0$ a time step. Let w be a given function. We wish to find a solution $v \in H^2(\mathbb{T}^d)$ to the semidiscrete equation

$$\frac{1}{\tau}(v^2 - w^2) = -\partial_{ij}^2\big(v\partial_{ij}^2 v - \partial_i v\partial_j v\big). \tag{4.1}$$

LEMMA 4.1. *Let $d \le 3$. Assume that w is a nonnegative measurable function on $\mathbb{T}^d$ with finite entropy $E_1(w^2) < +\infty$ and unit mass $\int_{\mathbb{T}^d} w^2 dx = 1$. Then there exists a nonnegative weak solution $v \in H^2(\mathbb{T}^d)$ to (4.1). Furthermore, v^2 has unit mass, the physical entropy is dissipated in the sense*

$$E_1(v^2) + 2\tau\kappa_1\int_{\mathbb{T}^d}\|\nabla^2 v\|^2 dx \le E_1(w^2), \tag{4.2}$$

and the entropies $E_\gamma(v^2)$ and $E_\gamma(w^2)$ are related by

$$(1 + 16\pi^4\tau\gamma^2\kappa_\gamma)E_\gamma(v^2) \le E_\gamma(w^2), \tag{4.3}$$

where $1 \le \gamma < (\sqrt{d}+1)^2/(d+2)$ and κ_γ is defined in (2.6).

Proof.

Step 1. Definition of the regularized problem. The solution to (4.1) is obtained as the limit of solutions to a regularized problem. For this, recall that (1.4) can be written as

$$\partial_t(v^2) = -\frac{1}{2}\partial_{ij}^2(v^2\partial_{ij}^2 y) \quad \text{with } y = \log(v^2).$$

We regularize (4.1) in the above formulation by adding a strongly elliptic operator in y:

$$\frac{1}{\tau}(v^2 - w^2) = -\frac{1}{2}\partial_{ij}^2(v^2\partial_{ij}^2 y) - \varepsilon(\Delta^2 y + y) + \varepsilon\mathrm{div}\big(|\nabla \log[v]_\mu|^2\nabla y\big), \tag{4.4}$$

where ε, $\mu > 0$ are regularization parameters and $[v]_\mu = \max\{v, \mu\}$. The fourth-order operator $\varepsilon(\Delta^2 y + y)$ guarantees coercivity of the above right-hand side with respect to y. The nonlinear second-order operator allows to derive the a priori estimates for the general entropy E_γ.

Step 2. Solution of the regularized problem. In order to solve (4.4) we employ the Leray–Schauder fixed-point theorem (see Theorem B.5 in [23]). Let $\sigma \in [0,1]$ and $\bar{v} \in W^{1,4}(\mathbb{T}^d) \hookrightarrow L^\infty(\mathbb{T}^d)$, and introduce for y, $z \in H^2(\mathbb{T}^d)$,

$$\begin{aligned} a(y,z) &= \frac{1}{2}\int_{\mathbb{T}^d} \bar{v}^2\partial_{ij}^2 y\partial_{ij}^2 z dx + \varepsilon\int_{\mathbb{T}^d}(\Delta y\Delta z + yz + |\nabla \log[\bar{v}]_\mu|^2\nabla y\cdot\nabla z)dx, \\ f(z) &= \frac{\sigma}{\tau}\langle \bar{v}^2 - w^2, z\rangle_{H^{-2},H^2}. \end{aligned}$$

Since $\bar{v} \in W^{1,4}(\mathbb{T}^d)$, also $\log[\bar{v}]_\mu \in W^{1,4}(\mathbb{T}^d)$, hence $|\nabla \log[\bar{v}]_\mu|^2\nabla y\cdot\nabla z$ is integrable. The bilinear form a is continuous and coercive since, by the Gagliardo–Nirenberg inequality (2.11),

$$a(y,y) \ge \varepsilon\int_{\mathbb{T}^d}\big((\Delta y)^2 + y^2\big)dx \ge C\varepsilon\|y\|_{H^2}^2. \tag{4.5}$$

Moreover, w^2 has finite physical entropy, so $w^2 \in L^1(\mathbb{T}^d) \hookrightarrow H^{-2}(\mathbb{T}^d)$ in space dimensions $d \le 3$, yielding continuity of the linear form f. Consequently, Lax–Milgram's lemma provides the existence of a unique solution to

$$-a(y,z) = f(z) \quad \text{for all } z \in H^2(\mathbb{T}^d).$$

Define the fixed-point operator $S : W^{1,4}(\mathbb{T}^d)\times[0,1] \to W^{1,4}(\mathbb{T}^d)$ by $S(\bar{v},\sigma) := v = e^{y/2}$. Since $y \in H^2(\mathbb{T}^d) \hookrightarrow L^\infty(\mathbb{T}^d)$, we have indeed that $v \in H^2(\mathbb{T}^d) \hookrightarrow W^{1,4}(\mathbb{T}^d)$.

We shall now verify the hypotheses of the Leray–Schauder theorem; the latter provides a solution v of $S(v,1) = v$. The operator S is constant at $\sigma = 0$, $S(\bar{v},0) = 1$. By standard results for elliptic equations, S is continuous and compact since the embedding $H^2(\mathbb{T}^d) \hookrightarrow W^{1,4}(\mathbb{T}^d)$ is compact. It remains to show a uniform bound for all fixed points of $S(\cdot,\sigma)$. This bound is obtained from the production of the physical entropy and Lemma 2.4.

Let $v \in H^2(\mathbb{T}^d)$ be a fixed point of $S(\cdot,\sigma)$ for some $\sigma \in [0,1]$. Then v is a solution to (4.4) with σ/τ instead of $1/\tau$, and with $v = e^{y/2} > 0$, $y \in H^2(\mathbb{T}^d)$. Since $\phi(s) = s(\log s - 1) + 1$ is convex, $\phi(s_1) - \phi(s_2) \le \phi'(s_1)(s_1 - s_2)$ for all s_1, $s_2 \ge 0$. Hence,

$$\begin{aligned} \frac{\sigma}{\tau}(E_1(v^2) - E_1(w^2)) &= \frac{\sigma}{\tau}\int_{\mathbb{T}^d}(\phi(v^2) - \phi(w^2))dx \\ &\le \frac{\sigma}{2\tau}\int_{\mathbb{T}^d}(v^2 - w^2)\log(v^2)dx = -a(y,y) \\ &\le -\frac{1}{2}\int_{\mathbb{T}^d} v^2\|\nabla^2\log(v^2)\|^2 dx - \varepsilon\int_{\mathbb{T}^d}((\Delta y)^2 + y^2)dx. \end{aligned} \tag{4.6}$$

The estimate of Lemma 2.4 shows that

$$\frac{\sigma}{\tau}(E_1(v^2)-E_1(w^2))+2\kappa_1\int_{\mathbb{T}^d}\|\nabla^2 v\|^2dx\le 0.$$

As a consequence,

$$E_1(v^2)\le E_1(w^2)\quad\text{and}\quad \|\nabla^2 v\|_{L^2}^2\le\frac{1}{2\tau\kappa_1}E_1(w^2).$$

In particular, $\nabla^2 v$ is uniformly bounded in $L^2(\mathbb{T}^d)$. Together with the elementary inequality $s\le\phi(s)+(e-1)$ for all $s\ge 0$, we obtain

$$\|v\|_{L^2}^2\le\int_{\mathbb{T}^d}(\phi(v^2)+e-1)dx=E_1(v^2)+e-1.$$

This means that v is uniformly bounded in $L^2(\mathbb{T}^d)$. Then the Gagliardo–Nirenberg inequality gives the desired uniform bound for v:

$$\|v\|_{H^2}^2\le C(\|\nabla^2 v\|_{L^2}^2+\|v\|_{L^2}^2)\le C\Big(1+\frac{1}{2\tau\kappa_1}\Big)E_1(w^2)+2C. \tag{4.7}$$

The Leray–Schauder fixed-point theorem provides a solution v to $S(v,1)=v$, which we denote by v_ε. Obviously, v_ε satisfies (4.4).

Step 3. Lower bound for v_ε. By construction of v_ε, there exists $y_\varepsilon\in H^2(\mathbb{T}^d)$ such that $v_\varepsilon=e^{y_\varepsilon/2}$. Going back to (4.6), we see that

$$\frac{1}{2\tau}(E_1(v_\varepsilon^2)-E_1(w^2))\le-\varepsilon\int_{\mathbb{T}^d}((\Delta y_\varepsilon)^2+y_\varepsilon^2)dx\le-\varepsilon C\|y_\varepsilon\|_{H^2}^2,$$

using the Gagliardo–Nirenberg inequality. Hence,

$$\|y_\varepsilon\|_{H^2}\le\left(\frac{E_1(w^2)}{2\varepsilon\tau C}\right)^{1/2}\le c\varepsilon^{-1/2}, \tag{4.8}$$

where $c>0$ is, here and in the following, a generic constant independent of ε. In combination with the embedding $H^2(\mathbb{T}^d)\hookrightarrow L^\infty(\mathbb{T}^d)$, this gives $\|y_\varepsilon\|_{L^\infty}\le c\varepsilon^{-1/2}$. Consequently, v_ε is strictly positive:

$$v_\varepsilon=\exp\Big(\frac{y_\varepsilon}{2}\Big)\ge\exp\Big(-\frac{c}{2\varepsilon^{1/2}}\Big)=\mu(\varepsilon)>0.$$

Thus, with $\mu:=\mu(\varepsilon)$, it holds that $[v_\varepsilon]_\mu=v_\varepsilon$, and the respective fixed point $v_\varepsilon\in H^2(\mathbb{T}^d)$ satisfies

$$\begin{aligned}\frac{1}{\tau}(v_\varepsilon^2-w^2)&=-\partial_{ij}^2(v_\varepsilon\partial_{ij}^2v_\varepsilon-\partial_i v_\varepsilon\partial_j v_\varepsilon)\\&\quad-\varepsilon(\Delta^2\log v_\varepsilon+\log v_\varepsilon)+\varepsilon\operatorname{div}(|\nabla\log v_\varepsilon|^2\nabla\log v_\varepsilon).\end{aligned} \tag{4.9}$$

Step 4. The limit $\varepsilon\to 0$. The estimate (4.7) shows that the sequence (v_ε) is bounded in $H^2(\mathbb{T}^d)$. Thus, for a subsequence which is not relabeled, $v_\varepsilon\rightharpoonup v$ weakly in $H^2(\mathbb{T}^d)$ and $v_\varepsilon\to v$ strongly in $W^{1,4}(\mathbb{T}^d)$ and $L^\infty(\mathbb{T}^d)$ as $\varepsilon\to 0$ for some $v\in H^2(\mathbb{T}^d)$. For the first expression on the right-hand side in (4.9), we thus obtain

$$v_\varepsilon\partial_{ij}^2v_\varepsilon-\partial_i v_\varepsilon\partial_j v_\varepsilon\rightharpoonup v\partial_{ij}^2v-\partial_i v\partial_j v\quad\text{weakly in }L^2(\mathbb{T}^d).$$

In order to prove that v is indeed a solution to (4.1), we verify that the expressions involving the factor ε vanish as $\varepsilon \to 0$. From the refined coercivity estimate

$$a(y_\varepsilon, y_\varepsilon) \geq \varepsilon(c\|y_\varepsilon\|_{H^2}^2 + \|\nabla y_\varepsilon\|_{L^4}^4),$$

we learn that

$$\|\nabla y_\varepsilon\|_{L^4} \leq c\varepsilon^{-1/4}.$$

In combination with (4.8), this gives

$$\begin{aligned}
&\big|\big\langle \varepsilon\big(\Delta^2 \log v_\varepsilon + \log v_\varepsilon - \operatorname{div}(|\nabla \log v_\varepsilon|^2 \nabla \log v_\varepsilon)\big), z\big\rangle_{H^{-2},H^2}\big| \\
&\quad \leq \varepsilon\big(\|\log v_\varepsilon\|_{H^2}\|z\|_{H^2} + \|\log v_\varepsilon\|_{L^2}\|z\|_{L^2} + \|\nabla \log v_\varepsilon\|_{L^4}^3\|z\|_{W^{1,4}}\big) \\
&\quad \leq c(\varepsilon^{1/2} + \varepsilon^{1/4})\|z\|_{H^2}
\end{aligned}$$

for any test function $z \in H^2(\mathbb{T}^d)$. Therefore,

$$\varepsilon\big(\Delta^2 \log v_\varepsilon + \log v_\varepsilon - \operatorname{div}(|\nabla \log v_\varepsilon|^2 \nabla \log v_\varepsilon)\big) \rightharpoonup 0 \quad \text{weakly in } H^{-2}(\mathbb{T}^d),$$

so v satisfies (4.1).

Step 5. Verification of (4.2) and (4.3). Conservation of mass follows from the weak formulation of (4.1) by using $z \equiv 1$ as a test function. From (4.6) and Lemma 2.4 it follows that

$$E_1(v_\varepsilon^2) + 2\tau\kappa_1 \int_{\mathbb{T}^d} \|\nabla^2 v_\varepsilon\|^2 dx \leq E_1(w^2).$$

In the limit $\varepsilon \to 0$, this inequality gives (4.2) since (a subsequence of) v_ε converges weakly to v in $H^2(\mathbb{T}^d)$ and the L^2 norm of the Hessian of v_ε constitutes a weakly lower semicontinuous functional on $H^2(\mathbb{T}^d)$.

Next, we prove (4.3). Recall that the solutions v_ε of the regularized equation (4.4) are strictly positive and bounded in modulus. Hence $\log v_\varepsilon$ and v_ε^p, for arbitrary exponents $p \in \mathbb{R}$, are well-defined functions in $H^2(\mathbb{T}^d)$. Using the test function $\phi_\gamma'(v_\varepsilon)/2 = (v_\varepsilon^{2(\gamma-1)} - 1)/2(\gamma-1)$ in (4.9) gives (see (3.1) for the definition of ϕ_γ)

$$\begin{aligned}
\frac{1}{2\tau}&(E_\gamma(v_\varepsilon^2) - E_\gamma(w^2)) = \frac{1}{2\tau}\int_{\mathbb{T}^d} \big(\phi_\gamma(v_\varepsilon^2) - \phi_\gamma(w^2)\big)dx \\
&\leq \frac{1}{2\tau}\int_{\mathbb{T}^d} \phi_\gamma'(v_\varepsilon^2)(v_\varepsilon^2 - w^2)dx = \frac{1}{2(\gamma-1)\tau}\int_{\mathbb{T}^d}(v_\varepsilon^2 - w^2)\partial_{ij}^2(v_\varepsilon^{2(\gamma-1)})dx \\
&= -\frac{1}{2(\gamma-1)}\int_{\mathbb{T}^d}(v_\varepsilon\partial_{ij}^2 v_\varepsilon - \partial_i v_\varepsilon \partial_j v_\varepsilon)\partial_{ij}^2(v_\varepsilon^{2(\gamma-1)})dx \\
&\quad - \frac{\varepsilon}{\gamma-1}\int_{\mathbb{T}^d}\big(\Delta(v_\varepsilon^{2(\gamma-1)})\Delta(\log v_\varepsilon) + |\nabla \log v_\varepsilon|^2 \nabla(\log v_\varepsilon)\cdot\nabla(v_\varepsilon^{2(\gamma-1)})\big)dx \\
&\quad - \frac{\varepsilon}{2(\gamma-1)}\int_{\mathbb{T}^d} v_\varepsilon^{\gamma-1}\log v_\varepsilon dx \\
&= A_1 - \varepsilon A_2 - \varepsilon A_3.
\end{aligned}$$

Now, by Lemma 2.2,

$$A_1 \leq -\kappa_\gamma \int_{\mathbb{T}^d} (\Delta u^{\gamma/2})^2 dx.$$

Furthermore, by Lemma 2.5, applied to $f = u^{\gamma/2}$ and $p = \gamma$, and since u has unit mass, we obtain

$$\frac{\gamma}{\gamma-1}\left(\int_{\mathbb{T}^d} u^\gamma dx - 1\right) \le \frac{1}{8\pi^4}\int_{\mathbb{T}^d}(\Delta u^{\gamma/2})^2 dx,$$

so finally,

$$A_1 \le -\frac{8\pi^4\gamma\kappa_\gamma}{\gamma-1}\left(\int_{\mathbb{T}^d} u^\gamma dx - 1\right) = -8\pi^4\gamma^2\kappa_\gamma E_\gamma(v_\varepsilon^2).$$

Now, we show that A_2 and A_3 are bounded from below, uniformly in $\varepsilon > 0$. This is clear for A_3 since $\gamma > 1$. The remaining integral can be written as

$$\begin{aligned} A_2 &= 2\int_{\mathbb{T}^d} v_\varepsilon^{2(\gamma-1)}\Big(\Big(\frac{\Delta v_\varepsilon}{v_\varepsilon}\Big)^2 - 2(2-\gamma)\frac{\Delta v_\varepsilon}{v_\varepsilon}\Big|\frac{\nabla v_\varepsilon}{v_\varepsilon}\Big|^2 + 2(2-\gamma)\Big|\frac{\nabla v_\varepsilon}{v_\varepsilon}\Big|^4\Big)dx \\ &= 2\int_{\mathbb{T}^d} v_\varepsilon^{2(\gamma-1)}\Big(\Big(\frac{\Delta v_\varepsilon}{v_\varepsilon} - (2-\gamma)\Big|\frac{\nabla v_\varepsilon}{v_\varepsilon}\Big|^2\Big)^2 + \gamma(2-\gamma)\Big|\frac{\nabla v_\varepsilon}{v_\varepsilon}\Big|^4\Big)dx \\ &\ge 0 \end{aligned}$$

since $\gamma < (\sqrt{d}+1)^2/(d+2) \le 3/2$. These estimates give

$$\frac{1}{\tau}(E_1(v_\varepsilon^2) - E_1(w^2)) \le -16\pi^4\gamma^2\kappa_\gamma E_\gamma(v_\varepsilon^2).$$

We pass to the limit $\varepsilon \to 0$ in this inequality. As $v_\varepsilon \to v$ strongly in $L^\infty(\mathbb{T}^d)$, integration and limit commute and we conclude that

$$\frac{1}{\tau}(E_\gamma(v^2) - E_\gamma(w^2)) \le -16\pi^4\gamma^2\kappa_\gamma E_\gamma(v^2)$$

from which (4.3) follows. This finishes the proof. ◻

4.2. A priori estimates. Let an arbitrary terminal time $T > 0$ be fixed in the following. Define the step function $v^{(\tau)} : [0,T) \to L^2(\mathbb{T}^d)$ recursively as follows. Let $v_0 = \sqrt{u_0}$, and for given $k \in \mathbb{N}$, let $v_k \in H^2(\mathbb{T}^d)$ be the nonnegative solution (according to Lemma 4.1) to (4.1) with $w = v_{k-1}$. Now define $v^{(\tau)}(t) := v_k$ for $(k-1)\tau < t \le k\tau$. Then $v^{(\tau)}$ satisfies

$$\frac{1}{\tau}\big((v^{(\tau)})^2 - (\sigma_\tau v^{(\tau)})^2\big) = -\partial_{ij}^2\big(v^{(\tau)}\partial_{ij}^2 v^{(\tau)} - \partial_i v^{(\tau)}\partial_j v^{(\tau)}\big), \tag{4.10}$$

where σ_τ denotes the shift operator $(\sigma_\tau v^{(\tau)})(t) = v^{(\tau)}(t-\tau)$ for $\tau \le t < T$. In order to pass to the continuum limit $\tau \to 0$ in (4.10), we need the following a priori estimate.

LEMMA 4.2. *The function $v^{(\tau)}$ satisfies*

$$\|(v^{(\tau)})^2\|_{L^{11/10}(0,T;H^2(\mathbb{T}^d))} + \tau^{-1}\|(v^{(\tau)})^2 - (\sigma_\tau v^{(\tau)})^2\|_{L^{11/10}(0,T;H^{-2}(\mathbb{T}^d))} \le c, \tag{4.11}$$

where the constant $c > 0$ is independent of τ.

Proof. From Lemma 4.1 we know that

$$\|v^{(\tau)}\|_{L^\infty(0,T;L^2(\mathbb{T}^d))} = \|u_0\|_{L^1(\mathbb{T}^d)}^{1/2} = 1, \quad \|\nabla^2 v^{(\tau)}\|_{L^2(0,T;L^2(\mathbb{T}^d))} \le c.$$

In order to derive (4.11), we employ the Gagliardo–Nirenberg and Hölder inequalities. The former inequality shows that

$$
\begin{aligned}
\|v^{(\tau)}\|^{8/3}_{L^{8/3}(0,T;L^\infty(\mathbb{T}^d))} &\le C\int_0^T \|v^{(\tau)}(t,\cdot)\|^{8\theta/3}_{H^2}\|v^{(\tau)}(t,\cdot)\|^{8(1-\theta)/3}_{L^2}dt \\
&\le C\|v^{(\tau)}\|^{8(1-\theta)/3}_{L^\infty(0,T;L^2(\mathbb{T}^d))}\int_0^T \|v^{(\tau)}(t,\cdot)\|^{8\theta/3}_{H^2}dt,
\end{aligned}
\tag{4.12}
$$

where $\theta = d/4$. Since $8\theta/3 = 2d/3 \le 2$ in dimensions $d \le 3$, the right-hand side is uniformly bounded. Applying Hölder's inequality with respect to t, for $p = 9/5$ and $p' = 9/4$, we infer that

$$
\begin{aligned}
\|v^{(\tau)}\partial^2_{ij}v^{(\tau)}\|^{11/10}_{L^{11/10}(0,T;L^2(\mathbb{T}^d))} &\le C\int_0^T \|v^{(\tau)}(t,\cdot)\|^{11/10}_{H^2}\|v^{(\tau)}(t,\cdot)\|^{11/10}_{L^\infty}dt \\
&\le C\|v^{(\tau)}\|^{11/10}_{L^{11p/10}(0,T;H^2(\mathbb{T}^d))}\|v^{(\tau)}\|^{11/10}_{L^{11p'/10}(0,T;L^\infty(\mathbb{T}^d))}.
\end{aligned}
\tag{4.13}
$$

Since $11p/10 = 99/50 \le 2$ and $11p'/10 = 99/40 \le 8/3$, the right-hand side is uniformly bounded in view of the boundedness of $v^{(\tau)}$ in $L^{8/3}(0,T;L^\infty(\mathbb{T}^d))$. On the other hand, by the Gagliardo–Nirenberg inequality,

$$
\begin{aligned}
\|v^{(\tau)}\|^{16/7}_{L^{16/7}(0,T;W^{1,4}(\mathbb{T}^d))} &\le C\int_0^T \|v^{(\tau)}(t,\cdot)\|^{16\theta/7}_{H^2}\|v^{(\tau)}(t,\cdot)\|^{16(1-\theta)/7}_{L^2}dt \\
&\le C\|v^{(\tau)}\|^{16(1-\theta)/7}_{L^\infty(0,T;L^2(\mathbb{T}^d))}\|v^{(\tau)}\|^{16\theta/7}_{L^{16\theta/7}(0,T;H^2(\mathbb{T}^d))},
\end{aligned}
\tag{4.14}
$$

where $\theta = (d+4)/8$. As $16\theta/7 = 2(d+4)/7 \le 2$ in dimensions $d \le 3$, $v^{(\tau)}$ is uniformly bounded in $L^{16/7}(0,T;W^{1,4}(\mathbb{T}^d))$. As a straightforward conclusion,

$$
\begin{aligned}
\|\partial_i v^{(\tau)}\partial_j v^{(\tau)}\|^{11/10}_{L^{11/10}(0,T;L^2(\mathbb{T}^d))} &\le \int_0^T \|\nabla v^{(\tau)}(t,\cdot)\|^{22/10}_{L^4}dt \\
&\le \|v^{(\tau)}\|^{22/10}_{L^{22/10}(0,T;W^{1,4}(\mathbb{T}^d))} \le c,
\end{aligned}
\tag{4.15}
$$

since $22/10 < 16/7$. Estimates (4.13) and (4.15) together yield

$$
\begin{aligned}
&\|\nabla^2(v^{(\tau)})^2\|_{L^{11/10}(0,T;L^2(\mathbb{T}^d))} \\
&\qquad\le 2\sum_{i,j=1}^d \|v^{(\tau)}\partial_{ij}v^{(\tau)} + \partial_i v^{(\tau)}\partial_j v^{(\tau)}\|_{L^{11/10}(0,T;L^2(\mathbb{T}^d))} \le c.
\end{aligned}
$$

Moreover, by (4.12) and (4.14), since $22/10 < 8/3$ and $22/10 < 16/7$,

$$
\|\nabla(v^{(\tau)})^2\|_{L^{11/10}(0,T;L^2(\mathbb{T}^d))} \le 2\|v^{(\tau)}\|_{L^{22/10}(0,T;L^4(\mathbb{T}^d))}\|\nabla v^{(\tau)}\|_{L^{22/10}(0,T;L^4(\mathbb{T}^d))}.
$$

The right-hand side is bounded by the considerations above. This estimates the first term in (4.11). To obtain a uniform bound on the second term in (4.11), we combine again (4.13) and (4.15):

$$
\begin{aligned}
&\frac{1}{\tau}\|(v^{(\tau)})^2 - (\sigma_\tau v^{(\tau)})^2\|_{L^{11/10}(0,T;H^{-2}(\mathbb{T}^d))} \\
&\quad\le \sum_{i,j=1}^d \Big(\|v^{(\tau)}\partial^2_{ij}v^{(\tau)}\|_{L^{11/10}(0,T;L^2(\mathbb{T}^d))} + \|\partial_i v^{(\tau)}\partial_j v^{(\tau)}\|_{L^{11/10}(0,T;L^2(\mathbb{T}^d))}\Big) \le c. \qquad \square
\end{aligned}
$$

4.3. The limit $\tau \to 0$. The a priori estimates of the previous subsection are sufficient to pass to the limit $\tau \to 0$.

LEMMA 4.3. *There exists some nonnegative function $u \in W^{1,1}(0,T;H^{-2}(\mathbb{T}^d))$ with $\sqrt{u} \in L^2(0,T;H^2(\mathbb{T}^d))$ such that, for a subsequence of $(v^{(\tau)})$, which is not relabeled, as $\tau \to 0$,*

$$\begin{aligned} \frac{1}{\tau}\left((v^{(\tau)})^2 - \sigma_\tau(v^{(\tau)})^2\right) &\rightharpoonup \partial_t u \quad \textit{weakly in } L^{11/10}(0,T;H^{-2}(\mathbb{T}^d)),\\ v^{(\tau)}\partial_{ij}^2 v^{(\tau)} &\rightharpoonup \sqrt{u}\partial_{ij}^2\sqrt{u} \quad \textit{weakly in } L^1(0,T;L^2(\mathbb{T}^d)),\\ \partial_i v^{(\tau)}\partial_j v^{(\tau)} &\rightharpoonup \partial_i\sqrt{u}\partial_j\sqrt{u} \quad \textit{weakly in } L^1(0,T;L^2(\mathbb{T}^d)). \end{aligned}$$

Moreover, u is a weak solution to (1.4)–(1.5).

Proof. Estimate (4.11) allows to apply the Aubin lemma [22], showing that, up to a subsequence, $(v^{(\tau)})^2 \to u$ in $L^{11/10}(0,T;L^\infty(\mathbb{T}^d))$ as $\tau \to 0$ for some limit function u. Here, we have used that $H^2(\mathbb{T}^d)$ embeds compactly into $L^\infty(\mathbb{T}^d)$ in dimensions $d \le 3$. In particular, $(v^{(\tau)})$ converges pointwise a.e. Since obviously, $(v^{(\tau)})^2$ is nonnegative, so is u, and we can define $\sqrt{u} \in L^{22/10}(0,T;L^\infty(\mathbb{T}^d))$; note that $v^{(\tau)}$ converges strongly to $\sqrt{u}$ in this space.

Now, the first claim follows directly from (4.11) and the construction of $v^{(\tau)}$. Estimate (4.11) further yields weak convergence of $v^{(\tau)}$ in $L^2(0,T;H^2(\mathbb{T}^d))$. The weak limit necessarily coincides with $\sqrt{u}$, the strong limit from above.

By Hölder's inequality,

$$\begin{aligned} \|v^{(\tau)} - \sqrt{u}\|^2_{L^2(0,T;L^\infty(\mathbb{T}^d)} &\le \|(v^{(\tau)} - \sqrt{u})^2\|_{L^{11/10}(0,T;L^\infty(\mathbb{T}^d))} \cdot T^{1/11}\\ &\le \|(v^{(\tau)})^2 - u\|_{L^{11/10}(0,T;L^\infty(\mathbb{T}^d))} \cdot T^{1/11}. \end{aligned} \tag{4.16}$$

In the last step, we have used that $(a-b)^2 \le |a^2 - b^2|$ for arbitrary nonnegative $a, b \in \mathbb{R}$. Now, by the Gagliardo–Nirenberg and Hölder inequalities,

$$\begin{aligned} &\|\nabla(v^{(\tau)} - \sqrt{u})\|^2_{L^2(0,T;L^4(\mathbb{T}^d))}\\ &\qquad \le C\|v^{(\tau)} - \sqrt{u}\|_{L^2(0,T;H^2(\mathbb{T}^d))}\|v^{(\tau)} - \sqrt{u}\|_{L^2(0,T;L^\infty(\mathbb{T}^d))}. \end{aligned}$$

The first term in the product is bounded (cf. estimate (4.2)); the second term converges to zero by (4.16) above. Thus $v^{(\tau)} \to \sqrt{u}$ strongly in $L^2(0,T;W^{1,4}(\mathbb{T}^d))$ and

$$\partial_i v^{(\tau)}\partial_j v^{(\tau)} \rightharpoonup \partial_i\sqrt{u}\partial_j\sqrt{u} \quad \text{weakly in } L^1(0,T;L^2(\mathbb{T}^d)).$$

The remaining limit follows from (4.16) and weak convergence of $v^{(\tau)}$ to $\sqrt{u}$ in $L^2(0,T;H^2(\mathbb{T}^d))$. Finally, since $L^2(0,T;H^2(\mathbb{T}^d)) \hookrightarrow L^2(0,T;L^\infty(\mathbb{T}^d))$, one verifies that $u = \sqrt{u}\cdot\sqrt{u} \in L^1(0,T;H^2(\mathbb{T}^d))$ by the Hölder and Gagliardo–Nirenberg estimates. ∎

5. Decay rates for nonnegative weak solutions. We prove Theorem 1.2 for the solutions constructed in the previous section.

First, we show that $\kappa_\gamma > 0$ for $1 \le \gamma < (\sqrt{d}+1)^2/(d+2)$. Indeed, by definition, $\kappa_\gamma > 0$ if $p(\gamma) > 0$, with the quadratic polynomial $p(\gamma)$ given in (2.6). But $p(\gamma) > 0$ if and only if $\gamma_- < \gamma < \gamma_+$ where $\gamma_\pm$ are the two roots of p. Now, a computation yields $\gamma_\pm = (\sqrt{d} \pm 1)^2/(d+2)$, and it is immediately seen that $\gamma_- < 1 < \gamma_+$.

Next, set $t_n = n\tau$ for $n = 0, \dots, M$. From (4.3) we know that

$$E_\gamma\big(v^{(\tau)}(t_{n+1},\cdot)^2\big) - E_\gamma\big(v^{(\tau)}(t_n,\cdot)^2\big) \le -(2\pi)^4\tau\gamma^2\kappa_\gamma E_\gamma\big(v^{(\tau)}(t_{n+1},\cdot)^2\big).$$

Summation over $n = 0, \dots, M-1$ gives

$$\begin{aligned} E_\gamma\big(v^{(\tau)}(t_M,\cdot)^2\big) - E_\gamma(u_0) &\le -(2\pi)^4\tau\gamma^2\kappa_\gamma \sum_{j=1}^{M} E_\gamma\big(v^{(\tau)}(t_j,\cdot)^2\big) \\ &\le -(2\pi)^4\tau\gamma^2\kappa_\gamma \int_\tau^{t_M} E_\gamma\big(v^{(\tau)}(s,\cdot)^2\big)ds. \end{aligned}$$

Keep t fixed; perform the limits $\tau \to 0$ and $M \to \infty$ such that $t_M = M\tau \to t$. Since $v^{(\tau)} \to \sqrt{u}$ strongly in $L^2(0,T;L^\infty(\mathbb{T}^d))$ as $\tau \to 0$,

$$E_\gamma\big(u(t,\cdot)\big) \le E_\gamma(u_0) - (2\pi)^4\tau\gamma^2\kappa_\gamma \int_0^t E_\gamma\big(u(s,\cdot)\big)ds.$$

Gronwall's lemma leads to the desired decay estimate. Decay in the L^1 norm follows immediately from the Csiszár–Kullback inequality. This finishes the proof of Theorem 1.2.

6. Nonuniqueness of solutions. In dimensions $d \le 3$, we provide a family of initial conditions for which the DLSS equation (1.4)–(1.5) has at least two solutions in the class $L^1(0,T;H^2(\mathbb{T}^d))$ for all $T > 0$. Namely, for arbitrary integers $n_1, \dots, n_d$, let

$$\hat{u}(t,x) = \cos^2(n_1\pi x_1)\cdots\cos^2(n_d\pi x_d), \quad x = (x_1,\dots,x_n)^\top \in \mathbb{T}^d.$$

This function is C^∞ smooth, time-independent, spatially multiperiodic, and has finite physical entropy, $\int_{\mathbb{T}^d}(\hat{u}(\log\hat{u}-1)+1)dx < +\infty$. Moreover, a simple calculation shows that the distribution

$$\partial^2_{ij}\big(\sqrt{\hat{u}}\partial^2_{ij}\sqrt{\hat{u}} - \partial_i\sqrt{\hat{u}}\partial_j\sqrt{\hat{u}}\big)$$

is identically zero. In other words, $\hat{u}$ is a weak solution of the stationary and, hence, also of the transient equation. This time-independent function is clearly not physical: it does not converge to the homogeneous steady state and it does not dissipate the physical (or any other) entropy.

On the other hand, Theorems 1.1 and 1.2 provide the existence of a weak solution $u(t,\cdot)$ to (1.4) with initial datum $u_0(x) = \hat{u}(0,x)$ which converges to the constant steady state as $t \to \infty$. Thus, $u \ne \hat{u}$. Hence, we have found two weak solutions to (1.4)–(1.5) in the class of nonnegative functions in $L^1(0,T;H^2(\mathbb{T}^d))$.

Moreover, the above observation makes clear that one cannot expect strict positivity of weak solutions for $t > 0$ if the initial conditions attain zero somewhere. On the other hand, numerical experiments (see, e.g., [8]) lead to the conjecture that for strictly positive initial data, the solutions are also strictly positive.

We remark that the stationary solution $\hat{u}$ does *not* have the regularity stated in the conclusions of Theorem 1.1: observe that $\sqrt{\hat{u}} \notin L^2(0,T;H^2(\mathbb{T}^d))$. Whether the condition $\sqrt{u} \in L^2(0,T;H^2(\mathbb{T}^d))$ is sufficient to obtain entropy-dissipative solutions (or perhaps even uniqueness and positivity) remains an open question.

Acknowledgment. The authors thank the anonymous referee for a simplification of the proof of Lemma 2.1.

REFERENCES

[1] A. Arnold, P. Markowich, G. Toscani, and A. Unterreiter, *On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck-type equations*, Comm. Partial Differential Equations, 26 (2001), pp. 43–100.

[2] E. Beretta, M. Bertsch, and R. Dal Passo, *Nonnegative solutions of a fourth-order nonlinear degenerate parabolic equation*, Arch. Ration. Mech. Anal., 129 (1995), pp. 175–200.

[3] W. Beckner, *A generalized Poincaré inequality for Gaussian measures*, Proc. Amer. Math. Soc., 105 (1989), pp. 397–400.

[4] P. Bleher, J. Lebowitz, and E. Speer, *Existence and positivity of solutions of a fourth-order nonlinear PDE describing interface fluctuations*, Comm. Pure Appl. Math., 47 (1994), pp. 923–942.

[5] S. Bobkov and P. Tetali, *Modified log-Sobolev inequalities in discrete settings*, J. Theoret. Probab., to appear.

[6] M. Cáceres, J. Carrillo, and G. Toscani, *Long-time behavior for a nonlinear fourth order parabolic equation*, Trans. Amer. Math. Soc., 357 (2004), pp. 1161–1175.

[7] J. A. Carrillo, J. Dolbeault, I. Gentil, and A. Jüngel, *Entropy-energy inequalities and improved convergence rates for nonlinear parabolic equations*, Discrete Contin. Dyn. Systet. Ser. B, 6 (2006), pp. 1027–1050.

[8] J. A. Carrillo, A. Jüngel, and S. Tang, *Positive entropic schemes for a nonlinear fourth-order equation*, Discrete Contin. Dynam. Syst. Ser. B, 3 (2003), pp. 1–20.

[9] I. Csiszár, *Information-type measures of difference of probability distributions and indirect observations*, Studia Sci. Math. Hungar., 2 (1967), pp. 299–318.

[10] B. Derrida, J. Lebowitz, E. Speer, and H. Spohn, *Dynamics of an anchored Toom interface*, J. Phys. A, 24 (1991), pp. 4805–4834.

[11] B. Derrida, J. Lebowitz, E. Speer, and H. Spohn, *Fluctuations of a stationary nonequilibrium interface*, Phys. Rev. Lett., 67 (1991), pp. 165–168.

[12] J. Dolbeault, I. Gentil, and A. Jüngel, *A logarithmic fourth-order parabolic equation and related logarithmic Sobolev inequalities*, Commun. Math. Sci., 4 (2006), pp. 275–290.

[13] U. Gianazza, G. Savaré, and G. Toscani, *The Wasserstein Gradient Flow of the Fisher Information and the Quantum Drift-Diffusion Equation*, Arch. Ration. Mech. Anal., 2008, to appear.

[14] M. P. Gualdani, A. Jüngel, and G. Toscani, *A nonlinear fourth-order parabolic equation with non-homogeneous boundary conditions*, SIAM J. Math. Anal., 37 (2006), pp. 1761–1779.

[15] D. Henry, *How to remember the Sobolev inequalities*, in Differential Equations, São Paolo, 1981, Lecture Notes in Math. 957, Springer, Berlin, 1982, pp. 97–109.

[16] A. Jüngel and D. Matthes, *An algorithmic construction of entropies in higher-order nonlinear PDEs*, Nonlinearity, 19 (2006), pp. 633–659.

[17] A. Jüngel and R. Pinnau, *Global nonnegative solutions of a nonlinear fourth-oder parabolic equation for quantum systems*, SIAM J. Math. Anal., 32 (2000), pp. 760–777.

[18] A. Jüngel and R. Pinnau, *A positivity preserving numerical scheme for a nonlinear fourth-order parabolic equation*, SIAM J. Numer. Anal., 39 (2001), pp. 385–406.

[19] A. Jüngel and G. Toscani, *Exponential decay in time of solutions to a nonlinear fourth-order parabolic equation*, Z. Angew. Math. Phys., 54 (2003), pp. 377–386.

[20] S. Kullback, *A lower bound for discrimination information in terms of variation*, IEEE Trans. Inform. Theory, 4 (1967), pp. 126–127.

[21] M. Ledoux, *On Talagrand's deviation inequalities for product measures*, ESAIM Probab. Statist., 1 (1995/97), pp. 63–87.

[22] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.

[23] M. Taylor, *Partial Differential Equations.* III. *Nonlinear Equations*, Springer, New York, 1997.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 2016–2032

A FLUID DYNAMIC MODEL FOR T-JUNCTIONS*

ALESSIA MARIGO† AND BENEDETTO PICCOLI†

Abstract. Motivated by real road junctions, we consider a new fluid dynamic model for traffic flow on networks. In particular at T-junctions, beside some flows distribution and/or merging, there happen some interactions of cars coming from different roads and going to different destinations. After determining some rules to uniquely solve Riemann problems, we prove existence of solutions on complete networks for initial data with bounded variation (and their limits in L^1_{loc}).

Key words. traffic flow on networks, conservation laws, T-junctions

AMS subject classifications. 90B20, 35L65

DOI. 10.1137/060673060

1. Introduction.

In recent years, many authors contributed to the development of fluid dynamic theory for flows on networks, see [4, 6, 7, 10, 11, 13, 14, 15, 16, 19]. The most important applications are in urban car traffic [5, 12], telecommunication data networks [9, 21, 22], gas pipelines [2], supply chains [1], and others.

The main approach used for car traffic is based on the idea of a junction (corresponding to the nodes of the network), with no particular relationships between incoming and outgoing roads. Then, real crossings are modeled splitting them into many junctions of lower complexity; see, e.g., [6, 13]. This means that the traffic flow from incoming roads distributes to outgoing roads according to certain preferences, which in [7] were modeled by a traffic distribution matrix. Then, in general, to solve uniquely Riemann problems, i.e., Cauchy problems with constant initial data on each road, one also needs to impose the maximization of a functional, e.g., the total flux.

In this paper we consider a different point of view for junctions, inspired by modeling need. To illustrate our approach, let us focus on a simple example of T-junction, represented in Figure 1. Here we have three roads with both directions of traffic. Then we can individuate the incoming flows, denoted by 1, 2, and 3, and the outgoing ones, denoted by A, B, and C. Each incoming flux at the junction splits into two parts depending on the final destination. Thus flux 1 is split into fluxes $1B$ and $1C$ (assuming that U-turns are not possible). As Figure 1 shows, there are many interactions among the various fluxes at the junction. However, not all such interactions can be considered in the same way. In fact, for instance, fluxes $1B$ and $3B$ must flow to the same final direction, thus clearly their sum cannot exceed the possible outgoing flow towards B. On the contrary, fluxes $1B$ and $3A$ share conflicting trajectories, but they do not share the same final destination, thus their sum is bounded only by the junction capacity.

To capture this situation, we model the T-junction as in Figure 2. More precisely, to encompass the whole dynamic happening at the T-junction, we use nine virtual junctions denoted by letters G, H, and K. The three junctions G are formed by an

*Received by the editors October 23, 2006; accepted for publication (in revised form) September 14, 2007; published electronically March 21, 2008. This research was partially supported by the INDAM Strategic Project 2005 "Traffic flows and optimization on complex networks" and by the HyCon Network of Excellence, E.C. grant IST-511368.
http://www.siam.org/journals/sima/39-6/67306.html
†Istituto per le Applicazioni del Calcolo "Mauro Picone," C.N.R., Viale del Policlinico 137, 00161 Roma, Italy (marigo@iac.cnr.it, piccoli@iac.cnr.it).

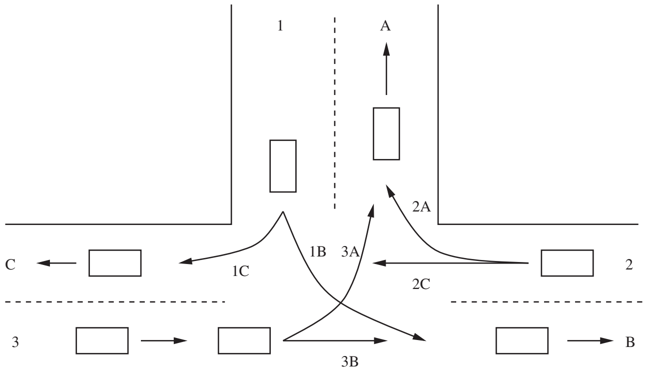


Fig. 1. *An example of T-junction.*

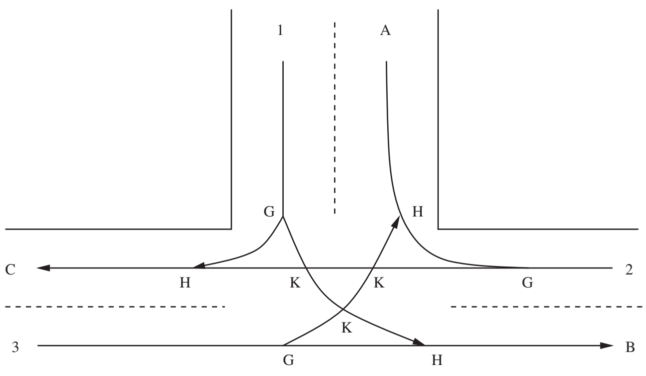


Fig. 2. *The model for a T-junction.*

incoming road and two outgoing roads, thus the described phenomenon is simply a flux split. Such junctions were already modeled in [7].

Instead, the three junctions H are formed by two incoming and one outgoing roads. In this case, clearly some right of way or yielding rule should show up to describe the traffic distribution. This is in fact the case and the theory was first developed in [6].

The junctions K have a quite different meaning. In this case there are two incoming and two outgoing roads. However, the traffic from each incoming road goes to a precise outgoing road, while sharing the junction space. Our main aim is then to model these new types of junctions. Let us illustrate the mathematical counterpart of this example.

We use the Lighthill–Whitham–Richards model (see [20, 23]), which consists of a single conservation law:

$$\rho_t + f(\rho)_x = 0, \tag{1.1}$$

where $\rho \in [0, \rho_{\max}]$ is the car density and $f(\rho) = \rho v(\rho)$ is the flux with $v(\rho)$ the average velocity. A junction J is called a crossing junction if it has the same characteristic of junctions K above. Thus J has n incoming roads, denoted for simplicity by $I_1, \dots, I_n$, and n outgoing roads, denoted by $I_{n+1}, \dots, I_{2n}$. Also, we denote by $\rho_i(t,x)$, $i = 1, \dots, n$, and $\rho_j(t,x)$, $j = n+1, \dots, 2n$, the traffic densities, respectively, on the incoming roads and the outgoing ones. To describe the dynamics at J, we assume the following.

(1) The flux from road I_i is the same of the corresponding exiting road I_{n+i}.

(2) The total flux through J does not exceed its maximum capacity Γ_J.

(3) The total flux through J is maximal respecting rules (1) and (2).

In case of high traffic, rules (1), (2), and (3) are not enough to isolate a unique solution to Riemann problems at J. This need of a new rule is not at all surprising, since such rules were used in car traffic for junctions with more incoming than outgoing roads, see [6], and also for telecommunication networks; see [9].

Therefore, we introduce a *flux proportion rule*, which is active only when the maximal incoming fluxes overcome the maximal junction capacity Γ_J. More precisely, we assume that there exist ideal equilibrium flux proportions among incoming roads. Thus, there exist coefficients $\bar{r}_i$ so that the following holds.

(FPR) The flux from road I_{i+1} is $\bar{r}_i$ times the flux from road I_i, for $i = 1, \dots, n-1$.

The rule (FPR) well captures the situation in the example of the T-junction above. Indeed, for instance, the flux $3A$ must yield to the flux $2C$, thus the corresponding flux proportion coefficient will be less than 1. While, usually the flux $1B$ must give precedence to the flux $3A$, unless yielding signs are deciding the contrary.

Let us further illustrate the role of rule (FPR), for simplicity restricting to the case of two incoming and two outgoing roads. First, (FPR) is used only if the sum of incoming maximal fluxes exceeds Γ_J. Then, to respect (FPR), we should set the incoming fluxes γ_1 and γ_2 so that $\gamma_2 = \bar{r}_1\gamma_1$. However, this may be in contrast with (3) (if, for example, $\bar{r}_1 = 1$, the maximal flux $\gamma_1^{\max}$ from road 1 exceeds Γ_J and the maximal flux from road 2 is less than $\Gamma_J/2$). In the latter case, we set the proportion between incoming fluxes so to respect rule (3) (i.e., summing up to Γ_J) and be as close as possible to the value prescribed by (FPR).

We first show how to define the solution to Riemann problems for the new type of junctions: the *crossing junctions*. The procedure to define the solution is based on rules (1), (2), (3), and (FPR). The obtained solution effectively defines a Riemann solver with consistency properties; see Proposition 2.

Then we provide estimates on the total variation of the flux for a wave interacting with a crossing junction, having two incoming and two outgoing roads. Such estimates are the key point to prove bounded variation (BV) estimates on the flux along wave front tracking approximate solutions.

More precisely, a wave front tracking algorithm is defined as in [7], i.e., approximating initial data with a piecewise constant function and solving Riemann problems (RPs) for interactions between waves and of waves with junctions. To provide a well-defined construction, estimates on the number of waves and interactions are in order. The latter are obtained with a careful analysis based on the special properties of the introduced Riemann solver. Then, to pass to the limit, BV estimates on the flux are used, together with standard weak compactness arguments. The final result is existence of solutions to Cauchy problems on networks.

The paper is organized as follows. Section 2 provides the basic definitions and results from previous papers, while in section 2.1 we describe the Riemann solver for crossing junctions. Section 3 contains flux variation estimates for waves interacting with crossing junctions. Finally, in section 4, we prove existence of solutions on the whole network for L^1_{loc} initial data, which can be approximated by BV functions with uniformly bounded variation.

2. Basic definitions. We use the same approach as in [16, 7, 11]. For reader convenience, we recall the main notation and results.

We consider a network formed by a collection $\mathcal{I}$ of unidirectional roads I_i, modeled by real (possibly unbounded) intervals $[a_i, b_i]$, whose natural order respects the direction of the road. Roads meet at junctions: each junction J is given by a collection of incoming roads and outgoing roads, and we indicate by $\mathcal{J}$ the collection of junctions. Thus the network can be identified with a directed graph. On each road the evolution is given by (1.1) and we assume

(F) The flux f is a smooth, strictly concave function (with $f(0) = f(\rho_{\max}) = 0$), thus there exists a unique $\sigma \in [0, \rho_{\max}]$ such that $f'(\sigma) = 0$ and it is the maximum of f over $[0, \rho_{\max}]$.

For notational simplicity, we assume that $\rho_{\max} = 1$.

DEFINITION 1. *We let $\tau : [0, 1] \to [0, 1]$ be the map such that $f(\rho) = f(\tau(\rho))$ and $\tau(\rho) \neq \rho$ if $\rho \neq \sigma$. Thus τ sends ρ to the other density value with the same flux (and $\tau(\sigma) = \sigma$).*

We restrict to crossing junctions as explained in the introduction. Then each junction J has $n = n(J)$ incoming roads and n outgoing roads.

Let us fix now a junction J and for simplicity assume that the incoming roads are $I_1, \ldots, I_n$ and the outgoing roads are $I_{n+1}, \ldots, I_{2n}$. A Riemann problem for system (1.1), on the real line, is a Cauchy problem with Heaviside-type initial data. We define a Riemann problem at J to be a Cauchy problem with initial datum constant on each road. Thus let us fix an initial condition $\rho_0 = (\rho_{1,0}, \ldots, \rho_{2n,0})$. We look for centered self-similar solutions (as it is natural for conservation laws, see [3]), thus we want to determine a $(2n)$-tuple $\hat{\rho} = (\hat{\rho}_1, \ldots, \hat{\rho}_{2n}) \in [0, 1]^{2n}$, so that the following holds. On each incoming road I_i, $i = 1, \ldots, n$, the solution consists of the single wave solution to the Riemann problem $(\rho_{i,0}, \hat{\rho}_i)$, while on each outgoing road I_j, $j = n + 1, \ldots, 2n$, the solution consists of the single wave $(\hat{\rho}_j, \rho_{j,0})$.

We consider waves with negative speed on incoming roads and positive on outgoing ones, thus

$$\hat{\rho}_i \in \begin{cases} \{\rho_{i,0}\} \cup \,]\tau(\rho_{i,0}), 1] & \text{if } 0 \leq \rho_{i,0} \leq \sigma, \\ [\sigma, 1] & \text{if } \sigma \leq \rho_{i,0} \leq 1, \end{cases} \tag{2.1}$$

$i = 1, \ldots, n$, and

$$\hat{\rho}_j \in \begin{cases} [0, \sigma] & \text{if } 0 \leq \rho_{j,0} \leq \sigma, \\ \{\rho_{j,0}\} \cup [0, \tau(\rho_{j,0})[& \text{if } \sigma \leq \rho_{j,0} \leq 1, \end{cases} \tag{2.2}$$

$j = n + 1, \ldots, 2n$. As a consequence, not every flux can be obtained on each road. More precisely, we define

$$\gamma_i^{\max} = \begin{cases} f(\rho_{i,0}) & \text{if } \rho_{i,0} \in [0, \sigma], \\ f(\sigma) & \text{if } \rho_{i,0} \in \,]\sigma, 1], \end{cases} \quad i = 1, \ldots, n, \tag{2.3}$$

and

$$\gamma_j^{\max} = \begin{cases} f(\sigma) & \text{if } \rho_{j,0} \in [0, \sigma], \\ f(\rho_{j,0}) & \text{if } \rho_{j,0} \in \,]\sigma, 1], \end{cases} \quad j = n + 1, \ldots, 2n. \tag{2.4}$$

The quantities $\gamma_i^{\max}$ and $\gamma_j^{\max}$ represent the maximum flux that can be obtained by a single wave solution on each road.

Remark 1. We may consider waves with zero speed on incoming or outgoing roads. However, this would generate shocks which stay at the intersection, without entering roads. Since solutions are usually considered in L^1, such shocks are not part of the solution on roads (because they affect the value only at one point).

Remark 2. The maximum fluxes on incoming and outgoing roads can be interpreted as maximal demands and supplies, according to the theory introduced by Lebacque; see [17, 18, 11]. Notice that for each road I_i and possible flux $\hat{\gamma}_i$, there exists a unique $\hat{\rho}_i$, satisfying (2.1) or (2.2), such that $f(\hat{\rho}_i) = \hat{\gamma}_i$. Moreover, by rule (1), outgoing fluxes are obtained once incoming fluxes are fixed. Then we obtain the following.

PROPOSITION 1. *To solve a Riemann problem at a crossing junction J, it is enough to determine the fluxes on incoming roads.*

We aim at finding a systematic way of solving RP at junctions as described by next definition.

DEFINITION 2. *Given a crossing junction J, we call Riemann solver for J a map $RS : [0,1]^n \times [0,1]^n \to [0,1]^n \times [0,1]^n$ that associates with Riemann data $\rho_0 = (\rho_{1,0}, \ldots, \rho_{2n,0})$ at J a vector $\hat{\rho} = (\hat{\rho}_1, \ldots, \hat{\rho}_{2n})$ so that the solution on an incoming road I_i, $i = 1, \ldots, n$, is given by the wave $(\rho_{i,0}, \hat{\rho}_i)$ and on an outgoing one I_j, $j = n+1, \ldots, 2n$, is given by the wave $(\rho_{j,0}, \hat{\rho}_j)$. We require the consistency condition:*

(CC) $RS\,(RS\,(\rho_0)) = RS\,(\rho_0)$.

Once a Riemann solver is introduced for every junction J, we can define the concept of solution on the network as in [11].

2.1. Riemann solver at crossing junctions. The aim of this section is to describe the solution to a Riemann problem at a crossing junction J, using rules (1), (2), (3), and (FPR).

Fix a crossing junction J with n incoming roads, $I_1, \ldots, I_n$ and n outgoing roads $I_{n+1}, \ldots, I_{2n}$. We denote by $\rho_i(t,x)$, $i = 1, \ldots, n$, and $\rho_j(t,x)$, $j = n+1, \ldots, 2n$, the traffic densities, respectively, on the incoming roads and the outgoing ones and by $(\rho_{i,0}, \rho_{j,0})$ the initial data of a Riemann problem. The rules (1), (2), and (3) can be rewritten as

(1) $f(\hat{\rho}_i) = f(\hat{\rho}_{n+i})$ for each $i = 1, \ldots, n$,

(2) $\sum_i f(\hat{\rho}_i) \le \Gamma_J$,

(3) $\sum_i f(\hat{\rho}_i)$ is maximal respecting rules (1) and (2).

The rules (1) and (2) alone do not give a unique solution $(\hat{\rho}_1, \ldots, \hat{\rho}_{2n})$. Moreover, since the solution $(\hat{\rho}_1, \ldots, \hat{\rho}_{2n})$ must satisfy conditions (1) and (2), we denote by Ω the admissible region for the fluxes $\gamma_i = f(\rho_i)$

$$\Omega = \{(\gamma_1, \ldots, \gamma_n) : \sum \gamma_i \le \Gamma_J, 0 \le \gamma_i \le \gamma_i^\wedge, i = 1, \ldots, n\},$$

where $\gamma_i^\wedge = \min\{\gamma_i^{\max}, \gamma_{n+i}^{\max}\}$.

Let us now quantify rule (FPR). We can rewrite the rule as

(FPR) $\bar{r}_i$ is the ratio among the fluxes on two successive roads $f(\hat{\rho}_i)$ and $f(\hat{\rho}_{i+1})$.

Now, we want to determine a unique solution to the Riemann problem using rules (1), (2), (3), and (FPR). More precisely, we try to fit rule (FPR) as much as possible respecting rules (1), (2), and (3).

Recall that, by Proposition 1, to solve the Riemann problem, it is enough to determine the fluxes $\hat{\gamma}_i = f(\hat{\rho}_i), i = 1, \ldots, n$. (Then $\hat{\gamma}_{n+i} = f(\hat{\rho}_i), i = 1, \ldots, n$.)

Then, let us determine $\hat{\gamma}_i, i = 1, \ldots, n$. We denote $\Gamma = \sum_i \gamma_i^\wedge$. We have to distinguish two cases:

(I) $\Gamma \le \Gamma_J$,

(II) $\Gamma > \Gamma_J$.

In the first case we set $\hat{\gamma}_i = \gamma_i^{\wedge},\ i = 1, \dots, n$. Let us analyze the second case in which we use the flux proportion coefficients $\bar{r}_1, \dots, \bar{r}_{n-1}$.

Consider the space $(\gamma_1, \dots, \gamma_n)$ and denote by γ_r the point that satisfy the following system of equations:

$$\begin{cases} \sum_i \gamma_i = \Gamma_J, \\ \gamma_{i+1} = \bar{r}_i \gamma_i. \end{cases} \tag{2.5}$$

Recall that the final fluxes should belong to Ω. We distinguish two cases:

(a) γ_r belongs to Ω,

(b) γ_r is outside Ω.

In the first case the solution is $\hat{\gamma} = \gamma_r$, which, by direct computations, is given by

$$\begin{aligned} \hat{\gamma}_1 &= \hat{\gamma}_{n+1} = \frac{1}{p(r)} \Gamma_J, \\ \hat{\gamma}_2 &= \hat{\gamma}_{n+2} = \frac{\bar{r}_1}{p(r)} \Gamma_J, \\ \hat{\gamma}_3 &= \hat{\gamma}_{n+3} = \frac{\bar{r}_1 \bar{r}_2^{\min}}{p(r)} \Gamma_J, \\ &\vdots \\ \hat{\gamma}_n &= \hat{\gamma}_{2n} = \frac{\bar{r}_1 \bar{r}_2 \cdots \bar{r}_{n-1}}{p(r)} \Gamma_J, \end{aligned} \tag{2.6}$$

where $p(r) = 1 + \bar{r}_1 + \bar{r}_1 \bar{r}_2 + \dots + \bar{r}_1 \bar{r}_2 \cdots \bar{r}_{n-1}$.

In the second case, we project γ_r on the admissible region. More precisely, for the case of two incoming and outgoing roads, i.e., for $n = 2$, the solution is as follows. If $\gamma_{r1} > \gamma_1^{\wedge}$, we set

$$\begin{aligned} \hat{\gamma}_1 &= \hat{\gamma}_3 = \gamma_1^{\wedge}, \\ \hat{\gamma}_2 &= \hat{\gamma}_4 = \Gamma_J - \gamma_1^{\wedge}. \end{aligned}$$

If otherwise $\gamma_{r2} > \gamma_2^{\wedge}$, we set

$$\begin{aligned} \hat{\gamma}_2 &= \hat{\gamma}_4 = \gamma_2^{\wedge}, \\ \hat{\gamma}_1 &= \hat{\gamma}_3 = \Gamma_J - \gamma_2^{\wedge}. \end{aligned}$$

For reader convenience, we illustrate the cases I, IIa, and IIb for the case of $n = 2$ in Figure 3.

For $n > 2$ we chose $\hat{\gamma}$ to be the projection $pr_{\widehat{\Omega}}(\gamma_r)$ on the convex set $\widehat{\Omega} = \Omega \cap \{\gamma : \sum_{i=1}^n \gamma_i = \Gamma_J\}$. (Notice that γ_r already belongs to the hyperplane $\{\gamma : \sum_{i=1}^n \gamma_i = \Gamma_J\}$.)

For future convenience we make a little abuse of notation as follows.

Notation. We define

$$\tilde{\Omega} = \{(\gamma_1, \dots, \gamma_n) : 0 \leq \gamma_i \leq \gamma_i^{\wedge}, i = 1, \dots, n\}$$

and denote by Case IIa the case where $\gamma_r \in \mathrm{Int}(\tilde{\Omega})$ and by Case IIb the case where $\gamma_r \notin \mathrm{Int}(\tilde{\Omega})$.

Roughly speaking, the case of γ_r belonging to the boundary of $\tilde{\Omega}$ is included in Case IIb, in fact this corresponds to have equality constraints on some flux solutions, as it will be more clear later.

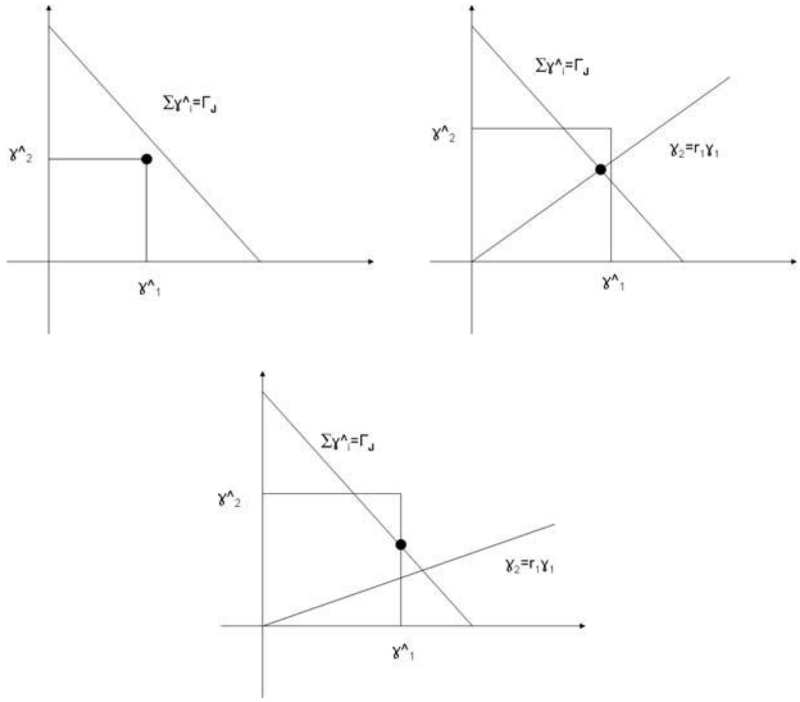


FIG. 3. *Solutions at junctions for $n = 2$: Case* I *(top), Case* IIa *(centre), Case* IIb *(bottom). The bold point represents the values of the fluxes in the solution to the Riemann problem.*

Moreover, we introduce the following.

Notation. For all $i \in \{1, \dots, n\}$ we denote

Case H0. $\rho_i \in [0, \sigma]$ and $\rho_{n+i} \in [\sigma, 1]$.

Case H1. $\rho_i \in [0, \sigma[$ and $\rho_{n+i} \in [0, \sigma[$.

Case H2. $\rho_i \in]\sigma, 1]$ and $\rho_{n+i} \in]\sigma, 1]$.

Case H3. $\rho_i \in]\sigma, 1]$ and $\rho_{n+i} \in [0, \sigma[$.

PROPOSITION 2. *Given an initial condition $(\rho_{i,0}, \rho_{j,0})$, there exists a unique $(\hat{\rho}_1, \dots, \hat{\rho}_{2n})$ satisfying rules* (1), (2), (3), *and* (FPR). *Such solution uniquely defines a Riemann solver at J, which respects the compatibility condition* (CC).

Proof. Assume first that $\Gamma \le \Gamma_J$, i.e., we are in Case I, then (2) is automatically satisfied. From (1) and (3), the solution to the RP is the point with coordinates $\hat{\gamma}_i = \hat{\gamma}_{n+i} = \min\{\gamma_i^{\max}, \gamma_{n+i}^{\max}\}$. Then there is no need to apply rule (FPR). The following situations represent an equilibrium point for the RP, i.e., $RS(\rho_0) = \rho_0$.

(i) Case H0 with $\rho_{n+i,0} = \tau(\rho_{i,0})$. Indeed $\hat{\gamma}_i = \hat{\gamma}_{n+i} = \gamma_i^{\wedge}$ and the unique solution $\hat{\rho}_i$ that satisfies $f(\hat{\rho}_i) = \hat{\gamma}_i$ and (2.1) is $\hat{\rho}_i = \rho_{i,0}$. The same can be said for $\hat{\rho}_{n+i}$.

(ii) Case H1 with $\rho_{i,0} = \rho_{n+i,0}$. Indeed $\hat{\gamma}_i = \hat{\gamma}_{n+i} = \gamma_1^{\max}$ and, as for Case H0, we have $\hat{\rho}_i = \rho_{i,0}$, while the unique $\hat{\rho}_{n+i}$, such that $f(\hat{\rho}_{n+i}) = \gamma_1^{\max}$, is $\hat{\rho}_{n+i} = \rho_{n+i,0}$.

(iii) Case H2 with $\rho_{i,0} = \rho_{n+i,0}$. This case is the opposite of (ii) and we reason in the same way.

Now the following cases may happen, for $i = 1, \dots, n$.

- If $\gamma_i^{\max} = \gamma_{n+i}^{\max}$ and the initial condition is described by Case H0, then we are in case (i).

- If $\gamma_i^{\max} = \gamma_{n+i}^{\max}$ and $(\rho_{i,0}, \rho_{n+i,0})$ is described by Case H3, then $f(\hat{\rho}_i) = \hat{\gamma}_i = \gamma_i^{\max} = f(\sigma)$ and $f(\hat{\rho}_{n+i}) = \hat{\gamma}_{n+i} = \gamma_{n+i}^{\max} = f(\sigma)$. Then the solution to the RP is given by $\hat{\rho}_i = \hat{\rho}_{n+i} = \sigma$, thus falling in (i).
- If $\gamma_i^{\max} < \gamma_{n+i}^{\max}$, then it must be $\rho_{i,0} \in [0, \sigma[$ and $\rho_{n+i,0} \in [0, \tau(\rho_{i,0})[$. Thus $f(\hat{\rho}_i) = \hat{\gamma}_i = \gamma_i^{\max} = f(\rho_{i,0})$, hence $\hat{\rho}_i = \rho_{i,0}$, and $f(\hat{\rho}_{n+i}) = \hat{\gamma}_i = \gamma_i^{\max} = f(\rho_{i,0})$ and $\hat{\rho}_{n+i,0} = \hat{\rho}_i$ since this is the unique solution that satisfies (2.2). The solution to the RP falls in case (ii).
- If $\gamma_i^{\max} > \gamma_{n+i}^{\max}$, then, reasoning as in the previous case, the solution falls in case (iii).

Finally, we proved that, for every ρ_0 satisfying Case I, $RS(\rho_0)$ is an equilibrium, therefore condition (CC) is satisfied.

Assume now Case IIa occurs, which means $\Gamma > \Gamma_J$ and γ_r, obtained by solving (2.5), belongs to $\mathrm{Int}(\tilde{\Omega})$. Then this is the only point satisfying all rules. Thus the solution is given by $\hat{\gamma} = \gamma_r$ (see (2.6)). The following situations represent an equilibrium point for the RP, i.e., $RS(\rho_0) = \rho_0$.

(iv) Case H3 with $\rho_{n+i,0} = \tau(\rho_{i,0})$ and $f(\rho_{i,0}) = \gamma_{r,i}$. Indeed $\hat{\gamma}_i = \hat{\gamma}_{n+i} = \gamma_{r,i}$ and the unique solution $(\hat{\rho}_i, \hat{\rho}_{n+i})$ such that $f(\hat{\rho}_i) = f(\hat{\rho}_{n+i}) = \gamma_{r,i}$ and (2.1) and (2.2) are satisfied is $(\rho_{i,0}, \rho_{n+i,0})$.

Now $\hat{\gamma} \in \tilde{\Omega}$ by assumption hence for $i = 1, \ldots, n$, it holds that $\hat{\gamma}_i < \gamma_i^{\wedge}$. Moreover, the unique solution $(\hat{\rho}_i, \hat{\rho}_{n+i})$ such that $f(\hat{\rho}_i) = f(\hat{\rho}_{n+i}) = \hat{\gamma}_i$ and (2.1) and (2.2) are satisfied is $\hat{\rho}_{n+i} = \tau(\hat{\rho}_i) \in [0, \sigma[$. This new point is described by case (iv).

Again, condition (CC) is satisfied.

Finally, assume that Case IIb holds true, i.e., $\Gamma > \Gamma_J$ and $\gamma_r \notin \mathrm{Int}(\tilde{\Omega})$. Then the solution given by $pr_{\widehat{\Omega}}(\gamma_r)$ maximizes the flux from each road, while respecting rules (1), (2), (3), and (FPR) (if not in contrast with previous rules). For $i = 1, \ldots, n$ we have either $\hat{\gamma}_i = \hat{\gamma}_{n+1} = \gamma_i^{\wedge}$ or $\hat{\gamma}_i = \hat{\gamma}_{n+i} < \gamma_1^{\wedge}$. In the first case we have the same analysis as for Case I, while in the second case we have the same analysis as for Case IIa.

Then condition (CC) is again satisfied and we are done. □

3. Interaction estimates for 2 × 2 crossing junctions. The aim of this section is to obtain estimates for the flux variation in case of interactions of a wave with a 2×2 junction J, i.e., a junction with two incoming and two outgoing roads.

For simplicity, we assume that the incoming roads are I_1 and I_2, while the outgoing roads are I_3 and I_4. We use the superscript "$-$" to indicate all quantities before the interaction and "$+$" to indicate all quantities after the interaction. Thus, for example, the initial data are given by $(\rho_1^-, \ldots, \rho_4^-)$, while the solution after the interaction is $(\rho_1^+, \ldots, \rho_4^+)$. We have four possible situations for the initial data:

Case I: $\Gamma \leq \Gamma_J$.

Case IIa: $\Gamma > \Gamma_J$ and $\gamma_r \in \mathrm{Int}(\tilde{\Omega})$.

Case IIb1: $\Gamma > \Gamma_J$, $\gamma_r \notin \mathrm{Int}(\tilde{\Omega})$, and $\gamma_{r1} \geq \gamma_1^{\wedge}$.

Case IIb2: $\Gamma > \Gamma_J$, $\gamma_r \notin \mathrm{Int}(\tilde{\Omega})$, and $\gamma_{r2} \geq \gamma_2^{\wedge}$.

We assume that the interaction with J happens for a wave (ρ_1, ρ_1^-) coming from road I_1, being the other cases entirely similar. In next sections, we provide estimates on the flux variation in the different cases.

3.1. Case I. We consider first the case $\Gamma \leq \Gamma_J$. As explained in the proof of Proposition 2, at the equilibrium before the interaction, for roads I_1 and I_3, and for roads I_2 and I_4 the following cases may happen: (i), (ii), and (iii). Now we describe the equilibrium type after the interaction of the wave (ρ_1, ρ_1^-) in the different cases.

Case i. We have that $\rho_1^- = \tau(\rho_3^-) \le \sigma$. Now ρ_1 is such that the speed of the wave (ρ_1, ρ_1^-) is positive then it must be $\rho_1 \in [0,\sigma] \setminus \{\rho_1^-\}$.

- If $\rho_1 \in [0, \rho_1^-[$, then $\gamma_1^{\max} < \gamma_1^{\max,-} = \gamma_3^{\max,-}$. We fall in Case I.H1 with $\rho_3^+ = \rho_1^+$ where $\rho_1^+ = \rho_1$ since $f(\rho_1^+) = \gamma_1^{\max}$. The new equilibrium is then $(\rho_1, \rho_2^-, \rho_1, \rho_4^-)$.
- If otherwise $\rho_1 \in]\rho_1^-, \sigma]$, then $\gamma_1^{\max} > \gamma_1^{\max,-} = \gamma_3^{\max,-}$. We therefore fall in Case I.H2 with the equilibrium $\rho_1^+ = \rho_3^+ = \rho_3^-$. The new equilibrium is $(\tau(\rho_1^-), \rho_2^-, \rho_3^-, \rho_4^-)$.

Case ii. In this case we have $\rho_1^- \in [0,\sigma]$ and $\rho_3^- = \rho_1^-$. Again, since the wave (ρ_1, ρ_1^-) must have positive speed, ρ_1 must satisfy $\rho_1 \in [0,\sigma] \setminus \{\rho_1^-\}$. Different cases may arise. To distinguish among them we denote by $\rho_1^{\lim}$ the density for which $f(\rho_1^{\lim}) + \gamma_2^{\wedge,-} = \Gamma_J$. Notice that, since we are in the situation $\Gamma_J \ge \Gamma^-$, it must be $\rho_1^{\lim} \ge \rho_1^-$.

- $\rho_1 \in [0, \rho_1^-[$. Thus $\gamma_1^{\max} < \gamma_1^{\max,-} < \gamma_3^{\max,-}$. This is the case denoted by Case I.H1 with $\rho_3^+ = \rho_1^+$ and $\rho_1^+ = \rho_1$. The new equilibrium is then given by $(\rho_1, \rho_2^-, \rho_1, \rho_4^-)$.
- $\rho_1 \in]\rho_1^-, \rho_1^{\lim}]$. Thus $\gamma_3^{\max,-} = f(\sigma) \ge \gamma_1^{\max} > \gamma_1^{\max,-}$ and $\gamma_1^{\max} = \gamma_1^{\wedge} \le \Gamma_J - \gamma_2^{\wedge,-}$. We fall in Case I.H1 with $\rho_3^+ = \rho_1^+ = \rho_1$ and the new equilibrium is given by $(\rho_1, \rho_2^-, \rho_1, \rho_4^-)$.
- If $\rho_1 \in]\rho_1^{\lim}, \sigma]$, then $\gamma_3^{\max,-} = f(\sigma) \ge \gamma_1^{\max} > \gamma_1^{\max,-}$ and $\gamma_1^{\max} = \gamma_1^{\wedge} > \Gamma_J - \gamma_2^{\wedge,-}$. This situation is described by the following cases:
  - if $\gamma_1^{\max} \le \gamma_{r1}$ and $\gamma_2^{\wedge,-} > \gamma_{r2}$, then we fall in Case IIb1.H1. The new equilibrium is $(\rho_1, \rho_2^+, \rho_1, \rho_4^+)$, where $f(\rho_2^+) = f(\rho_4^+) = \Gamma_J - f(\rho_1)$. Notice that $f(\rho_2^+) = f(\rho_4^+) < \gamma_2^{\wedge,-}$ therefore, for roads I_2 and I_4 we are in Case H3;
  - if $\gamma_1^{\max} > \gamma_{r1}$ and $\gamma_2^{\wedge,-} \le \gamma_{r2}$, then we fall in Case IIb2.H3. The new equilibrium is $(\rho_1^+, \rho_2^-, \rho_3^+, \rho_4^-)$ with $f(\rho_1^+) = f(\rho_3^+) = \Gamma_J - \gamma_2^{\wedge}$;
  - if $\gamma_1^{\max} > \gamma_{r1}$ and $\gamma_2^{\wedge,-} > \gamma_{r2}$, then we fall in Case IIa.H3 and then the new equilibrium is $(\rho_1^+, \rho_2^+, \rho_1^+, \rho_4^+)$, where $f(\rho_1^+) = \gamma_{r1}$ and $f(\rho_2^+) = f(\rho_4^+) = \Gamma_J - f(\rho_1^+) = \gamma_{r2}$. Roads I_2 and I_4 also fall in the situation described by Case H3.

Case iii. We have that $\rho_1^- = \rho_3^- > \sigma$. Since ρ_1 is such that the speed of wave (ρ_1, ρ_1^-) is positive, it must be $\rho_1 \in [0, \tau(\rho_3^-)[$. Thus $\gamma_1^{\max} < \gamma_3^{\max,-}$ and we fall in Case H1 with new equilibrium $(\rho_1, \rho_2^-, \rho_1, \rho_4^-)$.

In Case i, Case ii with $0 \le \rho_1 \le \rho_1^{\lim}$, and Case iii we have that $\rho_2^+ = \rho_2^-$ and $\rho_4^+ = \rho_4^-$ hence the variation of the flux at the junction is given only by the variation of the flux in roads I_1 and I_3. Only in Case ii with $\rho_1^{\lim} < \rho_1 \le \sigma$ we may have that $\rho_2^+ \ne \rho_2^-$ and $\rho_4^+ \ne \rho_4^-$. In this case, it holds that

$$f(\rho_2^+) = f(\rho_4^+) = \Gamma_J - f(\rho_1^+).$$

Then for the flux variation we have

$$0 \le f(\rho_2^-) - f(\rho_2^+) = f(\rho_2^-) - \Gamma_J + f(\rho_1^+) \le f(\rho_1^+) - f(\rho_1^-) \le f(\rho_1) - f(\rho_1^-),$$

and similarly for road I_4.

3.2. Case IIa. Assume now that $\Gamma > \Gamma_J$ with $\gamma_r \in \mathrm{Int}(\tilde{\Omega})$. Then at the equilibrium we are in Case iv.

We have $\rho_1^- = \tau(\rho_3^-) \in]\sigma, 1]$. Then it must be $\rho_1 \in [0, \tau(\rho_1^-)[$ for the wave (ρ_1, ρ_1^-) to have positive speed. In this case we get that $\gamma_1^{\max} < f(\sigma) = \gamma_1^{\max,-} = \gamma_3^{\max,-}$. Define again $\rho_1^{\lim}$ by $f(\rho_1^{\lim}) + \gamma_2^{\wedge,-} = \Gamma_J$, then we may have the following:

- $\rho_1 < \rho_1^{\lim}$, then $\gamma_1^{\max} < \gamma_1^{\max,-} = \gamma_3^{\max,-}$ and $\Gamma \le \Gamma_J$. Then we fall in Case I.H1 and the equilibrium is given by $(\rho_1, \rho_2^+, \rho_1, \rho_4^+)$ with $f(\rho_2^+) = \gamma_2^{\wedge,-}$. The situation in the roads I_2 and I_4 falls in one among Case H0, Case H1, and Case H2. For the flux variation on roads I_2 and I_4 we have

$$f(\rho_2^+) = \gamma_2^{\wedge,-} = \Gamma_J - f(\rho_1^{\lim}) < \Gamma_J - f(\rho_1)$$

and $f(\rho_2^-) = \Gamma_J - f(\rho_1^-)$. Thus

$$f(\rho_2^+) - f(\rho_2^-) < f(\rho_1^-) - f(\rho_1).$$

- $\rho_1 = \rho_1^{\lim}$. This case is similar to the previous one: we fall in Case I.H1 with equilibrium $(\rho_1, \rho_2^+, \rho_1, \rho_4^+)$, with $f(\rho_2^+) = f(\rho_4^+) = \gamma_2^{\wedge,-} = \Gamma_J - f(\rho_1)$ and the flux variation on roads I_2 and I_4 is

$$f(\rho_2^+) - f(\rho_2^-) = f(\rho_1^-) - f(\rho_1).$$

- $\rho_1^{\lim} < \rho_1 < \gamma_{r1}$, then we fall in Case IIb1.H1 with equilibrium $(\rho_1, \rho_2^+, \rho_1, \rho_4^+)$ where $f(\rho_2^+) = f(\rho_4^+) = \Gamma_J - f(\rho_1)$. The situation in roads I_2 and I_4 remains in Case H3 and the flux variation is

$$f(\rho_2^+) - f(\rho_2^-) = \Gamma_J - f(\rho_1) - (\Gamma_J - f(\rho_1^-)) = f(\rho_1^-) - f(\rho_1).$$

- $\gamma_{r1} \le \rho_1 < \tau(\rho_1^-)$, then we remain in Case IIa.H3 with equilibrium $(\rho_1^-, \rho_2^-, \rho_3^-, \rho_4^-)$ and no flux variation occurs.

3.3. Case IIb1. Assume finally that $\Gamma > \Gamma_J$ with $\gamma_{r1} \ge \gamma_1^{\wedge,-}$. Then roads I_1 and I_3 are in one among Case i, Case ii, and Case iii, while roads I_2 and I_4 are in Case iv. Let us consider now the equilibrium after the interaction of a wave (ρ_1, ρ_1^-).

Case i. In this case $\rho_1^- \in [0, \sigma]$ and $\rho_3^- = \tau(\rho_1^-)$. Then since (ρ_1, ρ_1^-) must have positive speed, $\rho_1 \in [0, \sigma]$.

- If $\rho_1 \in [0, \rho_1^-[$, we have $\gamma_1^{\max} < \gamma_1^{\max,-} = \gamma_3^{\max,-}$ and we have to distinguish among $\rho_1 < \rho_1^{\lim}$, $\rho_1 = \rho_1^{\lim}$, and $\rho_1 > \rho_1^{\lim}$. In all cases the new equilibrium is given by $(\rho_1, \rho_2^+, \rho_1, \rho_4^+)$ with $f(\rho_2^+) = f(\rho_4^+) = \Gamma_J - f(\rho_1)$ or $f(\rho_2^+) = f(\rho_4^+) = \gamma_2^{\wedge,-}$ and, since $f(\rho_2^-) = \Gamma_J - f(\rho_1^-)$, the flux variation satisfies

$$f(\rho_2^+) - f(\rho_2^-) \le f(\rho_1^-) - f(\rho_1).$$

- If $\rho_1 \in]\rho_1^-, \sigma]$, then $\gamma_1^{\max} > \gamma_1^{\max,-} = \gamma_3^{\max,-}$. This is Case IIb1.H2 and the new equilibrium is given by $(\tau(\rho_1^-), \rho_2^-, \rho_3^-, \rho_4^-)$.

Case ii. Now $\rho_1^- \in [0, \sigma[$ and $\rho_3^- = \rho_1^-$. Then it must be $\rho_1 \in [0, \sigma]$.

- If $\rho_1 \in [0, \rho_1^-[\cup]\rho_1^-, f^{-1}(\gamma_{r1})]$, then we distinguish among the cases $\rho_1 < \rho_1^{\lim}$, $\rho_1 = \rho_1^{\lim}$, and $\rho_1 > \rho_1^{\lim}$. The new equilibrium is $(\rho_1, \rho_2^+, \rho_1, \rho_4^+))$ with $f(\rho_2^+) = f(\rho_4^+) = \Gamma_J - f(\rho_1)$ or $f(\rho_2^+) = f(\rho_4^+) = \gamma_2^{\wedge,-}$. For the flux variation we have

$$f(\rho_2^+) - f(\rho_2^-) \le f(\rho_1^-) - f(\rho_1).$$

- If $\rho_1 \in]f^{-1}(\gamma_{r1}), \sigma]$, then we fall in Case IIa.H3 and the new equilibrium is $(\rho_1^+, \rho_2^+, \tau(\rho_1^+), \tau(\rho_2^+))$ with $f(\rho_1^+) = \gamma_{r1}$ and $f(\rho_2^+) = \Gamma_J - f(\rho_1^+) = \gamma_{r1}$. Since $\rho_1 > \rho_1^+$, then $f(\rho_1) > f(\rho_1^+)$ and the flux variation is

$$f(\rho_2^-) - f(\rho_2^+) = f(\rho_1^+) - f(\rho_1^-) < f(\rho_1) - f(\rho_1^-).$$

Case iii. In this case we have that $\rho_1^- \in]\sigma, 1]$ and $\rho_3^- = \rho_1^-$. Then it must be $\rho_1 \in [0, \tau(\rho_1^-)]$ with $\gamma_1^{\max} < \gamma_3^{\max,-}$. We may have the following cases:

- $\rho_1^{\lim} \leq \rho_1 \leq \tau(\rho_1^-)$. Then we fall in Case IIb1.H1 and the new equilibrium is $(\rho_1, \rho_2^+, \rho_1, \rho_4^+))$ with $f(\rho_2^+) = f(\rho_4^+) = \Gamma_J - f(\rho_1)$. The flux variation is given by

$$f(\rho_2^+) - f(\rho_2^-) = f(\rho_1^-) - f(\rho_1).$$

- $0 \leq \rho_1 \leq \rho_1^{\lim}$. This is the situation described by Case I.H1 with equilibrium $(\rho_1, \rho_2^+, \rho_1, \rho_4^+))$ with $f(\rho_2^+) = f(\rho_4^+) = \gamma_2^{\wedge,-}$. Then the flux variation is

$$f(\rho_2^+) - f(\rho_2^-) = \gamma_2^{\wedge,-} - f(\rho_2^-) = -f(\rho_1^{\lim}) + f(\rho_1^-) < f(\rho_1^-) - f(\rho_1).$$

3.4. Case IIb2. Assume finally that $\Gamma > \Gamma_J$ with $\gamma_{r2} \geq \gamma_2^{\wedge,-}$. Then roads I_2 and I_4 are in one among Case i, Case ii, and Case iii while roads I_1 and I_3 are in Case iv. Let us consider now the equilibrium after the interaction of a wave (ρ_1, ρ_1^-). We have $\rho_1^- = \tau(\rho_3^-) \in]\sigma, 1]$ then it must be $\rho_1 \in [0, \tau(\rho_1^-)[$. In this case, since $f(\rho_1^-) = \Gamma_J - \gamma_2^{\wedge,-}$, $\gamma_1^{\max} + \gamma_2^{\wedge,-} < \Gamma_J$ and we fall in Case I.H1 with equilibrium $(\rho_1, \rho_2^-, \rho_1, \rho_4^-)$. No flux variation occurs on roads I_2 and I_4.

3.5. Interaction estimates. By the analysis of previous subsections, we have the following. Define $\mathcal{I}(\gamma, \gamma')$ to be the closed interval with extreme points γ and γ', then we have the following.

PROPOSITION 3. *For the interaction of a wave (ρ_1, ρ_1^-) with the junction J, we get $f(\rho_1^+) \in \mathcal{I}(f(\rho_1^-), f(\rho_1))$. Moreover, $|f(\rho_2^+) - f(\rho_2^-)| \leq |f(\rho_1) - f(\rho_1^-)|$.*

Let us denote by $TV_{1,3}^{\pm}(f)$ the sum of flux variations on roads I_1 and I_3 before and after the interaction and define similarly $TV_{2,4}^{\pm}(f)$. Then

$$TV_{1,3}^-(f) = f(\rho_1) - f(\rho_1^-).$$

Recalling that $f(\rho_3^{\pm}) = f(\rho_1^{\pm})$,

$$TV_{1,3}^+(f) = |f(\rho_1) - f(\rho_1^+)| + |f(\rho_3^+) - f(\rho_3^-)| = |f(\rho_1) - f(\rho_1^+)| + |f(\rho_1^+) - f(\rho_1^-)|.$$

Now, from Proposition 3, it follows that

$$TV_{1,3}^+(f) = TV_{1,3}^-(f).$$

Moreover, for the total variation of f on roads I_2 and I_4, we have

$$TV_{2,4}^+(f) = 2|f(\rho_2^+) - f(\rho_2^-)| \leq 2|f(\rho_1) - f(\rho_1^-)| = 2TV_{1,3}^-(f).$$

Finally, we get the following.

PROPOSITION 4. *The total variation of the flux on the whole junction satisfies*

$$TV(f)^+ = TV_{1,3}^+(f) + TV_{2,4}^+(f) \leq 3TV_{1,3}^-(f) = 3TV(f)^-.$$

4. Existence of solutions. The aim of this section is to prove the existence of solutions for networks with only crossing junction having (at most) two incoming and two outgoing roads. The existence is obtained via wave-front tracking algorithm.

Fix a decreasing sequence $\delta_\nu > 0$ such that $\delta_\nu \to 0$ as $\nu \to \infty$. Given an initial data $\rho_{i,0}$ on each road $I_i \in \mathcal{I}$ with bounded total variation, one approximates $\rho_{i,0}$ by a piecewise constant function with smaller total variation. Then one solves the

Riemann problems on the roads and at junctions, using Proposition 2. Moreover, rarefaction waves are replaced by rarefaction fans, i.e., collection of small shocks of size $\leq \delta_\nu$ traveling with the speed of the right state. When two waves interact inside a road, or a wave interact with a junction, then one solves a new Riemann problem and so on. The obtained weak solution ρ_ν is called a wave-front tracking approximate solution. See [3, 11] for further details.

For the wave-front tracking algorithm to work we need three basic estimates:

(i) bound on the number of waves;

(ii) bound on the number of interactions among waves and of waves with junctions;

(iii) bound on the total variation of ρ_ν.

The first two estimates are necessary to construct the wave-front tracking approximation, while the third is necessary to pass to the limit in ν and obtain a solution on the network ρ.

Therefore one has to prove (i), (ii), and (iii) for networks with crossing junctions having two incoming (and two outgoing roads). From now on, we fix a wave-front tracking approximate solution $\rho = \rho_\nu$ and provide estimates on it. Notice that (i), (ii), and (iii) hold on each road if no wave is generated by the junction J. In fact the total variation decreases in time. Thus the crucial point is to estimate the number of interactions with junctions and total variation increase due to such interactions.

4.1. Estimates on number of waves and interactions. The first estimate is achieved via a series of lemmas: the obtained results permits to avoid explosions in the number of interactions with a junction and thus on the number of waves.

We first need to recall some results from [11].

THEOREM 1. *Consider a network $(\mathcal{I}, \mathcal{J})$, and a Riemann solver RS_J, assigned for every $J \in \mathcal{J}$. Let ρ_0 be a piecewise constant (taking a finite number of values) initial datum and assume:*

(H*) *For every vertex J, consider a network formed by only the vertex J, replacing incoming and outgoing edges by infinite length ones. Then there exists a constant C_J such that, using the corresponding Riemann solver RS_J, the following holds. For every wave-front tracking approximate solution, denoting by M the number of waves in the initial datum, at most $C_J M$ waves are produced by the vertex J and there are at most $C_J M$ interactions of waves with the vertex J.*

Then, for every $T > 0$, we can construct a wave-front tracking approximate solution on $[0, T]$.

Theorem 1 essentially says that it is enough to consider each junction separately. In fact, via finite speed of waves, the interactions of waves with different junctions may happen only in a time interval whose length is bounded below.

Then for the rest of the section, we fix a network formed by a single junction J with two incoming roads with infinite length, namely I_1 and I_2, and two outgoing roads I_3 and I_4 with infinite length: our aim is to prove (H*).

Again, we recall some definitions and results from [11].

DEFINITION 3. *A wave (ρ_-, ρ_+) is a* big shock *if $\rho_- < \rho_+$ and*

$$sgn(\rho_- - \sigma) \cdot sgn(\rho_+ - \sigma) < 0.$$

DEFINITION 4. *We say that an incoming road I_i has a* good datum *at J at time $t > 0$ if*

$$\rho_i(t, b_i-) \in [\sigma, 1]$$

and a bad datum *otherwise. We say that an outgoing road I_j has a* good datum *at J at time $t > 0$ if*

$$\rho_j(t, a_j+) \in [0, \sigma]$$

and a bad datum *otherwise.*

DEFINITION 5. *For every road I_i, $i = 1, \ldots, 4$, we indicate by*

$$(\rho_-^\theta, \rho_+^\theta), \quad \theta \in \Theta = \Theta(\rho, t, i), \quad \Theta \text{ finite set,}$$

the discontinuities on road I_i at time t and by $x^\theta(t)$, $\lambda^\theta(t)$, $\theta \in \Theta$, respectively, their positions and velocities at time t. We also refer to the wave θ to indicate the discontinuity $(\rho_-^\theta, \rho_+^\theta)$.

For each discontinuity $(\rho_-^\theta, \rho_+^\theta)$ at time $\bar{t}$ on road I_i, we call $y^\theta(t)$, $t \in [\bar{t}, t_\theta]$, the trace of the wave so defined. We start with $y^\theta(\bar{t}) = x^\theta(\bar{t})$ and we continue up to the first interaction with another wave or a junction. If at time $\tilde{t}$ an interaction with a wave or a junction occurs, then either a single new wave $(\rho_-^{\tilde{\theta}}, \rho_+^{\tilde{\theta}})$ on road I_i is produced or no wave is produced. In the latter case we set $t_\theta = \tilde{t}$, otherwise we set $y^\theta(\tilde{t}) = x^{\tilde{\theta}}(\tilde{t})$ and follow $x^{\tilde{\theta}}(t)$ for $t \geq t^{\tilde{\theta}}$ up to next interaction and so on.

We have the following lemma whose proof can be found in [11, Lemma A.1.1, page 122].

LEMMA 1. *Consider a trace of a wave y^θ such that*

(a) *y^θ is generated at time $\bar{t}$ from J,*

(b) *y^θ interacts at time $\tilde{t} > \bar{t}$ with J,*

(c) *y^θ does not interact with the junction J on the interval $]\bar{t}, \tilde{t}[$.*

Then

(i) *$y^\theta(t)$ is a big shock for some $t \in]\bar{t}, \tilde{t}[$,*

(ii) *if $y^\theta(\bar{t})$ is generated on an incoming road and $y^\theta(\tilde{t}) = (\rho_l, \rho_r)$, then ρ_l is a bad datum and $f(\rho_l) < f(\rho_r)$; for outgoing roads, ρ_r is a bad datum and it holds that $f(\rho_l) > f(\rho_r)$.*

The previous lemma essentially tells us that waves produced by the junction J can come back only if they are big shocks. Now we need two additional results describing the effect of a big shock interacting back with a junction.

LEMMA 2. *Let $(\rho_{1,0}, \ldots, \rho_{4,0})$ be an equilibrium at J and $(\rho_i, \rho_{i,0})$ a wave interacting from road I_i, $i \in \{1, 3\}$, such that $f(\rho_i) < f(\rho_{i,0})$ $(f(\rho_i) > f(\rho_{i,0}))$. Call $\hat{\rho}$ the solution after the interaction, then $f(\hat{\rho}_j) \geq f(\rho_{j,0})$ $(f(\hat{\rho}_j) \leq f(\rho_{j,0}))$ for $j = 2, 4$.*

The same conclusion holds for roads I_i, $i \in \{1, 3\}$, if the wave interacts from a road I_j, $j \in \{2, 4\}$.

Proof. It is enough to reason on the space of incoming fluxes γ_i $(i = 1, 2)$. At interaction time, we move from the equilibrium $(\gamma_{1,0}, \gamma_{2,0})$ to a new equilibrium $(\hat{\gamma}_1, \hat{\gamma}_2)$. Both equilibria can be in one of the Cases I, IIa, or IIb. However, notice that the new equilibrium is always obtained from the old one either moving along the line $\sum_i \gamma_i = \Gamma_J$ or along the line $\gamma_2 = \gamma_2^\wedge$. In both cases a decrease in the flux on roads I_1 and I_3 produces an increase of the flux on roads I_2 and I_4, and vice versa. Thus the conclusion follows. ∎

LEMMA 3. *If either on road I_1 or road I_3 there is a bad datum, then nothing can happen on such roads when a big shock comes back to J from roads I_2 and I_4.*

The same conclusion holds for roads I_2 and I_4.

Proof. Assume, to fix the notation, that ρ_1 is a bad datum and a big shock $(\bar{\rho}_2, \rho_2)$ interacts back to J. From Lemma 1, $f(\bar{\rho}_2) < f(\rho_2)$. Moreover, from Lemma 2, on roads I_1 and I_3 there is an increase of the flux. However, a bad datum can give rise only to a wave decreasing the flux; see formulas (2.1), (2.2). Thus no wave can be produced on roads I_1 and I_3. ☐

LEMMA 4. *Assumption* (H*) *holds true.*

Proof. We can reason in the following way. The increase in the number of waves may happen only by interactions with the junction J.

Let y^θ, $\theta \in \Theta_0$, be the traces of waves in the initial data and every road, thus the trace ends at the first time of interaction with J. Let $t' < t''$ be an interval between two consecutive interactions of traces y^θ with J or with waves produced by J. Our aim is to estimate the number of waves in the time interval $[t', t'']$.

First, by Lemma 1, an interaction with J may happen only with a big shock coming back to J. Assume, for instance, that the first of such interactions happens for a big shock coming from road I_1 at time t_1 with $t' \leq t_1 < t''$. Then the datum on road I_1 is bad after the interaction and no wave produced by J on road I_1. Nothing can happen on road I_1, unless a big shock interacts from road I_3, say at time t_3 with $t_1 < t_3 < t''$. In this case, a bad datum is on road I_3, after time t_3, and a big shock may be produced on road I_1. After time t_3, and before time t'', by Lemma 3 no wave is produced from J on roads I_1 and I_3. In particular, the (possible) big shock on road I_1 cannot interact with any wave produced by J and also with any wave of the initial data by definition of t''. Notice that at most four new waves on roads I_2 and I_4 are produced in this way and there are at most four interactions of waves with J.

The same reasoning can be done for roads I_2 and I_4.

Finally, in the time interval $[t', t'']$ at most four new waves are produced.

Since the estimate is valid on each interval of the type $[t', t'']$, we can bound the total number of waves by $4 \cdot \#(\Theta_0)$, where $\#(\Theta_0)$ indicates the cardinality of Θ_0, i.e., the number of waves in the initial data. Similarly, the number of interactions is bounded again by $5 \cdot \#(\Theta_0)$, thus the proof is finished. ☐

4.2. BV estimates. Reasoning as for Theorem 1, it is enough to consider every junction one at a time.

So, first we consider a network formed by a single junction J with two incoming roads with infinite length, namely I_1 and I_2, and two outgoing roads I_3 and I_4 with infinite length. Then we can prove the following.

LEMMA 5. *For every wave-front tracking approximate solution ρ, we have*

$$TV(f(\rho(t))) \leq 3TV(f(\rho(0))) + 2f(\sigma). \tag{4.1}$$

Proof. First notice that only interactions with J may increase the total variation of the flux.

We need to reason on waves coming back to J. So let t_1 be the first time in which a wave generated by J comes back to J, say from road I_1. On the time interval $[0, t_1]$, waves may interact with J only from one road; otherwise to have interaction from two different roads, at least one wave should have come back to J. Then, using Proposition 4, we have

$$TV(f(\rho(t_1))) \leq 3TV(f(\rho(0))).$$

Now, necessarily on road I_1, we have a bad datum at time t_1. By Lemma 3, road I_1 does not change its status because of interactions with J of waves from roads I_2 and I_4. If a wave interacts with J from I_1, then I_1 still has a bad datum (because of the positive velocity of the wave). If a wave interacts with J from I_3, then the following happens. If the wave increases the flux, then it is simply reflected back. While a wave decreasing the flux, again because of the velocity, necessarily brings road I_3 to a bad datum. Then, we proved that after time t_1, at least one of the two roads I_1 or I_3 has a bad datum. Then, again by Lemma 3, the interactions with J from roads I_2 and I_4 can only decrease the flux variation.

On the other side, interactions with J from I_1 and I_3 can only decrease the flux, because at least one road has a bad datum. Thus, by Lemma 2, the total variation produced on roads I_2 and I_4, after time t_1, is bounded by $2f(\sigma)$. Then we conclude. $\square$

Let us now pass to the case of a complete network.

PROPOSITION 5. *Consider a network $(\mathcal{I}, \mathcal{J})$ and a wave-front tracking approximate solution ρ. Let $\delta = \min\{b_i - a_i : I_i \in \mathcal{I}\}$, i.e., the minimum length of a road, and let $\bar\lambda = \max\{|f'(0)|, |f'(\rho_{\max})|\}$, i.e., the maximum velocity of a wave. For $t \in [(n-1)\ \delta/\bar\lambda,\ n\ \delta/\bar\lambda[$, it holds that*

$$TV(f(\rho(t))) \le 3^n\ TV(f(\rho(0))) + \sum_{\ell=1}^{n} 3^\ell\ 2\ f(\sigma).$$

Proof. On a time interval of length $\Delta t = \delta/\bar\lambda$, no wave generated from a junction may interact with another junction. Thus we can reason as if junctions were isolated, and so apply Lemma 5. Then defining $TV_k = TV(f(\rho(k\Delta t)))$, we get

$$TV_k \le 3\, TV_{k-1} + 2f(\sigma),$$

and the desired estimate is readily obtained. $\square$

Remark 3. Notice that Proposition 5 says that the increase in the flux total variation is exponential in time, we can in fact rewrite the estimate as

$$TV(f(\rho(t))) \sim e^{\frac{\bar\lambda t}{\delta}\ln(3)} \left(TV(f(\rho(0))) + \frac{\bar\lambda t}{\delta} 2f(\sigma) \right).$$

4.3. Existence of solutions. Once an estimate on the flux variation is obtained, we can prove existence of solutions by the same technique of [11].

DEFINITION 6. *Consider a road network $(\mathcal{I}, \mathcal{J})$ and consider an approximate wave-front tracking solution ρ. For every road I_i, we define two curves $Y_-^{i,\rho}(t)$, $Y_+^{i,\rho}(t)$, called boundary of external flux, briefly BEF, in the following way. We set the initial condition $Y_-^{i,\rho}(0) = a_i$, $Y_+^{i,\rho}(0) = b_i$ (if $a_i = -\infty$, then $Y_-^{i,\rho} \equiv -\infty$ and if $b_i = +\infty$, then $Y_+^{i,\rho} \equiv +\infty$). We let $Y_\pm^{i,\rho}(t)$ follow the generalized characteristic as defined in* [8]*, letting $Y_-^{i,\rho}(t) = a_i$ (resp., $Y_+^{i,\rho}(t) = b_i$) if the generalized characteristic reaches the boundary and $f'(\rho(t, a_i)) < 0$ (resp., $f'(\rho(t, b_i)) > 0$). (In this way $Y_\pm^{i,\rho}(t)$ may coincide with a_i or b_i for some time intervals.) Let $\bar t$ be the first time $\bar t$ such that $Y_-^{i,\rho}(\bar t) = Y_+^{i,\rho}(\bar t)$ (possibly $\bar t = +\infty$), then we let $Y_\pm^{i,\rho}$ be defined on $[0, \bar t]$. Finally, we define the sets*

$$D_1^i(\rho) = \left\{(t,x) : t \in [0, \bar t) : Y_-^{i,\rho}(t) \le x \le Y_+^{i,\rho}(t)\right\}$$

and

$$D_2^i(\rho) = [0, +\infty) \times [a_i, b_i] \setminus D_1^i(\rho).$$

Clearly $Y_\pm^i(t)$ bound the set on which the datum is not influenced by the other roads through the junctions.

LEMMA 6. *The curves $t \mapsto Y_{\pm}^i(t)$ are uniformly Lipschitz.*

Proof. The curves $Y_{\pm}^i$ are generalized characteristics, thus their velocity is uniformly bounded by $\max\{f'(0), |f'(1)|\}$. ∎

LEMMA 7. *For every $t \geq 0$, there exist at most two big waves on*

$$\left\{x : (t,x) \in D_2^i(\rho)\right\} \subseteq [a_i, b_i].$$

Proof. A big wave can originate at time t on road I_i from J only if road I_i has a bad datum at J at time t. If this happens, then road I_i does not have a bad datum at J up to the time in which a big wave is absorbed from I_i. Then we reach the conclusion. ∎

We are able to state and prove the main result.

THEOREM 2. *Fix a road network $(\mathcal{I}, \mathcal{J})$. Given $C > 0$ and $T > 0$, there exists an admissible solution defined on $[0,T]$ for every initial data $\bar{\rho} \in \mathrm{cl}\{\rho : TV(\rho) \leq C\}$, where* cl *indicates the closure in L^1_{loc}.*

Proof. We fix a sequence of initial data $\bar{\rho}_\nu$ piecewise constant such that $TV(\bar{\rho}_\nu) \leq C$ for every $\nu \geq 0$ and $\bar{\rho}_\nu \to \bar{\rho}$ in L^1_{loc} as $\nu \to +\infty$. For each $\bar{\rho}_\nu$ we consider an approximate wave-front tracking solution ρ_ν such that $\rho_\nu(0,x) = \bar{\rho}_\nu(x)$ and rarefactions are split in rarefaction shocks of size $\frac{1}{\nu}$.

For every road I_i, we notice that on $D_1^i(\rho_\nu)$, ρ_ν is not influenced by other roads and so the estimates of [3] hold. Since the curves $Y_{\pm}^{i,\rho_\nu}$ are uniformly Lipschitz continuous, they converge uniformly in time, up to a subsequence, to a limit curve. Therefore the regions $D_1^i(\rho_\nu)$ converge in measure to a limit region D_1^i. More precisely, we have the estimate

$$\mathrm{meas}(D_1^i(\rho_{\nu+1}) \Delta D_1^i(\rho_\nu)) \leq T \cdot \|Y_{\pm}^{i,\nu+1} - Y_{\pm}^{i,\nu}\|_{L^\infty},$$

where Δ is the symmetric difference of sets, namely $A \Delta B = (A \setminus B) \cup (B \setminus A)$. Then $\rho_\nu \to \rho$ in L^1_{loc} on D_1^i with ρ admissible solution to the Cauchy problem.

On $D_2^i := [0, +\infty[\times[a_i, b_i] \setminus D_1^i$, we have that, up to a subsequence, $\rho_\nu \rightharpoonup^* \rho$ weak* on L^1 and, using Proposition 5, $f(\rho_\nu) \to \bar{f}$ in L^1 for some $\bar{f}$. By Lemma 7, there are at most two big waves on D_2^i for every time, hence, splitting the domain D_2^i into a finite number of pieces where we can invert the function f, getting $\rho_\nu \to f^{-1}(\bar{f})$ in L^1. Together with $\rho_\nu \rightharpoonup^* \rho$ weak* on L^1, we conclude that $\rho_\nu \to \rho$ strongly in L^1.

The other requirements of the definition of admissible solution are clearly satisfied. ∎

5. Summary. We considered a fluid dynamic model for traffic flow on networks. On each road we used the established Lighthill–Witham–Richards model, while special type of junctions, called crossing junctions, are treated, coming from modeling of real junctions of T type.

We first defined a Riemann solver at crossing junctions, taking into account the maximal load of the junction and equilibrium flux proportions among incoming roads. Then we established existence of solutions on the whole network by a wave-front tracking algorithm.

Since the flux variation is not conserved for interactions of waves with junctions (see Proposition 4), we cannot expect Lipschitz continuous dependence of solutions; see [11, section 5.4, page 111].

Acknowledgment. The authors thanks the anonymous referees for indications to improve the paper.

REFERENCES

[1] D. Armbruster, P. Degond, and C. Ringhofer, *A model for the dynamics of large queuing networks and supply chains*, SIAM J. Appl. Math., 66 (2006), pp. 896–920.

[2] M. Banda, M. Herty, and A. Klar, *Gas flow in pipeline networks*, Networks and Heterogenous Media, 1 (2006), pp. 41–56.

[3] A. Bressan, *Hyperbolic Systems of Conservation Laws—The One-Dimensional Cauchy Problem*, Oxford University Press, Oxford, 2000.

[4] G. Bretti, R. Natalini, and B. Piccoli, *Numerical approximations of a traffic flow model on networks*, Networks and Heterogeneous Media, 1 (2006), pp. 57–84.

[5] N. Cascone, C. D'Apice, B. Piccoli, and L. Raritá, *Optimization of traffic on road networks*, M3AS Math. Models Methods Appl. Sci., 17 (2007), pp. 1587–1617.

[6] Y. Chitour and B. Piccoli, *Traffic circles and timing of traffic lights for cars flow*, Discrete Contin. Dyn. Syst. Ser. B, 5 (2005), pp. 599–630.

[7] G. Coclite, M. Garavello, and B. Piccoli, *Traffic flow on road networks*, SIAM J. Math. Anal., 36 (2005), pp. 1862–1886.

[8] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Berlin, 1999.

[9] C. D'Apice, R. Manzo, and B. Piccoli, *Packets flow on a telecommunication network*, SIAM J. Math. Anal., 38 (2006), pp. 717–740.

[10] M. Garavello and B. Piccoli, *Traffic flow on a road network using the Aw-Rascle model*, Comm. Partial Differential Equations, 31 (2006), pp. 243–275.

[11] M. Garavello and B. Piccoli, *Traffic Flow on Road Networks*, Applied Math Series 1, American Institute of Mathematical Sciences, Springfield, 2006.

[12] D. Helbing, J. Siegmeier, and S. Lämmer, *Self-organized network flows*, Networks and Heterogeneous Media, 2 (2007), pp. 193–210.

[13] D. Helbing, S. Lämmer, and J.-P. Lebacque, *Self-organized control of irregular or perturbed network traffic*, in Optimal Control and Dynamic Games, C. Deissenberg and R. F. Hartl, eds., Springer, Dordrecht, 2005, pp. 239–274.

[14] M. Herty and A. Klar, *Modelling, simulation and optimization of traffic flow networks*, SIAM J. Sci. Comp., 25 (2003), pp. 1066–1087.

[15] M. Herty, S. Moutari, and M. Rascle, *Optimization criteria for modelling intersections of vehicular traffic flow*, Networks and Heterogeneous Media, 1 (2006), pp. 275–294.

[16] H. Holden and N. H. Risebro, *A mathematical model of traffic flow on a network of unidirectional roads*, SIAM J. Math. Anal., 26 (1995), pp. 999–1017.

[17] M. M. Khoshyaran and J. P. Lebacque, *A reactive dynamic assignment scheme*, in Mathematics in Transport Planning and Control, J. D. Griffith, ed., Pergamon, 1998.

[18] J. P. Lebacque and M. M. Khoshyaran, *Macroscopic flow models (First order macroscopic traffic flow models for networks in the context of dynamic assignment)*, in Transportation Planning, the State of the Art, M. Patriksson and M. Labbé, ed., Klüwer Academic Press, Norwell, MA, 2002.

[19] J. P. Lebacque and J. B. Lesort, *The Godunov scheme and what it means for first order traffic flow models*, in Proceedings of the Internaional Symposium on Transportation and Traffic Theory No. 13, Lyon, Pergamon Press, Oxford, 1996, pp. 647–677.

[20] M. J. Lighthill and G. B. Whitham, *On kinetic waves.* II. *Theory of traffic flows on long crowded roads*, Proc. Roy. Soc. London Ser. A, 229 (1955), pp. 317–345.

[21] A. Marigo, *Optimal distribution coefficients for telecommunication networks*, Networks and Heterogeneous Media, 1 (2006), pp. 315–336.

[22] A. Marigo and B. Piccoli, *Optimal distribution coefficients for packets traffic on a telecommunication network*, in Proc. IEEE Joint Int. Conference on Decision and Control and European Control Conference (CDC - ECC), 2005.

[23] P. I. Richards, *Shock waves on the highway*, Oper. Res., 4 (1956), pp. 42–51.

SIAM J. MATH. ANAL.
Vol. 39, No. 6, pp. 2033–2052

HOPF BIFURCATION FROM VISCOUS SHOCK WAVES*

BJÖRN SANDSTEDE† AND ARND SCHEEL‡

Abstract. Using spatial dynamics, we prove a Hopf bifurcation theorem for viscous Lax shocks in viscous conservation laws. The bifurcating viscous shocks are unique (up to time and space translation), exponentially localized in space, periodic in time, and their speed satisfies the Rankine–Hugoniot condition. We also prove an "exchange of spectral stability" result for super- and subcritical bifurcations and outline how our proofs can be extended to cover degenerate, over-, and undercompressive viscous shocks.

Key words. viscous conservation law, Lax shock, Hopf bifurcation

AMS subject classifications. 35L65, 35B32, 35L67

DOI. 10.1137/060675587

1. Introduction. The purpose of this paper is to study Hopf bifurcation from viscous shock waves. While Hopf bifurcations from equilibria are well understood in ordinary differential equations (ODEs) and in dissipative partial differential equations (PDEs) on bounded domains, a variety of new phenomena and difficulties arise when studying Hopf bifurcations for PDEs on unbounded domains.

In particular, Hopf bifurcations from travelling waves are complicated by the presence of a neutral mode at the origin which is induced by spatial translation. If the essential spectrum of the linearization around the travelling wave is bounded away from the imaginary axis, appropriate center-manifold reductions and equivariant parametrizations as in [4, 6, 17] show that the bifurcation problem reduces to a standard Hopf bifurcation, and standard results on bifurcation and exchange of stability [2] immediately carry over to this setting [15, section 2]; the only effect of the translation mode is an adjustment of the wave speed. When the Hopf instability is caused by essential spectrum that crosses the imaginary axis, a variety of interesting new phenomena can occur, including failure of bifurcation [15] and bifurcation of multiple solution branches [16, section 2.3]. The situation becomes more involved when the instability caused by the essential spectrum is stationary, as the wave will then typically decay only algebraically at onset which leads to significant complications in the analysis [16, sections 2.1, 2.2, and 3].

From the preceding list, one can easily envision yet another possible scenario where the complex Hopf eigenvalues belong to the point spectrum, whilst the translation mode is embedded in the continuous spectrum. This situation arises, for instance, when the primary wave is not spatially localized, but the Hopf eigenfunctions are localized: Examples are Hopf bifurcations from coherent structures such as sources and sinks in one spatial dimension, and spiral waves in two dimensions. A model problem in higher space dimensions, but with a space-dependent potential, has recently been analyzed in [1]. Viscous shock waves provide another prominent example where the translation mode is embedded into the continuous spectrum. In fact, conservation

*Received by the editors November 21, 2006; accepted for publication (in revised form) September 25, 2007; published electronically March 21, 2008.
http://www.siam.org/journals/sima/39-6/67558.html

†Department of Mathematics, University of Surrey, Guildford, GU2 7XH, UK (b.sandstede@surrey.ac.uk). This author gratefully acknowledges a Royal Society Wolfson Research Merit Award.

‡School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (scheel@math.umn.edu). This author was partially supported by the NSF grant DMS-0203301.

laws can often be derived formally and rigorously in spatially extended systems where the primary pattern breaks the underlying continuous symmetry [3, 7].

In this article, we investigate Hopf bifurcations from viscous shock waves using the spatial-dynamics techniques we developed in [13, 14, 15] for Hopf bifurcations from fronts and pulses in reaction-diffusion systems. Our paper is strongly motivated by recent work of Texier and Zumbrun [19, 20] in which they analyzed oscillatory instabilities of viscous shocks using delicate estimates for the temporal period map of the linearized semigroup. Texier and Zumbrun proved the existence of a continuous branch of oscillatory viscous shocks with a $1/x$ decay estimate at spatial infinity. In a personal communication, Zumbrun asked us whether spatial-dynamics techniques can be used to obtain the same or stronger results than those in [19, 20]. We demonstrate here that the spatial-dynamics approach yields indeed sharper results, while simplifying the analysis and adding geometric insight into the problem: We show that the bifurcating oscillatory shocks are unique, exponentially localized, and depend smoothly on the bifurcation parameter, and we calculate the spectra of the linearization about the bifurcating oscillating shock waves, thereby confirming the expected exchange of stability. Instead of analyzing the temporal semigroup whose linearization has essential spectrum up to the imaginary axis, we consider the spatial evolution of temporally periodic functions for which we gain compactness of the resolvent due to the imposed time periodicity. While this method may appear nonintuitive, it is completely analogous to the usual phase-plane analysis used to prove existence of viscous shocks and to study their stationary bifurcations. After this paper was completed, Texier and Zumbrun were able to extend their approach to prove in [21] exponential localization of the bifurcating solutions for compressible Navier–Stokes and magnetohydrodynamics.

Outline. In section 2, we state our main result on bifurcation and spectral stability of modulated shocks. The bifurcation result is proved in section 3. In section 4, we review the precise characterization of spectra and prove stability and instability in the case of super- and subcritical bifurcations, respectively. We conclude with a discussion of several extensions and generalizations in section 5.

2. Setup and main results. Consider the viscous conservation law

$$u_t + f(u)_y = u_{yy}, \qquad y \in \mathbb{R},\ u \in \mathbb{R}^n, \tag{2.1}$$

where f is a smooth flux function. We are interested in viscous shocks $q^0(y - c^0 t)$ which connect the constant rest states $u^0_\pm$ at $y = \pm\infty$ so that

$$\lim_{x\to\pm\infty} q^0(x) = u^0_\pm.$$

Viscous shocks are stationary solutions in the moving reference frame $x = y - ct$ in which (2.1) becomes

$$u_t = \partial_x \left[u_x + cu - f(u)\right], \qquad x \in \mathbb{R},\ u \in \mathbb{R}^n, \tag{2.2}$$

and they therefore satisfy the integrated steady state equation

$$u_x = [f(u) - f(u^0_-)] - c[u - u^0_-], \tag{2.3}$$

where the speed c is given necessarily by the Rankine–Hugoniot condition

$$c = \frac{f_j(u^0_+) - f_j(u^0_-)}{u^0_{+,j} - u^0_{-,j}}, \qquad j = 1, \ldots, n. \tag{2.4}$$

In particular, $q^0(x)$ lies in the intersection of the unstable manifold $\widetilde{W}^{\mathrm{u}}_-$ of u^0_- and the stable manifold $\widetilde{W}^{\mathrm{s}}_+$ of u^0_+ for (2.3).

The most common viscous shocks are Lax shocks for which $u^0_\pm$ are hyperbolic equilibria of (2.3) with $\dim \widetilde{W}^{\mathrm{u}}_- = p+1$ and $\dim \widetilde{W}^{\mathrm{s}}_+ = n-p$ for some $p \in \{0, \dots, n-1\}$. We assume that the intersection of $\widetilde{W}^{\mathrm{u}}_-$ and $\widetilde{W}^{\mathrm{s}}_+$ is transverse along q^0 and that the Jacobian $f_u(u^0_\pm)$ has only real and distinct eigenvalues. If $u^\varepsilon_\pm$ are smooth curves that depend on a real parameter $\varepsilon \approx 0$, then we find a smooth family of Lax shocks $q^\varepsilon(x)$ with a smooth speed relation $c = c^\varepsilon$ given by the Rankine–Hugoniot condition. Since the eigenvalues of $f_u(u)$ are the characteristic speeds of propagation at u, the condition on the dimensions of $\widetilde{W}^{\mathrm{s,u}}_\pm$ merely states that $p+1$ characteristics enter the shock from the left and $n-p$ characteristics enter from the right.

We are interested in the scenario where the Lax shocks undergo a Hopf instability upon increasing ε through zero. We therefore consider the linearization at the shock which is given by the linear operator

$$\mathcal{L}^\varepsilon := \partial_x \left[\partial_x + c^\varepsilon - f_u(q^\varepsilon(x)) \right],$$

which we view as a closed unbounded operator on $L^2(\mathbb{R}, \mathbb{R}^n)$. Its essential spectrum is readily seen to be contained in the closed left half-plane, touching the imaginary axis only at the origin with a quadratic tangency. We assume that the point spectrum lies in the open left half-plane, bounded away from the imaginary axis, except for an isolated pair $\lambda(\varepsilon)$ and $\overline{\lambda(\varepsilon)}$ of simple complex eigenvalues with

$$\lambda(0) = \mathrm{i}\omega_0 \neq 0, \qquad \operatorname{Re} \lambda_\varepsilon(0) > 0. \tag{2.5}$$

THEOREM 2.1 (bifurcation). *Under the above assumptions, there are positive constants K, η, and δ and a smooth function*

$$\begin{aligned} [0, \delta) &\longrightarrow \mathcal{C}^2_{\mathrm{unif}}(\mathbb{R} \times S^1, \mathbb{R}^n) \times \mathbb{R} \times \mathbb{R}, \\ a &\longmapsto (q^*(a), \varepsilon(a), \omega(a)), \end{aligned}$$

so that $u^*(x,t;a) := q^*(x, \omega(a)t; a)$ satisfies (2.2) with $c = c^{\varepsilon(a)}$ for all a,

$$|q^*(x,\tau;a) - u^{\varepsilon(a)}_\pm| \leq K \mathrm{e}^{-\eta|x|}, \qquad q^*(x,\tau;0) = q^0(x), \qquad \omega(0) = \omega_0,$$

and $q^*(x,\cdot;a)$ has minimal period 2π in τ for each $a > 0$. Furthermore, any non-stationary time-periodic solution $u(x,t)$ of (2.2), which is pointwise close to $q^0(x)$ and converges to $u^\varepsilon_\pm$ as $x \to \pm\infty$, is in fact an appropriate space and time translation of u^*.

Note that $u^*(x,t;a)$ and $q^{\varepsilon(a)}(x)$ have the same asymptotic rest states and travel with the same (average) wave speed. Theorem 2.1 remains true if $f = f(u;\varepsilon)$ depends smoothly on the parameter ε.

Spectral stability of the modulated shocks $u^*(x,t;a)$ is determined by the Floquet spectrum

$$\Sigma = \{\lambda \in \mathbb{C};\ \mathrm{e}^{\lambda T} \in \text{ spectrum of } \Phi_T\},$$

where $T = 2\pi/\omega$ is the temporal period of u^*, and Φ_t is the evolution operator of the linearization

$$v_t = \partial_x[\partial_x + c^{\varepsilon(a)} - f_u(u^*(x,t;a))]v$$

of (2.2) about u^* on L^2 or $\mathcal{C}^0_{\mathrm{unif}}$.

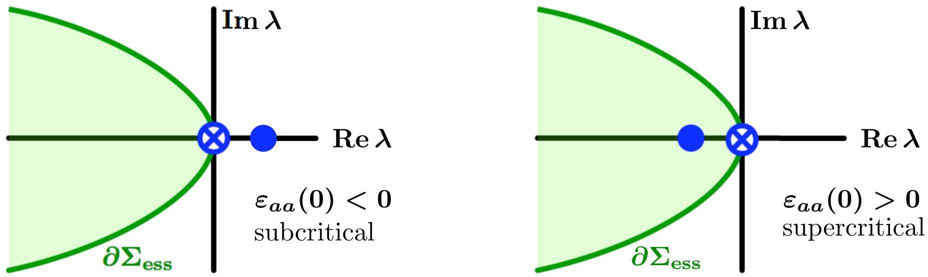


FIG. 1. *The Floquet spectra of the oscillatory shocks u^* from Theorem* 2.1 *is shown for $a > 0$: $\lambda = 0$ has geometric and algebraic multiplicity two, while the location of the remaining simple Floquet exponent near the origin depends on the sign of $\varepsilon_{aa}(0)$.*

THEOREM 2.2 (stability). *Assume that the hypotheses of Theorem* 2.1 *are met and that the Evans function associated with q^0 has a simple zero at the origin (see section* 4 *for details). If $\varepsilon_{aa}(0) \neq 0$, then the Floquet spectrum Σ of the oscillatory shock u^* given in Theorem* 2.1 *is as indicated in Figure* 1*.*

We refer the reader to [5, section 3.4] and [23, sections 9, 10] for explicit conditions which imply that the Evans function of q^0 has a simple zero at the origin.

3. Existence of modulated viscous shocks. In this section, we prove Theorem 2.1.

3.1. Preparations. We begin by collecting some properties of the linearization

$$\mathcal{L}^\varepsilon = \partial_x \left[\partial_x + c^\varepsilon - f_u(q^\varepsilon(x))\right]$$

about the viscous shocks that we need later on. Since we assumed in (2.5) that the Hopf eigenvalues $\lambda(0) = \mathrm{i}\omega_0 \neq 0$ and $\overline{\lambda(0)}$ of $\mathcal{L}^0$ are simple, we know that there are nonzero L^2-functions v_j and ψ_j for $j = 1, 2$ that form a basis of the eigenspaces of $\mathcal{L}^0$ and its adjoint $[\mathcal{L}^0]^*$, respectively, associated with these Hopf eigenvalues. We can choose these functions so that

$$\begin{aligned} \langle \psi_i, v_j\rangle_{L^2} = \langle \psi_i, \psi_j\rangle_{L^2} &= \delta_{ij}, \\ \mathcal{L}^0 v_1 = -\omega_0 v_2, \qquad \mathcal{L}^0 v_2 &= \omega_0 v_1. \end{aligned} \tag{3.1}$$

The result [10, Theorem 5.4 in Chapter II] gives the characterization

$$\mathrm{Re}\,\lambda_\varepsilon(0) = \frac{1}{2}\sum_{j=1}^{2} \langle \psi_j, \partial_\varepsilon \mathcal{L}^\varepsilon|_{\varepsilon=0} v_j\rangle_{L^2} \tag{3.2}$$

of the derivative $\mathrm{Re}\,\lambda_\varepsilon(0)$ which we assumed in (2.5) to be positive.

3.2. Spatial dynamics. To find time-periodic solutions of (2.2), we rescale time $\tau := \omega t$ to get

$$\omega\partial_\tau u + f(u)_x - cu_x = u_{xx},$$

which we then cast as the first-order system

$$\begin{pmatrix} u_x \\ v_x \end{pmatrix} = \begin{pmatrix} v \\ \omega\partial_\tau u + f(u)_x - cv \end{pmatrix} = \begin{pmatrix} v \\ \omega\partial_\tau u + f_u(u)v - cv \end{pmatrix}. \tag{3.3}$$

We view (3.3) as an equation for $U = (u, v)$ in $Y = H^1(S^1) \times H^{1/2}(S^1)$ with $S^1 = [0, 2\pi]/\sim$. The space Y is natural for various reasons: First, we wish to work on spaces of time-periodic functions and shall see later in (3.13), see also [13, section 3], that the spaces $H^{s+1/2}(S^1) \times H^s(S^1)$ for $s \geq 0$ are the only spaces compatible with the linear leading-order part of (3.3). We choose $s = 1/2$ since $H^1(S^1)$ embeds into $C^0(S^1)$ which guarantees that the nonlinearity $f_u(u)$ in (3.3) is well defined and differentiable on Y. Finally, we remark that we shall often use, for convenience, the scalar product

$$\langle U, V\rangle_X := \frac{1}{2\pi}\int_0^{2\pi} \langle U(\tau), V(\tau)\rangle_{\mathbb{R}^{2n}}\, \mathrm{d}\tau$$

of the space $X = L^2(S^1) \times L^2(S^1)$ to define complements and compute adjoints; this scalar product is also an inner product on Y as Y embeds continuously into X.

The system (3.3) is invariant under the S^1-action

$$\Gamma : S^1 \longrightarrow L(Y, Y), \qquad \sigma \longmapsto \Gamma_\sigma, \qquad [\Gamma_\sigma U](\tau) = U(\tau - \sigma). \tag{3.4}$$

We record that the fixed-point space $\operatorname{Fix}\Gamma \cong \mathbb{R}^n \times \mathbb{R}^n$ of this action consists precisely of all time-independent functions, and (3.3) restricted to $\operatorname{Fix}\Gamma$ becomes the usual travelling-wave ODE

$$\begin{pmatrix} u_x \\ v_x \end{pmatrix} = \begin{pmatrix} v \\ f_u(u)v - cv \end{pmatrix} \tag{3.5}$$

which is equivalent to (2.3). Equation (3.5) possesses the equilibria $U_{\mathrm{eq}} = (u, 0)$ for $u \in \mathbb{R}^n$ and the heteroclinic orbits $Q^\varepsilon(x) := (q^\varepsilon, q^\varepsilon_x)(x)$ for $c = c^\varepsilon$: $Q^\varepsilon(x)$ connects $U^-_{\mathrm{eq}}(\varepsilon) = (u^\varepsilon_-, 0)$ to $U^+_{\mathrm{eq}}(\varepsilon) = (u^\varepsilon_+, 0)$ with

$$T_{Q^\varepsilon(x)} W^{\mathrm{u}}_- + T_{Q^\varepsilon(x)} W^{\mathrm{cs}}_+ = \mathbb{R}^{2n}, \tag{3.6}$$

where $W^j_\pm := W^j(U^\pm_{\mathrm{eq}}(\varepsilon))$. The transversality of the intersection in (3.6) is a consequence of the following dimension count for $\varepsilon = 0$. Since $\dim W^{\mathrm{u}}_- = p + 1$ and $\dim W^{\mathrm{cs}}_+ = 2n - p$, it suffices to show that the only nontrivial elements in the intersection of the tangent spaces are multiples of $Q^0_x(x)$. This, in turn, can be seen as follows. Starting with any nontrivial bounded solution $(u, v)(x)$ of the variational equation

$$\begin{pmatrix} u_x \\ v_x \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ f_{uu}(q^0)[q^0_x, \cdot] & f_u(q^0) - c^0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \tag{3.7}$$

of (3.5) about $Q^0 = (q^0, q^0_x)$, we find that its first component $u(x)$ is a nontrivial bounded solution of

$$u_{xx} = [(f_u(q^0) - c^0)u]_x. \tag{3.8}$$

If a nontrivial bounded solution $(u, v)(x)$ of (3.7) lies, in addition, in $T_{Q^\varepsilon(x)} W^{\mathrm{u}}_-$, then $u(x)$ decays exponentially at $x = -\infty$ and is therefore also a nontrivial bounded solution of the variational equation

$$u_x = [f_u(q^0) - c^0]u \tag{3.9}$$

of (2.3) about q^0 since we can integrate (3.8) once, starting at $x=-\infty$. To complete the argument, we recall our assumption that q^0 lies in the transverse intersection of the unstable and stable manifolds of the hyperbolic rest states $u^0_\pm$ of (2.3); this hypothesis implies that each nontrivial bounded solution of (3.9) is necessarily exponentially localized and must, in fact, be a multiple of q^0_x as claimed.

Next, we linearize the full system (3.3) in the solution $Q^0=(q^0,q^0_x)$ for $\omega=\omega_0$ to get

$$V_x=\begin{pmatrix} 0 & 1 \\ \omega_0\partial_\tau+f_{uu}(q^0)[q^0_x,\cdot] & f_u(q^0)-c^0 \end{pmatrix}V, \qquad V\in Y. \tag{3.10}$$

For $x\to\pm\infty$, we obtain the asymptotic systems

$$V_x=\begin{pmatrix} 0 & 1 \\ \omega_0\partial_\tau & f_u(u^0_\pm)-c^0 \end{pmatrix}V, \qquad V\in Y, \tag{3.11}$$

whose properties we discuss first. Equations (3.10) and (3.11) leave each subspace $Y_k:=\{\mathrm{e}^{\mathrm{i}k\tau}\hat V;\ \hat V\in\mathbb{C}^{2n}\}$ invariant for $k\in\mathbb{Z}$. If we restrict (3.11) to Y_k, we obtain the system

$$\hat V_x=\begin{pmatrix} 0 & 1 \\ \mathrm{i}k\omega_0 & f_u(u^0_\pm)-c^0 \end{pmatrix}\hat V, \qquad \hat V\in\mathbb{C}^{2n}, \tag{3.12}$$

where $V=\mathrm{e}^{\mathrm{i}k\tau}\hat V$. For $k\neq 0$, the matrices in (3.12) are hyperbolic: $\nu=\mathrm{i}\kappa$ is an eigenvalue if and only if $\det[-\kappa^2-\mathrm{i}\kappa(f_u(u^0_\pm)-c^0)-\mathrm{i}k\omega_0]=0$, which is excluded since $f_u(u^0_\pm)$ was assumed to have only real eigenvalues.[1] For $|k|\to\infty$, the eigenvalues of the matrices in (3.12) are

$$\nu_j=\pm\sqrt{\mathrm{i}\omega_0 k}(1+\mathrm{O}(|k|^{-1/2})) \qquad \text{with eigenfunction } \begin{pmatrix}\nu_j e_j\\ e_j\end{pmatrix}, \tag{3.13}$$

where e_j denotes the canonical basis in $\mathbb{R}^n$. In particular, the stable and unstable eigenspaces have a uniform angle in $H^1(S^1)\times H^{1/2}(S^1)$ as $|k|\to\infty$, and therefore for all $k\neq 0$; see also [13, Lemma 3.3]. Thus, we can apply the results in [11, 13] to conclude that (3.11) restricted to $Y_\mathrm{h}:=\overline{\bigoplus_{k\neq 0}Y_k}$ has exponential dichotomies $\Phi^{\mathrm{s,u}}_{\pm,\mathrm{h}}(x,y)$ on $\mathbb{R}^\pm$ since the perturbation

$$\begin{pmatrix} 0 & 0 \\ f_{uu}(q^0(x))[q^0_x(x),\cdot] & f_u(q^0(x))-f_u(u^0_\pm) \end{pmatrix}: H^1\times H^{1/2}\longrightarrow H^1\times H^{1/2}$$

is bounded independently of x and converges to zero as $|x|\to\infty$. We define

$$\nu^\mathrm{s}_\pm:=-\frac{1}{2}\sup\left\{\mathrm{Re}\,\nu_j;\ \mathrm{Re}\,\nu_j<0,\ \nu_j \text{ is an eigenvalue of } (3.12)_\pm \text{ for some } k\in\mathbb{Z}\right\},$$
$$\nu^\mathrm{u}_\pm:=\frac{1}{2}\inf\left\{\mathrm{Re}\,\nu_j;\ \mathrm{Re}\,\nu_j>0,\ \nu_j \text{ is an eigenvalue of } (3.12)_\pm \text{ for some } k\in\mathbb{Z}\right\},$$

---

[1] Purely imaginary spatial eigenvalues $\nu=\mathrm{i}\kappa$ are actually equivalent to essential spectrum at $\lambda=\mathrm{i}\omega_0 k$ so that, for more general viscosity matrices and fluxes, the analysis goes through provided the Hopf eigenvalue $\mathrm{i}\omega_0$ is *not resonant* with essential spectrum on the imaginary axis.

and observe that $\nu^{\mathrm{s}}_{\pm}, \nu^{\mathrm{u}}_{\pm} > 0$ due to (3.13). The spaces

$$E^{\mathrm{cs}}_{+} = \left\{ V_0 \in Y;\ \exists \text{ solution } V(x) \text{ of (3.10) on } \mathbb{R}^+ \text{ with } V(0) = V_0,\ \sup_{x\geq 0} |V(x)| < \infty \right\},$$

$$E^{\mathrm{u}}_{-} = \left\{ V_0 \in Y;\ \exists \text{ solution } V(x) \text{ of (3.10) on } \mathbb{R}^- \text{ with } V(0) = V_0, \right.$$
$$\left. \sup_{x\leq 0} |V(x)| \mathrm{e}^{\nu^{\mathrm{u}}_{-}|x|} < \infty \right\}$$

are closed subspaces of Y.

CLAIM. *We have*

$$E^{\mathrm{cs}}_{+} \cap E^{\mathrm{u}}_{-} = \mathbb{R}Q^0_x(0) \oplus \mathbb{R}V_1(0) \oplus \mathbb{R}V_2(0), \tag{3.14}$$
$$Y = [E^{\mathrm{cs}}_{+} + E^{\mathrm{u}}_{-}] \oplus \mathbb{R}\Psi_1(0) \oplus \mathbb{R}\Psi_2(0), \tag{3.15}$$

where, using the definitions of v_j and ψ_j from section 3.1*,*

$$\begin{aligned} V_1(x) &:= \cos\tau \begin{pmatrix} v_1 \\ \partial_x v_1 \end{pmatrix}(x) + \sin\tau \begin{pmatrix} v_2 \\ \partial_x v_2 \end{pmatrix}(x), \\ V_2(x) &:= -\sin\tau \begin{pmatrix} v_1 \\ \partial_x v_1 \end{pmatrix}(x) + \cos\tau \begin{pmatrix} v_2 \\ \partial_x v_2 \end{pmatrix}(x) \end{aligned} \tag{3.16}$$

and

$$\begin{aligned} \Psi_1(x) &:= \cos\tau \begin{pmatrix} \widetilde{\psi}_1 \\ \psi_1 \end{pmatrix}(x) + \sin\tau \begin{pmatrix} \widetilde{\psi}_2 \\ \psi_2 \end{pmatrix}(x), \\ \Psi_2(x) &:= -\sin\tau \begin{pmatrix} \widetilde{\psi}_1 \\ \psi_1 \end{pmatrix}(x) + \cos\tau \begin{pmatrix} \widetilde{\psi}_2 \\ \psi_2 \end{pmatrix}(x) \end{aligned} \tag{3.17}$$

with $\widetilde{\psi}_j := -\partial_x \psi_j - [f_u^T(q^0) - c^0]\psi_j$ for $j = 1, 2$.

Proof. The characterization of E^{cs}_{+} and E^{u}_{-} is a consequence of the existence of exponential dichotomies on Y_{h} and the dynamics of the travelling-wave ODE (3.5). First, recall that the dynamics on the Fourier subspaces Y_k decouple, so that we can write

$$E^{\mathrm{cs}}_{+} = \bigoplus_{k\in\mathbb{Z}} (E^{\mathrm{cs}}_{+} \cap Y_k), \qquad E^{\mathrm{u}}_{-} = \bigoplus_{k\in\mathbb{Z}} (E^{\mathrm{u}}_{-} \cap Y_k).$$

We know that the strong unstable manifold $W^{\mathrm{u}}(U^{-}_{\mathrm{eq}}(0))$ and the center-stable manifold $W^{\mathrm{cs}}(U^{+}_{\mathrm{eq}}(0))$ of (3.5) intersect transversely along $Q^0(x)$; see (3.6). Thus,

$$\operatorname{span} Q^0_x(0) = E^{\mathrm{cs}}_{+} \cap E^{\mathrm{u}}_{-} \cap Y_0.$$

Next, $V_0 \in Y_{\mathrm{h}}$, the subspace of nonzero Fourier modes $k \neq 0$, lies in $E^{\mathrm{cs}}_{+} \cap E^{\mathrm{u}}_{-}$ if and only if $V(x)$ satisfies (3.11) on $\mathbb{R}$ with $V(x) \to 0$ exponentially as $|x| \to \infty$. Since (3.11) on Y decouples, we find that such a solution can be taken in the form $V(x) = \mathrm{e}^{\mathrm{i}k\tau}(v, v_x)(x)$ for some integer $k \neq 0$. In particular, $v(x)$ satisfies

$$\mathcal{L}^0 v = \mathrm{i}k\omega_0 v,$$

and is therefore an L^2-eigenfunction of $\mathcal{L}^0$ to the eigenvalue $\lambda = \mathrm{i}k\omega_0$. Inspecting our hypotheses on $\mathcal{L}^0$, (3.14) follows. To prove (3.15), we consider the adjoint equation

$$\Psi_x = -\begin{pmatrix} 0 & -\omega_0\partial_\tau + f_{uu}^T(q^0)[q_x^0,\cdot] \\ 1 & f_u^T(q^0) - c^0 \end{pmatrix}\Psi, \qquad \Psi \in Y, \tag{3.18}$$

of (3.11), taken with respect to the inner product in the space $X = L^2(S^1) \times L^2(S^1)$. We note that the functions $\Psi_j(x)$ from (3.17) satisfy (3.18). A calculation shows that

$$\frac{\mathrm{d}}{\mathrm{d}x}\langle V(x), \Psi(x)\rangle_X = 0 \qquad \text{for all } x \in \mathbb{R}$$

for solutions $V(x)$ of (3.11) and $\Psi(x)$ of (3.18); see [14] for similar arguments. Using the relation between (3.18) and $[\mathcal{L}^0]^*$, we conclude that (3.15) is met. □

Note that the direction $Q_x^0(0) \in E_+^{\mathrm{cs}} \cap E_-^{\mathrm{u}}$ corresponds to the flow direction. To remove it, we shall later use the hyperplane

$$\mathcal{S} := [\mathbb{R}Q_x^0(0)]^\perp \subset Y. \tag{3.19}$$

We are now ready to discuss the nonlinear equation (3.3) near the orbit $Q^0(x)$ for ω close to ω_0 and ε close to zero. It is convenient to set

$$\omega = \omega_0 + \Omega$$

and to consider

$$\begin{pmatrix} u_x \\ v_x \end{pmatrix} = \begin{pmatrix} v \\ (\omega_0 + \Omega)\partial_\tau u + f_u(u)v - c^\varepsilon v \end{pmatrix} \tag{3.20}$$

near the orbit $Q^0 = (q^0, q_x^0)$ for (ε, Ω) close to zero. We employ the smooth coordinate change

$$z = x\sqrt{\omega_0 + \Omega}, \qquad (\widetilde{u}, \widetilde{v}) = \left(u, v/\sqrt{\omega_0 + \Omega}\right)$$

which transforms (3.20) into the equation

$$\begin{pmatrix} \widetilde{u}_z \\ \widetilde{v}_z \end{pmatrix} = \begin{pmatrix} \widetilde{v} \\ \partial_\tau \widetilde{u} + (\omega_0 + \Omega)^{-1/2}[f_u(\widetilde{u}) - c^\varepsilon]\widetilde{v} \end{pmatrix}. \tag{3.21}$$

The advantage of (3.21) over (3.20) is that the Ω-dependent part of the right-hand side of (3.21) is a smooth mapping from Y into itself which depends smoothly on (ε, Ω) for (ε, Ω) near zero; in contrast, the dependence on Ω of the right-hand side of (3.20) is through the term $u \mapsto \Omega\partial_\tau u$ which is not even bounded from H^1 into $H^{1/2}$. Using the fact that the linearized equation (3.10) can be solved using exponential dichotomies (whose existence we established above), we can proceed as in [13, section 3.5] and [22] to prove the existence of unstable and center-stable manifolds for (3.21), and therefore for (3.20), near the viscous shock. More precisely, there exist constants $\delta > 0$ and $K > 0$ such that

$$\begin{aligned} \mathcal{W}^{\mathrm{u}}_{\varepsilon,\Omega} := \{&U_0 \in Y;\ \exists \text{ solution } U(x) \text{ of (3.20) on } \mathbb{R}^- :\ U(0) = U_0,\ |U_0 - Q^0(0)| < \delta, \\ &|U(x) - U_{\mathrm{eq}}^-(\varepsilon)| \le K\mathrm{e}^{-\nu_-^{\mathrm{u}}|x|} \text{ for } x \le 0\}, \\ \mathcal{W}^{\mathrm{cs}}_{\varepsilon,\Omega} := \{&U_0 \in Y;\ \exists \text{ solution } U(x) \text{ of (3.20) on } \mathbb{R}^+ :\ U(0) = U_0,\ |U_0 - Q^0(0)| < \delta, \\ &\exists U_{\mathrm{eq}}^+ \in Y_0 \text{ with } |U_{\mathrm{eq}}^+ - U_{\mathrm{eq}}^+(0)| < \delta \text{ so that } |U(x) - U_{\mathrm{eq}}^+| \le K\mathrm{e}^{-\nu_+^{\mathrm{s}}|x|} \\ &\text{for } x \ge 0\} \end{aligned}$$

are smooth manifolds that are invariant under the action of the group Γ defined in (3.4) and that depend smoothly on (ε, Ω) near zero (smoothness with respect to the parameters follows from [22] since the right-hand side of the rescaled equation (3.21) is smooth in the parameters). Moreover, $Q^\varepsilon(0) \in \mathcal{W}^{\mathrm{u}}_{\varepsilon,\Omega} \cap \mathcal{W}^{\mathrm{cs}}_{\varepsilon,\Omega}$, and the tangent spaces of the invariant manifolds at this point of intersection are given by

$$T_{Q^0(0)}\mathcal{W}^{\mathrm{u}}_{0,0} = E^{\mathrm{u}}_{-}, \qquad T_{Q^0(0)}\mathcal{W}^{\mathrm{cs}}_{0,0} = E^{\mathrm{cs}}_{+}.$$

Note that the center-stable manifold $\mathcal{W}^{\mathrm{cs}}_{\varepsilon,\Omega}$ is in effect given as the union of stable manifolds to the manifold $\{U^+_{\mathrm{eq}} = (u, 0);\ u \in \mathbb{R}^n\}$ of asymptotic states, and therefore unique.

Finding solutions of (2.2), with temporal frequency ω near ω_0, that converge asymptotically to constants as $x \to \pm\infty$ is therefore equivalent to finding elements U_0 in the intersection

$$\mathcal{W}^{\mathrm{u}}_{\varepsilon,\Omega} \cap \mathcal{W}^{\mathrm{cs}}_{\varepsilon,\Omega} \cap \left[Q^0(0) + \mathcal{S}\right] \tag{3.22}$$

for Ω close to zero, with $\mathcal{S}$ as in (3.19). Note that U_0 will have nontrivial time-τ dependence if and only if U_0 has a nonzero Y_{h}-component. The minimal period will be $2\pi/\omega$ if the component of U_0 in Y_1 does not vanish. We use Lyapunov–Schmidt reduction to determine the intersection (3.22). To this end, we write

$$E^{\mathrm{cs}}_{+} \cap \mathcal{S} = \widetilde{E}^{\mathrm{cs}}_{+} \oplus \mathrm{span}\{V_1(0), V_2(0)\}, \qquad E^{\mathrm{u}}_{-} \cap \mathcal{S} = \widetilde{E}^{\mathrm{u}}_{-} \oplus \mathrm{span}\{V_1(0), V_2(0)\}.$$

There are then unique smooth maps

$$\begin{aligned} G^{\mathrm{cs}}(\cdot;\varepsilon,\Omega):&\quad \widetilde{E}^{\mathrm{cs}}_{+} \oplus \mathrm{span}\{V_1(0), V_2(0)\} \longrightarrow \widetilde{E}^{\mathrm{u}}_{-} \oplus \mathrm{span}\{\Psi_1(0), \Psi_2(0)\},\\ G^{\mathrm{u}}(\cdot;\varepsilon,\Omega):&\quad \widetilde{E}^{\mathrm{u}}_{-} \oplus \mathrm{span}\{V_1(0), V_2(0)\} \longrightarrow \widetilde{E}^{\mathrm{cs}}_{+} \oplus \mathrm{span}\{\Psi_1(0), \Psi_2(0)\} \end{aligned}$$

with

$$Q^\varepsilon(0) + \mathrm{graph}\, G^j(\cdot;\varepsilon,\Omega) = \mathcal{W}^j_{\varepsilon,\Omega} \cap [Q^0(0) + \mathcal{S}], \qquad j = \mathrm{cs}, \mathrm{u},$$

and $D_U G^j(0;0,0) = 0$ for $j = \mathrm{cs}, \mathrm{u}$. In particular, both maps are equivariant under the S^1-action Γ. Thus, intersections of $\mathcal{W}^{\mathrm{u}}_{\varepsilon,\Omega}$ and $\mathcal{W}^{\mathrm{cs}}_{\varepsilon,\Omega}$ in $Q^0(0) + \mathcal{S}$ are in one-to-one correspondence with the zeroes of the mapping

$$\begin{aligned} G(\cdot;\varepsilon,\Omega):&\quad \mathbb{R} \times \widetilde{E}^{\mathrm{u}}_{-} \times \widetilde{E}^{\mathrm{cs}}_{+} \longrightarrow \widetilde{E}^{\mathrm{u}}_{-} \oplus \widetilde{E}^{\mathrm{cs}}_{+} \oplus \mathrm{span}\{\Psi_1(0), \Psi_2(0)\},\\ &(a, w^{\mathrm{u}}, w^{\mathrm{cs}}) \longmapsto w^{\mathrm{u}} + G^{\mathrm{u}}(w^{\mathrm{u}} + aV_1(0);\varepsilon,\Omega) - [w^{\mathrm{cs}} + G^{\mathrm{cs}}(w^{\mathrm{cs}} + aV_1(0);\varepsilon,\Omega)], \end{aligned}$$

where we factored out the nontrivial S^1-action on $\mathrm{span}\{V_1(0), V_2(0)\}$. Lyapunov–Schmidt reduction shows that there is a unique map

$$W: \quad U_\delta(0) \subset \mathbb{R}^3 \longrightarrow \widetilde{E}^{\mathrm{u}}_{-} \times \widetilde{E}^{\mathrm{cs}}_{+}, \quad (a,\varepsilon,\Omega) \longmapsto (W^{\mathrm{u}}(a,\varepsilon,\Omega), W^{\mathrm{cs}}(a,\varepsilon,\Omega)),$$

so that $G(a, w^{\mathrm{u}}, w^{\mathrm{cs}};\varepsilon,\Omega) = 0$ if and only if

$$\langle \Psi_j(0), G(a, W(a,\varepsilon,\Omega);\varepsilon,\Omega)\rangle_X = 0 \qquad \text{for } j = 1, 2.$$

Furthermore, W is smooth in (a,ε,Ω) and we have $D_{(a,\varepsilon,\Omega)}W(0,0,0) = 0$. In fact, since $G(0,0,0;\varepsilon,\Omega) \equiv 0$ due to $Q^\varepsilon(0) \in \mathcal{W}^{\mathrm{u}}_{\varepsilon,\Omega} \cap \mathcal{W}^{\mathrm{cs}}_{\varepsilon,\Omega}$ for all (ε,Ω), we have in addition that $W(0,\varepsilon,\Omega) = 0$ for all small (ε,Ω), so that

$$W(a,\varepsilon,\Omega) = a\mathrm{O}(|a| + |\varepsilon| + |\Omega|). \tag{3.23}$$

It suffices therefore to solve the reduced equations

$$\langle \Psi_j(0), G(a, W(a,\varepsilon,\Omega);\varepsilon,\Omega)\rangle_X = 0 \qquad \text{for } j = 1,2. \tag{3.24}$$

To derive an expression for (3.24), we write (3.20) as

$$U_x = F(U,\varepsilon,\Omega). \tag{3.25}$$

Using the coordinates $U = Q^\varepsilon + \widetilde{U}$, we find that $\widetilde{U}$ satisfies

$$\widetilde{U}_x = F_U(Q^0,0,0)\widetilde{U} + \mathcal{N}(\widetilde{U},\varepsilon,\Omega,x), \tag{3.26}$$

where

$$\begin{aligned} \mathcal{N}(\widetilde{U},\varepsilon,\Omega,x) &:= F(Q^\varepsilon + \widetilde{U},\varepsilon,\Omega) - F(Q^\varepsilon,\varepsilon,\Omega) - F_U(Q^0,0,0)\widetilde{U} \\ &= \mathrm{O}(|\widetilde{U}|(|\widetilde{U}| + |\varepsilon| + |\Omega|)). \end{aligned} \tag{3.27}$$

Using the variation-of-constant formula that captures unstable and center-stable manifolds (see, e.g., [22] or [13, Proposition 3.13]) and the fact that $\Psi_j(x)$ satisfies (3.18) together with [14, Lemma 5.1], we find that (3.24) is given by

$$\int_{-\infty}^{\infty} \left\langle \Psi_j(x), \mathcal{N}(\widetilde{U}^\pm(x),\varepsilon,\Omega,x)\right\rangle_X \mathrm{d}x = 0, \qquad j = 1,2, \tag{3.28}$$

where $\widetilde{U}^\pm(x)$ satisfies (3.26) on $\mathbb{R}^\pm$ with $\widetilde{U}^-(0) = aV_1(0) + W^{\mathrm{u}}(a,\varepsilon,\Omega)$ and $\widetilde{U}^+(0) = aV_1(0) + W^{\mathrm{cs}}(a,\varepsilon,\Omega)$. If we write (3.28) as $\Pi(a,\varepsilon,\Omega) = 0$, then we know from the preceding discussion that $\Pi(0,\varepsilon,\Omega) = 0$ for all (ε,Ω): this solution corresponds to the persisting Lax shocks in $\operatorname{Fix}\Gamma$. To obtain genuinely time-periodic solutions corresponding to $a \neq 0$, we write

$$\Pi(a,\varepsilon,\Omega) = a\widetilde{\Pi}(a,\varepsilon,\Omega) \tag{3.29}$$

and consider $\widetilde{\Pi}(a,\varepsilon,\Omega) = 0$, which can be solved by the implicit function theorem provided the 2×2 matrix $D_{(\varepsilon,\Omega)}\widetilde{\Pi}(0,0,0)$ is invertible. Equations (3.29) and (3.28) show that

$$\begin{aligned} D_{(\varepsilon,\Omega)}\widetilde{\Pi}(0,0,0) &= D_a D_{(\varepsilon,\Omega)}\Pi(0,0,0) \\ &= \left[D_a D_{(\varepsilon,\Omega)} \int_{-\infty}^{\infty} \langle \Psi_j(x), \mathcal{N}(\widetilde{U}^\pm(x),\varepsilon,\Omega,x)\rangle_X \,\mathrm{d}x \Big|_{(a,\varepsilon,\Omega)=0} \right]_{j=1,2} \end{aligned} \tag{3.30}$$

which we now compute. We know that

$$\widetilde{U}^+(x) = aV_1(x) + W^{\mathrm{cs}}(a,\varepsilon,\Omega)(x), \qquad \widetilde{U}^-(x) = aV_1(x) + W^{\mathrm{u}}(a,\varepsilon,\Omega)(x)$$

which we rewrite as

$$\widetilde{U}^\pm(x) = a[V_1(x) + \widetilde{W}^\pm(x;a,\varepsilon,\Omega)] \tag{3.31}$$

with

$$\begin{aligned} \widetilde{W}^-(x;a,\varepsilon,\Omega) &:= \frac{1}{a} W^{\mathrm{u}}(a,\varepsilon,\Omega)(x) = \mathrm{O}(|a| + |\varepsilon| + |\Omega|), \qquad x \in \mathbb{R}^-, \\ \widetilde{W}^+(x;a,\varepsilon,\Omega) &:= \frac{1}{a} W^{\mathrm{cs}}(a,\varepsilon,\Omega)(x) = \mathrm{O}(|a| + |\varepsilon| + |\Omega|), \qquad x \in \mathbb{R}^+, \end{aligned} \tag{3.32}$$

due to the estimate (3.23). Thus,

$$\frac{\mathrm{d}^2}{\mathrm{d}(\varepsilon,\Omega)\mathrm{d}a}\mathcal{N}(\widetilde{U}^{\pm}(x),\varepsilon,\Omega,x)|_{(a,\varepsilon,\Omega)=0} = D_{(\varepsilon,\Omega)}F_U(Q^{\varepsilon}(x),\varepsilon,\Omega)|_{(a,\varepsilon,\Omega)=0}V_1(x).$$

Upon comparing (3.25) with (3.20), we see that

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\Omega\mathrm{d}a}\mathcal{N}|_{(a,\varepsilon,\Omega)=0} &= \begin{pmatrix} 0 & 0 \\ \partial_\tau & 0 \end{pmatrix} V_1,\\ \frac{\mathrm{d}^2}{\mathrm{d}\varepsilon\mathrm{d}a}\mathcal{N}|_{(a,\varepsilon,\Omega)=0} &= \begin{pmatrix} 0 & 0 \\ \partial_\varepsilon(f_{uu}(q^\varepsilon)[q_x^\varepsilon,\cdot])|_{\varepsilon=0} & \partial_\varepsilon[f_u(q^\varepsilon)-c^\varepsilon]|_{\varepsilon=0} \end{pmatrix} V_1.\end{aligned}$$

Substituting the expressions (3.16) and (3.17) for V_1 and Ψ_j and using the normalization (3.1), we obtain

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}\Omega\mathrm{d}a}\int_{-\infty}^{\infty}\left\langle \Psi_j(x),\mathcal{N}(\widetilde{U}^{\pm}(x),\varepsilon,\Omega,x)\right\rangle_X \mathrm{d}x\Big|_{(\varepsilon,\Omega)=0}\right]_{j=1,2} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \in \mathbb{R}^2. \tag{3.33}$$

An analogous computation for the derivative with respect to ε gives

$$\begin{aligned}&\left[\frac{\mathrm{d}^2}{\mathrm{d}\varepsilon\mathrm{d}a}\int_{-\infty}^{\infty}\left\langle \Psi_1(x),\mathcal{N}(\widetilde{U}^{\pm}(x),\varepsilon,\Omega,x)\right\rangle_X \mathrm{d}x\Big|_{(\varepsilon,\Omega)=0}\right]\\ &= \frac{1}{2}\sum_{j=1}^{2}\left\langle \psi_j, \partial_\varepsilon\partial_x\left[f_u(q^\varepsilon)v_j - c^\varepsilon v_j\right]_{\varepsilon=0}\right\rangle_{L^2}\\ &= -\frac{1}{2}\sum_{j=1}^{2}\langle \psi_j, \partial_\varepsilon\mathcal{L}^\varepsilon|_{\varepsilon=0}v_j\rangle_{L^2}\\ &= -\operatorname{Re}\lambda_\varepsilon(0),\end{aligned}$$

where we used (3.2) to obtain the last step. Hence, we find that the Jacobian in (3.30) is given by

$$D_{(\varepsilon,\Omega)}\widetilde{\Pi}(0,0,0) = \begin{pmatrix} -\operatorname{Re}\lambda_\varepsilon(0) & 0 \\ \star & 1 \end{pmatrix} \tag{3.34}$$

which is invertible due to our hypothesis on the transverse crossing of the Hopf eigenvalues.

Upon applying the implicit function theorem to solve $\widetilde{\Pi}(a,\varepsilon,\Omega)=0$, we conclude that there exist unique functions $(\varepsilon_*,\Omega_*)(a)\in\mathbb{R}^2$ and $Q^*(0;a)\in Y$, defined for $|a|<\delta$, so that

$$Q^*(0;a)\in\mathcal{W}^{\mathrm{u}}_{(\varepsilon_*,\Omega_*)(a)}\cap\mathcal{W}^{\mathrm{cs}}_{(\varepsilon_*,\Omega_*)(a)}\cap\left[Q^0(0)+\mathcal{S}\right].$$

These functions are smooth and satisfy $\partial_a(\varepsilon_*,\Omega_*)|_{a=0}=0$, $Q^*(0;0)=Q^0(0)$, and

$$\begin{aligned}Q^*(x;a) &= Q^{\varepsilon_*(a)}(x) + a[V_1(x)+\widetilde{W}^{\pm}(x;a,\epsilon_*(a),\Omega_*(a))]\\ &=: Q^{\varepsilon_*(a)}(x) + a\widetilde{Q}^*(x;a).\end{aligned} \tag{3.35}$$

By construction, we have $Q^*(x;a)\to U^-_{\mathrm{eq}}(\varepsilon_*(a))$ as $x\to-\infty$. Furthermore, we have $Q^*(x;a)\in\mathcal{W}^{\mathrm{cs}}_{(\varepsilon_*,\Omega_*)(a)}$ from which we infer that there exists a $U^+_*(a)\in\mathbb{R}^n$ with

$|U_*^+(a) - U_{\rm eq}^+(\varepsilon_*(a))| < \delta$, so that $Q^*(x;a) \to U_*^+(a)$ exponentially as $x \to \infty$ with rate $\nu_+^{\rm s}$. We claim that $U_*^+(a) = U_{\rm eq}^+(\varepsilon_*(a))$. To prove this claim, consider the smooth functional

$$\mathcal{E}: \quad Y \longrightarrow \mathbb{R}^n, \quad (u,v) \longmapsto \int_0^{2\pi} [v - f(u) + cu]\,\mathrm{d}\tau. \tag{3.36}$$

This functional is conserved under the evolution of (3.3). If $U(x) = (u,v)(x) \in Y$ is a solution of (3.3), then $v = u_x$ and

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\mathcal{E}(U(x)) &= \frac{\mathrm{d}}{\mathrm{d}x}\int_0^{2\pi} [v - f(u) + cu]\,\mathrm{d}\tau \\ &= \int_0^{2\pi} [v_x - f_u(u)v + cv]\,\mathrm{d}\tau \\ &\overset{(3.3)}{=} \int_0^{2\pi} \omega u_\tau\,\mathrm{d}\tau = 0 \end{aligned} \tag{3.37}$$

since u is 2π-periodic in τ. Furthermore, for $U = (u_0, v_0) \in \operatorname{Fix}\Gamma \subset Y$, we have

$$D_U\mathcal{E}(u_0,v_0)\begin{pmatrix}\widetilde{u}\\ \widetilde{v}\end{pmatrix} = \int_0^{2\pi} [\widetilde{v} - f_u(u_0)\widetilde{u} + c\widetilde{u}]\,\mathrm{d}\tau = \widetilde{v}_0 - f_u(u_0)\widetilde{u}_0 + c\widetilde{u}_0 \in \mathbb{R}^n.$$

In particular, $D_U\mathcal{E}(u_\pm^0, 0)$ restricted to $\mathbb{R}^n \times \{0\} \subset Y_0$ is invertible, since we assumed that none of the characteristic speeds vanishes in the frame that moves with speed c^0. Since $Q^{\varepsilon_*(a)}(x)$ connects $U_{\rm eq}^-(\varepsilon_*(a))$ to $U_{\rm eq}^+(\varepsilon_*(a))$, they have the same $\mathcal{E}$-values, and the preceding argument shows that there is no other equilibrium $U_{\rm eq}^+$ near $U_{\rm eq}^+(\varepsilon_*(a))$ with the $\mathcal{E}$-value of $U_{\rm eq}^+(\varepsilon_*(a))$. Therefore, $Q^*(x;a) \to U_{\rm eq}^\pm(\varepsilon_*(a))$ for $x \to \pm\infty$.

This completes the existence proof of the bifurcating oscillatory viscous shock waves. The uniqueness statement in Theorem 2.1 is a consequence of our construction which captures all solutions that lie in the intersection of $\mathcal{W}_{\varepsilon,\Omega}^{\rm u}$ and $\mathcal{W}_{\varepsilon,\Omega}^{\rm cs}$. We remark that (3.37) also shows that any time-periodic localized travelling viscous shock wave satisfies the Rankine–Hugoniot condition (2.4).

To prepare the ground for the following spectral stability proof, we derive an expression for $\varepsilon_{aa}(0)$. First, we set $(\varepsilon, \Omega) = 0$ and compute the derivatives

$$\frac{\mathrm{d}^j}{\mathrm{d}a^j}\Pi_i(a,0,0) = \frac{\mathrm{d}^j}{\mathrm{d}a^j}\int_{-\infty}^{\infty} \left\langle \Psi_i(x), \mathcal{N}(\widetilde{U}^\pm(x),0,0,x)\right\rangle_X \mathrm{d}x$$

at $a = 0$ for $i = 1,2$. Using the expressions (3.27) for $\mathcal{N}$ and (3.31) for $\widetilde{U}^\pm$ together with the estimate (3.32) for $\widetilde{W}(x;a,0,0)$, we easily find that the first and second derivatives vanish at $a = 0$ for $i = 1,2$, while

$$\begin{aligned} \kappa_3 :=& \frac{\mathrm{d}^3}{\mathrm{d}a^3}\Pi_1(a,0,0) \\ =& \int_{-\infty}^{\infty} \Big\langle \Psi_1(x), F_{UUU}(Q^0(x),0,0)[V_1(x)]^3 + \\ & \qquad + 3F_{UU}(Q^0(x),0,0)[V_1(x), \widetilde{W}_a(x;0,0,0)]\Big\rangle_X \mathrm{d}x. \end{aligned} \tag{3.38}$$

A straightforward calculation using (3.34) then shows that

$$\varepsilon_{aa}(0) = \frac{\kappa_3}{3\operatorname{Re}\lambda_\varepsilon(0)}. \tag{3.39}$$

4. Stability of the bifurcating modulated viscous shocks. This section is devoted to the proof of Theorem 2.2. Our goal is to determine the Floquet spectrum

$$\Sigma = \{\lambda \in \mathbb{C};\ \mathrm{e}^{2\pi\lambda} \in \text{ spectrum of } \Phi_{2\pi}\}$$

associated with the evolution Φ_t of the linearization

$$\omega v_\tau = \partial_x[\partial_x + c - f_u(q^*(x,\tau;a))]v$$

of (2.2) about q^* on $\mathcal{C}^0_{\mathrm{unif}}$. Note that (ε,ω) and q^* depend smoothly on the parameter a introduced in section 3, and so do the wave speed $c = c^\varepsilon$ and the asymptotic rest states $u^\varepsilon_\pm$ through $\varepsilon = \varepsilon_*(a)$; we will suppress this dependence for most of the proof.

The Floquet spectrum Σ is the disjoint union of the essential spectrum Σ_{ess} and the point spectrum Σ_{pt}, which consists, by definition, of all isolated eigenvalues with finite multiplicity. Since the modulated shock $q^*(x,\tau;a)$ converges exponentially to the constants $u^\varepsilon_\pm$ as $x \to \pm\infty$, uniformly in τ, the set Σ_{ess} is bounded to the right by the essential spectra

$$\Sigma^\pm_{\mathrm{ess}} = \{\lambda \in \mathbb{C};\ \det(k^2 + \mathrm{i}k[f_u(u^\varepsilon_\pm) - c^\varepsilon] + \lambda) = 0 \text{ for some } k \in \mathbb{R}\}$$

of $u^\varepsilon_\pm$ (see, for instance, [14, Proposition 2.10]), which touch the imaginary axis at $\lambda = 0$ and lie otherwise in the open left half-plane due to our hypothesis that the eigenvalues of $f_u(u^\varepsilon_\pm)$ are real and simple.

It therefore suffices to locate point spectrum, that is, isolated Floquet exponents λ which are captured, via the ansatz $v(x,\tau) = \mathrm{e}^{\lambda\tau}u(x,\tau)$ with $u(x,\tau+2\pi) = u(x,\tau)$ for all τ, by the equation

$$\omega_*(a)u_\tau + \lambda u = \partial_x[\partial_x + c^{\varepsilon_*(a)} - f_u(q^*(x,\tau;a))]u$$

which we rewrite as

$$\begin{aligned} V_x &= \begin{pmatrix} 0 & 1 \\ \omega\partial_\tau + \lambda + f_{uu}(q^*)[q^*_x,\cdot] & f_u(q^*) - c \end{pmatrix} V \\ &= [F_U(Q^*(x;a), \epsilon_*(a), \Omega_*(a)) + \lambda\mathcal{B}]V, \qquad V \in Y, \end{aligned} \tag{4.1}$$

with $Q^*(x;a) = (q^*, q^*_x)(x,\cdot;a)$ from (3.35).

Since we assumed spectral stability for $\epsilon = 0$ except for the Hopf eigenvalues and the translational eigenvalue at the origin (which all contribute to the Floquet exponent $\lambda = 0$), it suffices to find all isolated Floquet exponents of (4.1) in a fixed small neighborhood of the origin. We choose an open set $\Omega \subset \mathbb{C}$ as indicated in Figure 2. Standard theory implies that $\lambda \in \Omega$ is a Floquet exponent if and only if (4.1) has a nontrivial exponentially decaying solution on $\mathbb{R}$. Taking the limit $x \to \pm\infty$ in (4.1), we obtain the asymptotic operators

$$\begin{pmatrix} 0 & 1 \\ \omega\partial_\tau + \lambda & f_u(u_\pm) - c \end{pmatrix}. \tag{4.2}$$

We denote the eigenvectors and eigenvalues of $[f_u(u_\pm) - c]$ by $r^\pm_j$ and $\nu^\pm_j$, respectively. As discussed in section 3, the operators in (4.2) are hyperbolic for $\lambda = 0$ except for the n-fold eigenvalue $\mathcal{V} = 0$ with eigenvectors $(r^\pm_j, 0) \in Y_0$. This eigenvalue and the associated eigenvectors become

$$\mathcal{V}^\pm_j = -\frac{\lambda}{\nu^\pm_j} + \mathrm{O}(\lambda^2), \qquad R^\pm_j = \begin{pmatrix} r^\pm_j \\ \mathcal{V}^\pm_j r^\pm_j \end{pmatrix}, \qquad j = 1,\dots,n, \tag{4.3}$$

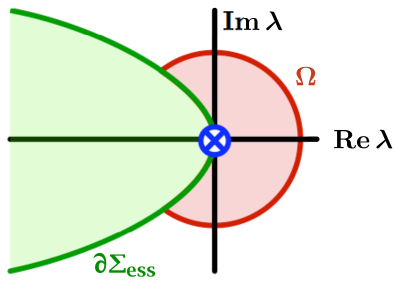


FIG. 2. *The definition of the open set $\Omega \subset \mathbb{C}$ in the complex Floquet plane is shown. The embedded Floquet exponent at the origin has multiplicity at least equal to two with eigenfunctions q_x^* and q_τ^*.*

for λ near zero. For $\lambda \in \Omega$, the unstable eigenspace $E_-^\infty(\lambda, a)$ at $x = -\infty$ and the stable eigenspace $E_+^\infty(\lambda, a)$ at $x = \infty$ are therefore given by

$$E_+^\infty := E_+^{\mathrm{s}} \oplus \mathcal{R}_+, \qquad E_-^\infty := E_-^{\mathrm{u}} \oplus \mathcal{R}_-, \qquad \mathcal{R}_\pm = \mathrm{span}\{R_j^\pm;\ \nu_j^\pm \gtrless 0\},$$

and these spaces depend smoothly on a and are analytic in λ for λ near zero. Note that $\dim \mathcal{R}^+ = p$ and $\dim \mathcal{R}^- = n - p - 1$ with p as in section 2.

LEMMA 4.1. *There are unique closed subspaces $E_\pm(\lambda, a)$ of Y, defined and analytic in λ near zero and smooth in $a \geq 0$, such that $V(x)$ is a bounded solution of (4.1) on $\mathbb{R}^\pm$ for some $\lambda \in \Omega$ if and only if $V(0) \in E_\pm(\lambda, a)$.*

Proof. We begin by considering (4.1) with Q^* replaced by Q^ε. In this case, (4.1) decouples on each Fourier space Y_k, and the claimed statement holds for Q^ε due to the Gap lemma [5, 9] applied in Y_0 and exponential dichotomy theory together with estimates as in [13, Lemma 3.3] in the other Fourier spaces. Since the difference of Q^ε and Q^* is small for all x and decays to zero exponentially as $|x| \to \infty$, these results carry over to (4.1) using, for instance, the integral formulation in [12, (4.12) in section 4.3]; see also [18, section 7.6] for a slightly different proof. ∎

Lemma 4.1 shows that Floquet exponents in Ω can be found by seeking nontrivial intersections of $E_-(\lambda, a)$ and $E_+(\lambda, a)$. To determine their intersections, we first set $(\lambda, a) = 0$ to see what these spaces look like at onset and then use perturbation theory to analyze the case when $(\lambda, a) \neq 0$.

Hence, let $\lambda = 0$, then (4.1) is simply the variational equation

$$V_x = F_U(Q^*(x; a), \epsilon_*(a), \Omega_*(a))V, \qquad V \in Y, \tag{4.4}$$

of the modulated wave Q^*. When $a = 0$, we have $Q^* = Q^0$, and (4.4) describes Floquet exponents at $\lambda = 0$ of the unperturbed viscous shock q^0. In particular, (4.4) decouples on each Fourier space Y_k, and our hypotheses on the Evans function and the spectral properties of the viscous shock imply that

$$\begin{aligned} E_+(0,0) \cap E_-(0,0) &= \mathrm{span}\{Q_x^0(0), V_1(0), V_2(0)\}, \\ [E_+(0,0) + E_-(0,0)]^\perp &= \mathrm{span}\{\Psi_0, \Psi_1(0), \Psi_2(0)\} \end{aligned} \tag{4.5}$$

for an appropriate nonzero vector $\Psi_0 \in Y_0$. Next, consider (4.4) for an arbitrary a near zero. First, note that the gradient of the jth component $\mathcal{E}_j$ of the conserved quantity $\mathcal{E}$ from (3.36), computed in the $X = L^2(S^1) \times L^2(S^1)$ scalar product, is given by

$$\nabla \mathcal{E}_j(u, v) = \nabla \langle \mathcal{E}(u, v), e_j \rangle_{\mathbb{R}^n} = \begin{pmatrix} -[f_u^T(u) - c]e_j \\ e_j \end{pmatrix}, \qquad j = 1, \dots, n, \tag{4.6}$$

where e_j denotes the jth canonical basis vector in $\mathbb{R}^n$. Our analysis of $\mathcal{E}$ in section 3 implies that these n gradients are linearly independent for $a = 0$, and we therefore have $\dim E^*(a) = n$ for all small a where

$$E^*(a) := \mathrm{span}\,\{\nabla\langle\mathcal{E}(Q^*(0;a)), e_j\rangle_{\mathbb{R}^n};\ j = 1, \ldots, n\}\,.$$

The gradients in (4.6) also satisfy the adjoint equation of (4.4), again computed in X, which shows that

$$\frac{\mathrm{d}}{\mathrm{d}x}\langle\nabla\langle\mathcal{E}(Q^*(x;a)), e_j\rangle_{\mathbb{R}^n}, V(x)\rangle_X \equiv 0, \qquad j = 1, \ldots, n, \tag{4.7}$$

for each solution $V(x)$ of (4.4); this can also be verified directly by evaluating (4.7). We denote by $\ell_j^{\pm}(a)$ the smooth eigenvectors of $[f_u^T(u_{\pm}^{\varepsilon}) - c^{\varepsilon}]$ at $\varepsilon_*(a)$ associated with the eigenvalues $\nu_j^{\pm}$ and define

$$E_{\pm}^*(a) = \mathrm{span}\left\{\nabla\langle\mathcal{E}(Q^*(0;a)), \ell_j^{\pm}(a)\rangle_{\mathbb{R}^n};\ \nu_j^{\pm} \lessgtr 0\right\} \subset E^*(a). \tag{4.8}$$

Set $a = 0$, then $\dim E_+^*(0) = n - p$ and $\dim E_-^*(0) = p + 1$. Equations (4.3), (4.6), and (4.7) imply that $E_{\pm}^*(0) \perp_X E_{\pm}(0,0)$ and, in fact, that $E_{\pm}^*(0)$ are perpendicular to each solution of (4.4) at $a = 0$ that decays exponentially at $x = -\infty$ or $x = \infty$. Equation (4.5) implies then that $E_+^*(0) \cap E_-^*(0) = \mathrm{span}\{\Psi_0\}$, and therefore $\dim[E_+^*(0) + E_-^*(0)] = n$. Since $E_{\pm}^*(a) \subset E^*(a)$ for all a, and the latter space is n-dimensional for all a, we conclude that $E_+^*(a) + E_-^*(a) = E^*(a)$, and the dimensions of the sum and intersection of $E_{\pm}^*(a)$ cannot change for a close to zero. Hence, we can choose a nonzero basis vector $\Psi_0^*(0;a)$ in the one-dimensional intersection $E_+^*(a) \cap E_-^*(a)$ that depends smoothly on a as well as linearly independent smooth elements $\Psi_j^{\pm}(0;a) \in E_{\pm}^*(a)$, with $j = 1, \ldots, n-p-1$ for the $+$ sign and $j = 1, \ldots, p$ for the $-$ sign, so that $\Psi_j^{\pm}(0;a) \perp_X \Psi_0^*(0;a)$ for all j. Using (4.3), (4.6), and (4.8), we see that

$$\Psi_j^{\pm}(0;a) \perp_X E_{\pm}(0,a), \qquad \Psi_0^*(0;a) \perp_X [E_+(0,a) + E_-(0,a)] \tag{4.9}$$

for all a. Lastly, we define

$$\Psi_{1,2}^*(0;a) := [1 - P(a)]\Psi_{1,2}(0;a), \tag{4.10}$$

where $P(a)$ is the orthogonal projection in X onto $E_+^*(a) + E_-^*(a)$.

Having prepared the ground for the forthcoming analysis, we now return to the full eigenvalue problem (4.1)

$$V_x = [F_U(Q^*(x;a), \epsilon_*(a), \Omega_*(a)) + \lambda\mathcal{B}]V.$$

We seek solutions $V^{\pm}(x)$ on $\mathbb{R}^{\pm}$ of the form

$$V^{\pm}(x) = b_0 Q_x^*(x;a) + b_1 V_1(x) + b_2 \widetilde{Q}_\tau^*(x;a) + \widetilde{V}^{\pm}(x;\lambda,a)b \tag{4.11}$$

with $b = (b_0, b_1, b_2)$, $Q^* = Q^{\varepsilon} + a\widetilde{Q}^*$ as in (3.35), and

$$\widetilde{V}^{\pm}(0;\lambda,a)b \perp \mathrm{span}\{Q_x^*(0;a), V_1(0), \widetilde{Q}_\tau^*(0;a)\} \tag{4.12}$$

for all b, so that

$$V^+(0) - V^-(0) \in \mathrm{span}\{\Psi_j^*(0;a);\ j = 0, 1, 2\}, \tag{4.13}$$

$$\mathrm{dist}\left(\frac{1}{|V^{\pm}(x)|_X}V^{\pm}(x), E_{\pm}^{\infty}(\lambda,a)\right) \to 0 \text{ as } x \to \pm\infty. \tag{4.14}$$

Using exponential dichotomies as in section 3, we can then easily show that the system (4.1), (4.11)–(4.14) has unique solutions for each $b \in \mathbb{R}^3$ and (λ, a) near zero and that these solutions depend analytically on λ and smoothly on a. In particular, $E_+(\lambda, a) \cap E_-(\lambda, a) \neq \{0\}$ if and only if $\det \mathcal{D}(\lambda, a) = 0$ where

$$\mathcal{D}(\lambda, a): \quad \mathbb{R}^3 \longrightarrow \mathbb{R}^3, \quad b \longmapsto \mathcal{D}(\lambda, a)b = \big(\langle \Psi_j^*(0; a), V^+(0) - V^-(0)\rangle_X\big)_{j=0,1,2}.$$

We will now compute the Taylor expansion of $\mathcal{D}$ to solve $\det \mathcal{D}(\lambda, a) = 0$.

First, we set $a = 0$ so that $Q^* = Q^0$, $Q_x^0 =: V_0$, and $\widetilde{Q}_\tau^* = V_2$. A calculation similar to the derivation of (3.28) gives

$$\begin{aligned} \partial_\lambda \mathcal{D}(0,0) &= \left(\int_{\mathbb{R}} \langle \Psi_i^*(x; 0), \mathcal{B}V_j(x)\rangle_X \, \mathrm{d}x \right)_{i,j=0,1,2} \\ &= \operatorname{diag}\left(\int_{\mathbb{R}} \langle \Psi_j(x), \mathcal{B}V_j(x)\rangle_X \, \mathrm{d}x \right) \\ &= \operatorname{diag}(M_0, 1, 1), \end{aligned}$$

where we used the normalization (3.1). Our hypothesis that $\lambda = 0$ is a simple zero of the Evans function of the viscous shock at $\varepsilon = 0$ implies that $M_0 \neq 0$.

Next, we set $\lambda = 0$ and compute derivatives with respect to a. Since $\lambda = 0$, the eigenvalue problem reduces to the variational equation (4.4). In particular, both $\partial_x Q^*(x; a)$ and $\partial_\tau \widetilde{Q}^*(x; a)$ are solutions of (4.4) that satisfy (4.14), and we can set $b_0 = b_2 = 0$ as they make no contribution to $\mathcal{D}(0, a)$. We focus therefore on $V^\pm(x) = V_1(x) + \widetilde{V}^\pm(x; 0, a)$ for which (4.9) and (4.14) together imply

$$\langle \Psi_0^*(0; a), V^+(0) - V^-(0)\rangle_X = 0 \tag{4.15}$$

for all a. The equation for $\widetilde{V}$ is

$$\begin{aligned} \widetilde{V}_x &= F_U(Q^0(x), 0, 0)\widetilde{V} \\ &\quad + [F_U(Q^*(x; a), \varepsilon_*(a), \Omega_*(a)) - F_U(Q^0(x), 0, 0)](V_1(x) + \widetilde{V}) \end{aligned} \tag{4.16}$$

and, proceeding as before and using (3.35), we obtain

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}a} \left\langle \Psi_j^*(0; a), \widetilde{V}^+(0; 0, a) - \widetilde{V}^-(0; 0, a) \right\rangle_X \Big|_{a=0} \\ &= \int_{\mathbb{R}} \left\langle \Psi_j(x), F_{UU}(Q^0(x), 0, 0)[V_1(x), V_1(x)] \right\rangle_X \, \mathrm{d}x \end{aligned}$$

for $j = 1, 2$. Inspecting (3.16) and (3.17), we see that the integrands vanish pointwise for each x. Summarizing the findings obtained so far, we have

$$\mathcal{D}(\lambda, a) = \begin{pmatrix} M_0\lambda & 0 & 0 \\ 0 & \lambda + \mathrm{O}(a^2) & 0 \\ 0 & \mathrm{O}(a^2) & \lambda \end{pmatrix} + \mathrm{O}(|\lambda|(|\lambda + |a|)).$$

Thus, it remains to compute the diagonal $\mathrm{O}(a^2)$ term. Expanding (4.16), we see that the second derivative with respect to a of this diagonal term is given by

$$\begin{aligned}\partial_a^2\mathcal{D}_{22}(0,0) &= \int_{-\infty}^{\infty}\langle\Psi_1(x), F_{UUU}(Q^0(x),0,0)[V_1(x)]^3\\ &\quad + 3F_{UU}(Q^0(x),0,0)[V_1(x),\partial_a\widetilde{V}(x;0,0)]\\ &\quad + \varepsilon_{aa}(0)D_\varepsilon(F_U(Q^\varepsilon(x),\varepsilon,0))|_{\varepsilon=0}V_1(x)\\ &\quad + \Omega_{aa}(0)\partial_\Omega F_U(Q^0(x),0,0)V_1\rangle_X\,\mathrm{d}x\\ &= \int_{-\infty}^{\infty}\langle\Psi_1(x), F_{UUU}(Q^0(x),0,0)[V_1(x)]^3\\ &\quad + 3F_{UU}(Q^0(x),0,0)[V_1(x),\partial_a\widetilde{V}(x;0,0)]\rangle_X\,\mathrm{d}x - \varepsilon_{aa}(0)\operatorname{Re}\lambda_\varepsilon(0),\end{aligned}$$

where the term involving Ω vanishes for the same reason that shows that the first component in (3.33) is zero. Comparing the integral term in the above expression with (3.38), we see that they coincide provided $\partial_a\widetilde{V}(x;0,0) = \partial_a\widetilde{W}(x;0,0,0)$. The following lemma, whose proof we postpone until after we finished the discussion of $\mathcal{D}(\lambda,a)$, states that this identity indeed holds.

LEMMA 4.2. *We have* $\partial_a\widetilde{V}(x;0,0) = \partial_a\widetilde{W}(x;0,0,0)$.

Thus, we can conclude that

$$\partial_a^2\mathcal{D}_{22}(0,0) = \kappa_3 - \varepsilon_{aa}(0)\operatorname{Re}\lambda_\varepsilon(0) \overset{(3.39)}{=} 2\varepsilon_{aa}(0)\operatorname{Re}\lambda_\varepsilon(0)$$

and consequently

$$\mathcal{D}(\lambda,a) = \begin{pmatrix} M_0\lambda & 0 & 0\\ 0 & \lambda+\varepsilon_{aa}(0)\operatorname{Re}\lambda_\varepsilon(0)a^2+\mathrm{O}(a^3) & 0\\ 0 & \mathrm{O}(a^2) & \lambda\end{pmatrix} + \mathrm{O}(|\lambda|(|\lambda+|a|)).$$

The equation $\det\mathcal{D}(\lambda,a) = 0$ has therefore precisely three solutions, counted with multiplicity, near zero which are given by $\lambda = 0$ with multiplicity two and a simple zero at

$$\lambda_*(a) = -\varepsilon_{aa}(0)\operatorname{Re}\lambda_\varepsilon(0)a^2 + \mathrm{O}(a^3),$$

so that $\lambda_*(a)$ and $\varepsilon_*(a)$ have opposite signs since we assumed that $\operatorname{Re}\lambda_\varepsilon(0) > 0$. Subject to establishing Lemma 4.2, this completes the proof of Theorem 2.2.

Proof of Lemma 4.2. Expanding the relevant equations for $\widetilde{V}$ and $\widetilde{W}$, we find that both $\partial_a\widetilde{V}^\pm(x;0,0)$ and $\partial_a\widetilde{W}^\pm(x;0,0,0)$ satisfy the linear inhomogeneous differential equation

$$V_x = F_U(Q^0(x),0,0)V + F_{UU}(Q^0(x),0,0)[V_1(x),V_1(x)].$$

The asymptotic boundary conditions in the hyperbolic directions coincide for both functions, but differ for the center directions. We shall show that the center components of $\partial_a\widetilde{V}^\pm(0;0,0)$ and $\partial_a\widetilde{W}^\pm(0;0,0,0)$ are equal to each other from which we can infer that the two solutions coincide as claimed.

We begin by discussing $\widetilde{V}^\pm(x;0,0)$. Equation (4.15) implies that the $\Psi_0^*(0;a)$ components of $\widetilde{V}^\pm(0;0,a)$ coincide for all a. Since $\Psi_j^\pm(0;a)$ is perpendicular to the space on the right-hand side of (4.12), we also have

$$\langle\Psi_j^\pm(0;a), \widetilde{V}^+(0;0,a) - \widetilde{V}^-(0;0,a)\rangle_X = 0$$

for all j and all a, and we conclude that the center components of $\widetilde{V}^{\pm}(0;0,a)$ coincide for all a. Since $V^{\pm}(0) \in E_{\pm}(0,a)$ for all a, (4.9) gives

$$\langle \Psi_j^{\pm}(0;a), V^{\pm}(0)\rangle_X = 0, \qquad \langle \Psi_0^{*}(0;a), V^{\pm}(0)\rangle_X = 0$$

for all a, and a Taylor expansion gives

$$\begin{aligned} &\langle \partial_a \Psi_j^{\pm}(0;0), V_1(0)\rangle_X + \langle \Psi_j^{\pm}(0;0), \partial_a \widetilde{V}^{\pm}(0;0,0)\rangle_X = 0 \qquad \text{for all } j,\\ &\langle \partial_a \Psi_0^{*}(0;0), V_1(0)\rangle_X + \langle \Psi_0^{*}(0;0), \partial_a \widetilde{V}^{\pm}(0;0,0)\rangle_X = 0. \end{aligned} \tag{4.17}$$

We now turn to $\partial_a \widetilde{W}^{\pm}(0;0,0,0)$. We set $(\varepsilon,\Omega)=0$ and consider the solution pieces

$$U^{\pm}(x) = Q^0(x) + a[V_1(x) + \widetilde{W}^{\pm}(x;a,0,0)]$$

from section 3. By construction, we have $U^{\pm}(x) \in \mathcal{W}^{\mathrm{u}}_{0,0}$, and the conserved quantity $\mathcal{E}(U^-(x))$ does, therefore, not depend on a. Its derivative with respect to a is given by

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}a}\mathcal{E}(Q^0(0) + a[V_1(0) + \widetilde{W}^-(0;a,0,0)])\\ &= \langle \nabla\mathcal{E}(Q^0(0) + a[V_1(0) + \widetilde{W}^-(0;a,0,0)]), V_1(0) + \widetilde{W}(0;a,0,0)\rangle_X\\ &= \langle \nabla\mathcal{E}(Q^0(0) + a[V_1(0) + \mathrm{O}(a)]), V_1(0) + a\widetilde{W}_a^-(0;0,0,0) + \mathrm{O}(a^2)\rangle_X\\ &= a\left[\left\langle \frac{\mathrm{d}}{\mathrm{d}a}\nabla\mathcal{E}(Q^0(0) + aV_1(0))|_{a=0}, V_1(0)\right\rangle_X + \langle \nabla\mathcal{E}(Q^0(0)), \widetilde{W}_a^-(0;0,0,0)\rangle_X\right]\\ &\quad + \mathrm{O}(a^2). \end{aligned}$$

Thus, we get

$$\begin{aligned} &\langle \partial_a \Psi_j^{\pm}(0;0), V_1(0)\rangle_X + \langle \Psi_j^{\pm}(0;0), \widetilde{W}_a^-(0;0,0,0)\rangle_X = 0,\\ &\langle \partial_a \Psi_0^{*}(0;0), V_1(0)\rangle_X + \langle \Psi_0^{*}(0;0), \widetilde{W}_a^-(0;0,0,0)\rangle_X = 0. \end{aligned} \tag{4.18}$$

We can proceed as before to show continuity of $\widetilde{W}_a^{\pm}(0;0,0,0)$ in the center components. This fact, together with the continuity of $\widetilde{V}_a(0;0,0)$ in the center directions and (4.17) and (4.18), shows that the center components of $\partial_a \widetilde{V}^{\pm}(0;0,0)$ and $\partial_a \widetilde{W}^{\pm}(0;0,0,0)$ are indeed equal to each other as claimed. ∎

5. Discussion. There are numerous possible generalizations and extensions of our results. The crucial ingredient is the existence of the evolution operators $\Phi_{\pm}^{\mathrm{s,u}}$ for the spatial dynamical system, which requires some hyperbolicity in the spatial dynamics. The results are clearly not dependent on the particular form of the viscosity matrix: *nonlinear viscosity* $B(u)u_{xx}$ is allowed as long as the essential spectrum is nonresonant with the Hopf eigenvalue (uniform positivity is typically sufficient). We can also allow *parameter-dependent fluxes*: the parameter ε may appear explicitly in the viscosity matrix and the flux $f = f(u;\varepsilon)$.

Under- and overcompressive shocks can be treated similarly. All viscous shocks can be viewed as heteroclinic orbits in the travelling-wave ODE (3.5)

$$\begin{pmatrix} u_x \\ v_x \end{pmatrix} = \begin{pmatrix} v \\ f_u(u)v - cv \end{pmatrix} \tag{5.1}$$

which connect families of equilibria at $x = \pm\infty$. To set up the problem, we can, for instance, prescribe the values of u on ingoing characteristics. Choose manifolds $\mathcal{S}_\pm$ of $\mathbb{R}^n$ so that $T_{u^0_\pm}\mathcal{S}_\pm \oplus \mathcal{I}_\pm = \mathbb{R}^n$, where $\mathcal{I}_\pm$ is the eigenspace belonging to eigenvalues $\nu^\pm \lessgtr 0$ of $f_u(u^0_\pm) - c^0$. We then seek viscous shock waves in the intersection of $\mathcal{W}^{\mathrm{u}}(\mathcal{S}_-)$ and $\mathcal{W}^{\mathrm{s}}(\mathcal{S}_+)$, where we regard $\mathcal{S}_\pm$ as subsets of the manifold $\mathbb{R}^n \times \{0\} \subset Y_0$ of equilibria of (5.1). Both manifolds are n-dimensional, and we will assume that their intersection along the viscous shock is transverse in the parameter c; this is equivalent to the assumption that $\lambda = 0$ is a simple root of the Evans function associated with the PDE linearization at the shock [5, 8, 23]. One can now vary $\mathcal{S}_\pm = \mathcal{S}^\varepsilon_\pm$ and continue the transverse intersection provided the speed $c = c^\varepsilon$ is adjusted appropriately. If a pair of complex eigenvalues crosses the imaginary axis at $\varepsilon = 0$, the analysis in this paper can be adapted easily to show that there is a unique family of oscillatory under- or overcompressive shocks bifurcating from the primary viscous shock. As for Lax shocks, the bifurcating oscillatory shocks converge exponentially to time-independent rest states as $|x| \to \infty$ due to the presence of the n conservation laws (3.36). We remark that undercompressive shocks occur as weak denotations in combustion, which makes them interesting from an applied viewpoint.

The analysis extends also to the case of *degenerate shock waves*, where we allow for an additional center direction within the travelling-wave ODE in $\mathrm{Fix}\,\Gamma$ at either u^0_- or u^0_+. Again, suitable transversality conditions on the intersections of $\mathcal{W}^{\mathrm{u}}_-$ and $\mathcal{W}^{\mathrm{s}}_+$ together with appropriate assumptions on the nonlinear behavior of the zero characteristic speed near the shock are needed.

Problems posed in *infinite cylinders*,

$$u_t = \Delta u + \sum_j \partial_{x_j} f_j(u), \qquad x \in \mathbb{R} \times \Omega,$$

for bounded cross sections $\Omega \subset \mathbb{R}^N$ and with Neumann boundary conditions on $\mathbb{R}\times\partial\Omega$, say, can also be treated. The existence of exponential dichotomies for this problem follows from [11, 14].

The major open problem that we did not address in this paper is nonlinear stability of the bifurcating oscillatory viscous shocks. It should be possible to establish nonlinear stability using a combination of the approach via pointwise estimates developed by Howard and Zumbrun in [8, 23] and our spatial-dynamics technique which can be used to obtain the necessary estimates for the Green's function; this will be pursued elsewhere.

REFERENCES

[1] T. Brand, M. Kunze, G. Schneider, and T. Seelbach, *Hopf bifurcation and exchange of stability in diffusive media*, Arch. Ration. Mech. Anal., 171 (2004), pp. 263–296.
[2] S.-N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer, New York, 1982.
[3] A. Doelman, B. Sandstede, A. Scheel, and G. Schneider, *The dynamics of modulated wave trains*, Mem. Amer. Math. Soc., to appear.
[4] B. Fiedler, B. Sandstede, A. Scheel, and C. Wulff, *Bifurcation from relative equilibria of noncompact group actions: Skew products, meanders, and drifts*, Doc. Math., 1 (1996), pp. 479–505.
[5] R. A. Gardner and K. Zumbrun, *The gap lemma and geometric criteria for instability of viscous shock profiles*, Comm. Pure Appl. Math., 51 (1998), pp. 797–855.
[6] M. Golubitsky, V. LeBlanc, and I. Melbourne, *Meandering of the spiral tip—an alternative approach*, J. Nonlinear Sci., 7 (1997), pp. 557–586.
[7] M. Haragus and A. Scheel, *Corner defects in almost planar interface propagation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 23 (2006), pp. 283–329.

[8] P. Howard and K. Zumbrun, *Stability of undercompressive shock profiles*, J. Differential Equations, 225 (2006), pp. 308–360.
[9] T. Kapitula and B. Sandstede, *Stability of bright solitary wave solutions to perturbed nonlinear Schrödinger equations*, Phys. D, 124 (1998), pp. 58–103.
[10] T. Kato, *Perturbation Theory for Linear Operators*, Springer, New York, 1980.
[11] D. Peterhof, B. Sandstede, and A. Scheel, *Exponential dichotomies for solitary wave solutions of semilinear elliptic equations on infinite cylinders*, J. Differential Equations, 140 (1997), pp. 266–308.
[12] B. Sandstede, *Stability of travelling waves*, in Handbook of Dynamical Systems II, B. Fiedler, ed., North–Holland, Amsterdam, 2002, pp. 983–1055.
[13] B. Sandstede and A. Scheel, *Essential instability of pulses and bifurcations to modulated travelling waves*, Proc. Roy. Soc. Edinburgh Sect. A, 129 (1999), pp. 1263–1290.
[14] B. Sandstede and A. Scheel, *On the structure of spectra of modulated travelling waves*, Math. Nachr., 232 (2001), pp. 39–93.
[15] B. Sandstede and A. Scheel, *Essential instabilities of fronts: Bifurcation and bifurcation failure*, Dyn. Syst., 16 (2001), pp. 1–28.
[16] B. Sandstede and A. Scheel, *Absolute instabilities of standing pulses*, Nonlinearity, 18 (2005), pp. 331–378.
[17] B. Sandstede, A. Scheel, and C. Wulff, *Dynamics of spiral waves on unbounded domains using center-manifold reduction*, J. Differential Equations, 141 (1997), pp. 122–149.
[18] A. Scheel, *Coarsening fronts*, Arch. Ration. Mech. Anal., 181 (2006), pp. 505–534.
[19] B. Texier and K. Zumbrun, *Relative Poincaré–Hopf bifurcation and galloping instability of travelling waves*, Methods Appl. Anal., 12 (2005), pp. 348–380.
[20] B. Texier and K. Zumbrun, *Galloping instability of viscous shock waves*, Phys. D, to appear.
[21] B. Texier and K. Zumbrun, *Hopf bifurcation of viscous shock solutions of compressible Navier–Stokes and MHD*, Arch. Ration. Mech. Anal., to appear.
[22] A. Vanderbauwhede and G. Iooss, *Center manifold theory in infinite dimensions*, Dynamics Reported, 1 (1992), pp. 125–163.
[23] K. Zumbrun and P. Howard, *Pointwise semigroup methods and stability of viscous shock waves*, Indiana Univ. Math. J., 47 (1998), pp. 741–871.